NeuroImage 94 (2014) 275–286

Contents lists available at ScienceDirect

NeuroImage

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

Manifold population modeling as a neuro-imaging biomarker: Application to ADNI and ADNI-GO

R. Guerrero *, R. Wolz, A.W. Rao, D. Rueckert, The Alzheimer's Disease Neuroimaging Initiative (ADNI) [1,2]

Department of Computing, Imperial College, London, UK

ARTICLE INFO

Article history:
Accepted 12 March 2014
Available online 21 March 2014

Keywords:
Alzheimer's disease (AD)
Mild cognitive impairment (MCI)
Classification
Sparse regression
Kernel density estimation
Manifold learning
Laplacian eigenmaps

ABSTRACT

We propose a framework for feature extraction from learned low-dimensional subspaces that represent inter-subject variability. The manifold subspace is built from data-driven regions of interest (ROI). The regions are learned via sparse regression using the mini-mental state examination (MMSE) score as an independent variable which correlates better with the actual disease stage than a discrete class label. The sparse regression is used to perform variable selection along with a re-sampling scheme to reduce sampling bias. We then use the learned manifold coordinates to perform visualization and classification of the subjects. Results of the proposed approach are shown using the ADNI and ADNI-GO datasets. Three types of classification techniques, including a new MRI Disease-State-Score (MRI-DSS) classifier, are tested in conjunction with two learning strategies. In the first case Alzheimer's Disease (AD) and progressive mild cognitive impairment (pMCI) subjects were grouped together, while cognitive normal (CN) and stable mild cognitive impaired (sMCI) subjects were also grouped together. In the second approach, the classifiers are learned using the original class labels (with no grouping). We show results that are comparable to other state-of-the-art methods. A classification rate of 71%, of arguably the most clinically relevant subjects, sMCI and pMCI, is shown. Additionally, we present classification accuracies between CN and early MCI (eMCI) subjects, from the ADNI-GO dataset, of 65%. To our knowledge this is the first time classification accuracies for eMCI patients have been reported.

Introduction

Alzheimer's disease (AD) is the most common form of dementia, usually associated with the elderly population (over 65 years of age). AD had a worldwide prevalence of around 26.6 million cases reported in 2006, and predictions suggest that this figure will increase fourfold to above 100 million by the year 2050 (Brookmeyer et al., 2007). If intervention could achieve even a modest one year delay of both disease onset and progression, there would be nearly nine million fewer cases of the disease by that time (Brookmeyer et al., 2007). Postulated interventions are more likely to be effective in early stages of the disease. These figures underline the huge impact advances in early diagnosis might have on the overall well-being of the population, the burden to caregivers and family members, as well as the associated financial costs to the world's health systems. Several studies over recent years have concluded and confirmed that AD can be diagnosed by clinical assessment alone accurately in 90% of the cases when validated against neuropathological standards (Ranginwala et al., 2008). However, by the time a patient is diagnosed he/she may already suffer from substantial loss of quality-of-life and chances for improvement, or even disease progression deceleration, may have deteriorated. Hence, the importance of very early diagnosis of the onset of dementia.

Several medications are currently approved by the U.S. Food and Drug Administration (FDA) to treat people who have been diagnosed with AD. Treating the symptoms of AD can provide patients with comfort, dignity, and independence for a longer period of time and can encourage and assist their caregivers as well. Disease modifying treatments are more likely to have a significant impact in the earlier stages of the disease. Population stratification is important to allow the recruitment of appropriate subjects for clinical trials, and explore the effects of novel treatments in subjects where results are expected to be most effective, hence, reducing overall costs of the trial by removing false positives in an earlier stage. Of special interest are subjects with amnestic mild cognitive impairment (MCI), which is a prodromal form of AD. Existing studies have suggested that about 10–12% of subjects with amnestic MCI progress to probable AD per year (Petersen et al., 1999). However, individual patients can remain in a stable MCI (sMCI) condition for years. From a clinical perspective it is therefore

* Corresponding author.
E-mail address: reg09@imperial.ac.uk (R. Guerrero).

[1] This project is partially funded by CONACyT, the Rabin Enzra trust, the 7th Framework Programme by the European Commission (http://cordis.europa.eu/ist/) and SEP-DGRI.

[2] Data used in the preparation of this article were obtained from the 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 the analysis or writing of this report. A complete listing of ADNI investigators is available at www.loni.ucla.edu/ADNI/Collaboration/ADNI_Authorship_list.pdf.

http://dx.doi.org/10.1016/j.neuroimage.2014.03.036
1053-8119/

particularly interesting to identify those subjects that are at immediate or medium-term risk of progressing from MCI to AD (pMCI).

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a study with the primary goal of testing whether serial MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and AD. Recent studies focus on identifying subjects at risk at a much earlier stage. In the ADNI Grand Opportunity (ADNI-GO) and ADNI-2 studies, a group of early MCI (eMCI) patients is included (Aisen et al., 2010) that represents individuals with milder degrees of cognitive and functional impairment than the MCI subjects. With a slower rate of progression, they form an especially interesting subject group as biomarker manifestation could potentially be different at such an early stage of the disease.

Imaging biomarkers play an increasingly important role in the early diagnosis of neurodegenerative diseases like AD. Magnetic resonance (MR) imaging examinations often form part of clinical assessment standards for patients with MCI. Biomarkers based on MR imaging are considered to be more sensitive to change after symptoms from amyloid-based biomarkers start to appear (i.e. accumulation of amyloid-β) (Frisoni et al., 2010). Imaging biomarker measurements can be key in the development of disease-modifying drugs. They can be used to explore the modifying effects these drugs may have on the disease trajectory through time, and also as a screening tool to select a more homogeneous prodromal patient population that are expected to have higher risk for rapid imminent clinical progression, thus increasing the efficiency of clinical trials (Hampel et al., 2010). Recently, there have been many studies with a main focus on automatically identifying such imaging biomarker. Many of the well-established and well-known biomarkers used in dementia that are derived from MR imaging are based on morphological measurements of specific brain structures (i.e. hippocampi, amygdalae, cortex, entorhinal cortex), such as volumes or shapes (Cho et al., 2012; Chupin et al., 2009; Coupé et al., 2012; Cuingnet et al., 2011; Koikkalainen et al., 2011; Lerch et al., 2008; Querbes et al., 2009; Westman et al., 2011; Wolz et al., 2010, 2011). However, neurodegeneration patterns may not necessarily follow strict standard definitions of anatomical structures or functional regions. Hence, limiting the analysis to predefined regions could potentially reduce the power of the biomarker to detect differences or changes over time.

More recently, the problem of learning clinically useful biomarkers has been cast as a machine learning problem. Models that derive from developments in the machine learning community have been put forward as an alternative to seek for discriminative features that could act as AD biomarkers independent from a predefined parcellation of structures (Davatzikos et al., 2011; Eskildsen et al., 2013; Fan et al., 2007, 2008; Gerardin et al., 2009; Misra et al., 2009; Vemuri et al., 2008; Wolz et al., 2012; Zhang et al., 2012). This independence could potentially lead to a better modeling of a disease trajectory for the whole brain and across time. Furthermore, this would account for the fact that the disease trajectory manifests itself at different regions at different times.

Some of the potential pitfalls when working with high-dimensional data, such as medical images, can be associated with the curse of dimensionality. This describes a general paradox that occurs in high-dimensional space, where if a neighborhood is considered “local”, then it will be most likely “empty”, while a “non-empty” neighborhood will probably be “non-local”. This implies that in high-dimensional space the variance-bias trade-off cannot be accomplished very well, unless there is a very large amount of samples available. That is, to keep the variance low the neighborhood has to be made large enough to include enough samples, but then a very large bias is introduced due to the large neighborhood, and vice versa (Scott, 1992).

Learning a low-dimensional subspace representation of complex very high-dimensional objects (i.e. images) is a central problem of machine learning and pattern recognition. Several methods have been proposed to find the underlying low-dimensional space of intrinsically low-dimensional data that is embedded in a high-dimensional space. A low-dimensional representation of the data allows the use of modeling techniques that suffer from the small sample size problem in high-dimensional spaces. There is a long history in the use of linear dimensionality reduction techniques, like principal component analysis (PCA) and multidimensional scaling (MDS) (Cox and Cox, 1994), across different fields. Recently, nonlinear techniques like principal curves (Hastie and Stuetzle, 1989), ISOMAP (Tenenbaum et al., 2000), locally linear embedding (LLE) (Roweis and Saul, 2000), or Laplacian eigenmaps (LE) (Belkin and Niyogi, 2002) have been proposed to better capture the variation of highly nonlinear data. For a comprehensive review of dimensionality reduction techniques see van der Maaten et al. (2009).

Working with brain MR images and using concepts from dimensionality reduction, Aljabar et al. (2009) applied spectral analysis to pairwise label overlaps obtained from a structural segmentation to discriminate AD patients from CN subjects. Klein et al. (2010) used vectors defined by the similarities between a given test subject and a set of training images as features from which to learn a classifier. Some dimensionality reduction techniques aim to model global variability over the whole dataset, which could potentially limit their generalization power of the learned subspace when dealing with complex datasets. In recent work, it has been suggested that this is indeed the case when dealing with brain images, and that nonlinear methods better capture the natural variability of such images (Gerber et al., 2010; Hamm et al., 2010). Wolz et al. (2012) propose to classify a subject's disease state in a manifold space that is learned from image similarities measured over an a-priori defined region of interest (ROI) and (clinical) meta-information related to the subject. However, as stated before, patterns of neurodegeneration may not necessarily be best observed in the predefined ROI, since useful information could potentially be ignored. On the other hand the ROI will most likely contain regions that are not associated with the neurodegeneration pattern, and this could confound the learned subspace. Furthermore, subject classification is performed by applying a support vector machine (SVM) approach to manifold coordinates. SVM finds a separation hyperplane defined by only a subset of subjects (support vectors) that lie close to the hyperplane in the learned subspace.

There are two fundamental problems when dealing with high-dimensional data such as 3D brain MR images: First, there is a large amount of variables (voxels) available in images, and not all contribute equally (or at all) to the modeling of the disease trajectory. Relevant variable selection from this large pool of predictor variables is a way to tackle this problem. We assume that the underlying disease trajectory manifests itself on a small subset of variables in an image, and so it can be modeled using a sparse set of voxels. The L_1 norm has been proposed as an effective solution to the variable selection problem (Tibshirani, 1996; Zou and Hastie, 2005). Secondly, the variable selection process often is an ill-posed problem, where the sample size is much smaller than the number of variables and variables are highly correlated. That is, the L_1 norm can only select up to N uncorrelated variables, where N is the number of samples. Although the dimensionality reduction techniques mentioned before can deal this issue, all variables contribute to the manifold estimation process.

We propose to use sparse regression to learn a ROI in which a distance measure allows us to define a manifold space that better describes the different stages of AD, by defining the ROI where the disease trajectory can be better observed and quantified. The resulting compact manifold representation has a sufficiently low dimensionality to allow us to model different populations directly from the learned manifold coordinates. The population distribution models of the observed data can be used to infer the disease state of a new patient by embedding it in the manifold and obtaining a probabilistic score on class correspondence as opposed to a discrete label as in classification approaches. This probabilistic estimation allows us to move away from a discrete decision based on hyperplanes to a continuous characterization or modeling of disease progression via the proposed MRI Disease-State-Score (MRI-

DSS) formulation, that fully takes advantage of subspace coordinates and yields a continuous variable on the disease trajectory.

This paper is organized as follows: In the following section we describe the characteristics and pre-processing steps of the datasets. We also provide a detailed description of the methodology used in this study (variable selection, manifold learning, population modeling and disease state score). The Results section summarizes the experiments carried out and the results obtained. Finally, a discussion of the results obtained is presented, followed by conclusions and directions for future work.

Material and methods

Data

Data used in this article was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public–private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early Alzheimer's disease (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, MD, VA Medical Center and University of California San Francisco. ADNI is the result of efforts of many co-investigators 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 aged 55 to 90, to participate in the research, approximately 200 CN older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years and 200 people with early AD to be followed for 2 years. For up-to-date information, see www.adni-info.org.

In this work, we used the subset of 523 subjects for which T1-weighted 1.5 T MR images were available at baseline, 12 and 24 month follow-up, as of October 2011. 12 of those subjects where discarded due to label ambiguities (subjects whose labels changed from MCI to CN or from AD to MCI). The remaining 511 subjects consisted of 106 patients diagnosed as probable AD, 230 as MCI (114 sMCI and 116 pMCI) and 175 CN (see Table 1 for a description of the demographics). The remaining 315 images (56 CN, 119 sMCI, 49 pMCI and 91 AD) that did not contain all time points were used as training data in the variable selection scheme (Section 2.2).

Additionally, experiments were carried out using the ADNI-GO dataset (Aisen et al., 2010). The purpose of the ADNI-GO study is to build upon the information obtained in the original ADNI, to examine how brain imaging can be used with other tests to measure the progression of MCI and early AD. ADNI-GO seeks to define and characterize the mildest symptomatic phase of AD, referred to in this study as early amnestic MCI (eMCI). However, generally no formal sub-categorization between eMCI and MCI (or late MCI) exists. The eMCI subjects represent individuals with milder degrees of cognitive and functional impairment than the MCI subjects, and their rate of progression is slower (Aisen et al., 2010). From this dataset, all the available images labeled as CN or eMCI were selected and preprocessed in the same way as with ADNI (see Table 2 for a description of the demographics).

Image preprocessing

In this study, all the images used were skull stripped using multi-atlas segmentation (Leung et al., 2011) and intensity normalized at a global scale using a piecewise linear function (Nyl and Udupa, 1999). Intensity normalization was carried out following an iterative scheme, where all images are normalized to a common template/subject, then the template was changed and all the images were re-normalized to the new template. This was repeated N times, where N is the number of subjects to aid in removing normalization bias (Coupé et al., 2012). Also, all images were transformed to a common space, the MNI152 template, and hence re-sliced and re-sampled to an isotropic voxel size of 1 mm. A coarse free-form-deformation (Rueckert et al., 1999), using a control point spacing of 10 mm, was carried out to remove gross anatomical variability while aligning anatomical structures in order to focus on more local variation. In order to account for disease manifestation and progression in left-handed and right-handed populations, and hence find more generalizable regions, the selected variables from the Relevant variable selection section are mirrored along left–right hemispheres prior to the subsequent steps.

Relevant variable selection

Regression techniques allow the modeling and analysis of several variables, where the focus is on modeling the relationship between a dependent variable and one or more independent variables. Over the years several regression methods have been proposed (Tibshirani, 1996; Tikhonov and Arsenin, 1977; Zou and Hastie, 2005), with arguably the simplest method being ordinary least squares regression (OLSR). In high-dimensional problems, however, the solution to OLSR is not unique and so some form of regularization is required in order for the model to generalize well beyond the training data. In ridge regression, this is achieved by incorporating an L_2 penalty into the OLSR objective function, which leads to a unique solution in which correlated predictors are given similar regression weights. LASSO regression, on the other hand, uses an L_1 penalty that regularizes the problem by encouraging a sparse solution in which most of the estimated regression weights are zero. This is a highly desirable trait when dealing with high-dimensional data because it allows for variable selection. Two of the main problems with LASSO are that it does not perform well in the presence of highly correlated variables (i.e. neighboring voxels in an image would probably be very well correlated) and that it can only select a number of variables that is up to the number of samples (a significant problem for high-dimensional data). Elastic net regression (Zou and Hastie, 2005) seeks to address the drawbacks of the LASSO (Tibshirani, 1996), i.e. it allows selecting a number of variables that is greater than the number of samples. This is done by adding an additional L_2 penalty term on the model's coefficients to the LASSO.

Table 1
Subject groups mean age, sample size, MMSE scores, gender, CDR scores and weight data (with standard deviation in brackets) from ADNI.

	N	Age	MMSE	Men	CDR	Weight
CN	175	76.34 ± 5.11	29.17 ± 0.97	52% (91)	0 ± 0.1	74.43 ± 15.57
sMCI	114	75.12 ± 6.67	27.29 ± 2.25	66% (75)	0.49 ± 0.05	77.02 ± 12.83
pMCI	116	74.73 ± 6.93	26.64 ± 1.7	63% (73)	0.5 ± 0.05	74.56 ± 14.41
AD	106	75.4 ± 7.39	23.25 ± 1.97	53% (56)	0.77 ± 0.25	72.58 ± 13.81

Table 2
Subject groups mean age, sample size, MMSE scores, gender and weight data (with standard deviation in brackets) from ADNI-GO.

	N	Age	MMSE	Men	Weight
CN	134	73.77 ± 10.85	28.99 ± 1.23	51% (68)	75.68 ± 15.08
eMCI	229	67.42 ± 18.61	28.29 ± 1.53	54% (124)	81.47 ± 15.89

Elastic net

The elastic net performs automatic variable selection and continuous shrinkage. Additionally, it encourages the grouping of highly correlated variables. It can be formulated as follows:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \|\mathbf{1} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_R \|\boldsymbol{\beta}\|_2^2 + \lambda_L \|\boldsymbol{\beta}\|_1 \right\}. \quad (1)$$

Here $\mathbf{X}$ is an N by p matrix containing N vectorized images, $\boldsymbol{\beta}$ is an N by p matrix of the regression coefficients, $\mathbf{1}$ is the vector of response variables, λ_R and λ_L are the ridge regression and the LASSO regression penalty weights, respectively. In Eq. (1), the L_1 term encourages solutions that are sparse, while the L_2 term promotes the grouping of correlated variables. Grouping correlated variables can be viewed as desirable in some applications. For example, as mentioned before neighboring voxels in an image are expected to be correlated, hence their grouped selection as predictor variables can be seen as an ROI learning algorithm. Viewed as an image regression problem, the elastic net finds regions of interest (predictor variables) within the images $\mathbf{X}$ that are useful to regress a variable $\mathbf{1}$ associated to each image. This could be the clinical label or the mini-mental state examination (MMSE) score associated to a patient. The elastic net objective function, Eq. (1), can be solved efficiently using the LARS-EN algorithm (Zou and Hastie, 2005), which allows for the number of steps or number of variables selected to be easily incorporated as early termination criteria. It is worth noting that Eq. (1), in the special cases where λ_R and λ_L are set to zero, becomes the ordinary least square regression. If λ_L is set to zero then it describes a ridge regression and if λ_R is set to zero we obtain a LASSO regression. Another special case arises when $\lambda_R \to \infty$: It can be shown (Zou and Hastie, 2005) that for each predictor variable $\mathbf{x}_i$, minimizing Eq. (1) has a closed-form solution that can be written as:

$$\underset{\lambda_R \to \infty}{\hat{\boldsymbol{\beta}}_i} = \left(\left| \mathbf{I}^T \mathbf{x}_i \right| - \frac{\lambda_L}{2} \right)_+ \operatorname{sign}\left(\mathbf{I}^T \mathbf{x}_i \right), \qquad i = 1, 2, \ldots, p. \quad (2)$$

where $(\cdot)_+$ refers to the positive part.

This can be solved very efficiently, since $\mathbf{I}^T\mathbf{x}_i$ is the univariate regression coefficient of the ith predictor, the estimates $\hat{\beta}_i$ are obtained by applying a soft threshold to the univariate regression coefficients. Eq. (2) is also known as univariate soft thresholding.

As stated before, the L_2 regularization term (ridge) encourages the selection of correlated variables. In medical images it can be expected that voxels (variables) that belong to the same anatomical structure will have a high degree of correlation within each other. Choosing $\lambda_R \to \infty$ imposes a maximal grouping condition on Eq. (1). In this setting elastic net regression can be used as a ROI learning algorithm.

As $\lambda_R \to \infty$, we are left with only one free parameter λ_L, from which we will drop the subindex and refer to it only as λ from now on. Eq. (2) can be solved for a range of regularization parameters λ when we find the full regularization path, $\lambda_{\min} \le \lambda \le \lambda_{\max}$ up to the desired stopping criteria in the same way as one would using the LARS-EN algorithm. In our case we limit the step size on the path such that we ensure that at each step we add only one variable.

Training re-sampling

In order to increase robustness against sampling errors from the dataset, we adopt a re-sampling scheme. In this approach, the regularization path is found on B random subsets, solving Eq. (2) over a range of values $\lambda \in [\lambda_{max}, \lambda_{min}]$, such that zero variables are included at λ_{max}, K variables are included at λ_{min} and with each step only one more variable is added. At each step k on a subset b we obtain a set of regression coefficients $\hat{\boldsymbol{\beta}}_{b,k}(\lambda_{b,k})$, where $b = 1, 2, \ldots, B$ and $k = 1, 2, \ldots, K$. We define an indicator variable $\Psi_{bk}(\lambda_{b,k})$ which is set to one if the coefficient corresponding to variable x_j is non-zero, and is set to zero otherwise. The relevance of each variable is measured by defining the probability of it being selected by the regressor as

$$P_{v_j}\left(\lambda_{B,K}\right) = \frac{1}{BK} \sum_{b=1}^{B} \sum_{k=1}^{K} \Psi_{b,k}\left(\lambda_{b,k}\right), \quad j = 1, 2, \ldots, p. \quad (3)$$

Thresholding the probabilities P_v at τ to keep the most relevant voxels, yields a mask that defines a ROI that correlates with the disease progression.

Manifold learning

One of the aims of this work is to produce continuous models of the different discrete stages. For this purpose the learned ROI is still relatively high-dimensional and hence the curse of dimensionality (Scott, 1992) would generally still impede the estimation of generalizable continuous model.

Manifold learning in general refers to a set of machine learning techniques that aim at finding a low-dimensional representation of high-dimensional data while trying to faithfully represent the intrinsic geometry of the data. For example, if we have an image dataset and each image is considered a single point in a very high-dimensional space, then this high-dimensional space is probably overcomplete in the sense that a sub-manifold of far fewer dimensions (that is most likely to be non-linear) may represent most of the variation in the dataset. In manifold learning, a similarity or distance matrix is typically used to represent the relations between pairs of data items, which can be assumed to be either the original images or some set of features derived from the images. This matrix may be viewed as a graph in which each node corresponds to an image and the weight of each edge encodes a similarity or distance between the images or derived features.

In our framework, given a set of N vectors of length D that define the most relevant voxels (variables) $\mathbf{V} = \{v_1, v_2, \ldots, v_N\} \in \mathbb{R}^D$ from a set of images, the aim is to learn the underlying manifold in $\mathbb{R}^d$ ($d \ll D$) that best represents the population from $\mathbf{V}$. Here $v_i = \{v_1, v_2, \ldots, v_D\}$ are the weighted most relevant voxels in image i.

Laplacian eigenmaps

Laplacian eigenmaps can be used to derive a low-dimensional representation of the data $f: \mathbf{V} \to \mathbf{Y}, \mathbf{y}_i = f(\mathbf{v}_i)$ while preserving the local geometric properties of the manifold (Belkin and Niyogi, 2002). Laplacian eigenmaps uses a sparse, local neighborhood graph to approximate geodesic distances among data points. In Belkin and Niyogi (2002) a distance (dissimilarity) measure is used to identify the k-neighborhood around each point. From these distances, a sparse neighborhood graph G is constructed. A weight matrix $\mathbf{W}$ that converts distances to similarities and assigns a value to each edge in G is computed using a Gaussian heat kernel

$$w_{i,j} = K\left(\mathbf{v}_i, \mathbf{v}_j\right) = \exp\left(\frac{-\left\| \mathbf{v}_i - \mathbf{v}_j \right\|^2}{2\sigma^2} \right), \quad (4)$$

with standard deviation σ.

In our work we use the cross-correlation (similarity) measure to identify the k-neighborhood around each point and do not use a heat kernel. From these similarities, the sparse neighborhood graph G is constructed and the edge weight matrix W simply uses the cross-correlation values. Avoiding the use of the heat kernel eliminates the need of its parameter σ.

Laplacian eigenmaps aims to place points $\mathbf{v}_i$ and $\mathbf{v}_j$ close together in the low-dimensional space if their similarity or weight $w_{i,j}$ is high, i.e. if

they are close in the original, high-dimensional space. This is achieved by means of minimizing the following cost function

$$\phi(\mathbf{Y}) = \underset{\mathrm{y}}{\operatorname{argmin}} \sum_{i,j} \left\|\mathbf{y}_i - \mathbf{y}_j\right\|^2 w_{i,j}, \tag{5}$$

under the constraint that $\mathbf{y}^T\mathbf{D}\mathbf{y} = 1$ which removes an arbitrary scaling factor in the embedding and prevents the trivial solution where all $\mathbf{y}_i$ are zero. The minimization of Eq. (5) can be formulated as an eigenproblem (Anderson and Morley, 1985). This can be calculated from the weight matrix $\mathbf{W}$, the degree matrix $\mathbf{M}$ and the graph Laplacian $\mathbf{L} = \mathbf{M} - \mathbf{W}$. The degree matrix is a diagonal matrix that contains information about the degree of each vertex of $\mathbf{W}$, where $m_{i,i} = \sum_j w_{i,j}$. Hence the low-dimensional manifold $\mathbf{Y}$ that represents all the data points can be obtained via solving a generalized eigenproblem

$$\mathbf{L}\nu = \mu\mathbf{M}\nu, \tag{6}$$

where ν and μ are the eigenvectors and eigenvalues, respectively. In turn the d eigenvectors corresponding to the smallest (non-zero) eigenvalues represent the new coordinate system.

Population distribution modeling

It is now widely accepted that both pathological processes and clinical decline occur gradually over time, with AD being the end stage of the accumulation and progression of these pathological changes. Additionally, the current consensus on AD is that these changes begin years before the earliest clinical symptoms occur (Jack et al., 2010). Hence, AD biomarkers need to reflect this temporal progression, and imaging biomarkers are not an exception.

As stated before, part of the aim in this work is to produce models of the different discrete stages using continuous probabilistic Gaussian mixture models in order to make predictions of group assignment or disease evolution of unseen samples. This modeling is done using the coordinates of the low-dimensional representation found using Laplacian eigenmaps in order to avoid the curse of dimensionality associated with the high-dimensional space. In Parzen kernel density estimation (KDE), each observation sample is treated as a component in a mixture model. See Scott (1992) and Wand and Jones (1994) for a detailed description of multivariate kernel density estimation. Each sample in the manifold can be viewed as a single Dirac delta function, which can be written as a Gaussian with zero covariance, with its probability concentrated at the point itself, we can define a multivariate and N-component Gaussian mixture model of the sample distribution as (Kristan et al., 2011):

$$P_s(y) = \sum_{i=1}^{N} = \alpha_i \phi_{\Sigma_i}(y - y_i), \tag{7}$$

where ϕ_{Σ_i} defines a Gaussian of mean y_i and covariance Σ_i that belong to the sample mixture model distribution.

Defining the KDE, $\hat{P}_{KDE}(y)$, as the convolution between the sample distribution $P_s(y)$ and a Gaussian kernel with a covariance matrix (also known as the bandwidth) $\mathbf{H}$, we get:

$$\hat{P}_{KDE}(y) = \phi_{\mathbf{H}}(y) * P_s(y) = \sum_{i=1}^{N} \alpha_i \phi_{\mathbf{H}+\Sigma_i}(y - y_i), \tag{8}$$

where * denotes a convolution.

Considering the case where the sample distribution P_s is a Gaussian mixture model, with $\Sigma_i = 0$ (Dirac delta functions), Eq. (8) can be rewritten as

$$\hat{P}_{KDE}(y) = \sum_{i=1}^{N} \alpha_i \phi_{\mathbf{H}}(y - y_i). \tag{9}$$

The asymptotic mean integrated squared error (AMISE) allows us to measure the fit of the estimated distribution $\hat{P}_{KDE}(y)$ to the unknown underlying distribution $P_u(y)$, and it is defined as

$$\mathrm{AMISE} = (4\pi)^{-d/2}|\mathbf{H}|^{-1/2}N_\alpha^{-1} + \frac{1}{4}d^2 \int tr^2\left\{\mathbf{H}\mathcal{G}_{P_u}(y)\right\}dy \tag{10}$$

where $\mathrm{tr}\{\cdot\}$ is the trace, $\mathcal{G}_{P_u}(y)$ is the Hessian of the unknown probability $P_u(y)$ and $N_\alpha = (\sum_{i=1}^{N} \alpha_i^2)^{-1}$.

We can use AMISE to determine the optimal bandwidth $\mathbf{H}$ (according to the observable data) of the kernel used in $\hat{P}_{KDE}(y)$ to estimate $P_u(y)$. Defining $\mathbf{H}$ in terms of scale ξ and structure $\mathbf{F}$ as $\mathbf{H} = \xi^2\mathbf{F}$ then the AMISE measure is minimized at

$$\xi_{opt} = \left[d(4\pi)^{d/2}N_\alpha|\mathbf{F}|^{1/2}R(P_u, \mathbf{F})\right]^{-1/(d+4)}, \tag{11}$$

with

$$R(P_u, \mathbf{F}) = \int tr^2\left\{\mathbf{F}\mathcal{G}_{P_u}(y)\right\}dy, \tag{12}$$

and since P_u is unknown, $R(P_u, \mathbf{F})$ is approximated as

$$\hat{R}(P_u, \mathbf{F}, G) = \int \mathrm{tr}\left\{\mathbf{F}\mathcal{G}_{P_G}(y)\right\}\mathrm{tr}\left\{\mathbf{F}\mathcal{G}_{P_s}(y)\right\}, \tag{13}$$

where P_s is the sample and P_G is the so-called pilot distribution with covariance matrix $\Sigma_{gj} = G + \Sigma_{sj}$ and G is the pilot bandwidth estimated using the multivariate normal scale rule (Duong and Hazelton, 2003; Wand and Jones, 1994) as

$$G = \hat{\Sigma}_{smp}\left(\frac{4}{(d+2)N_\alpha}\right)^{2/(d+4)} \tag{14}$$

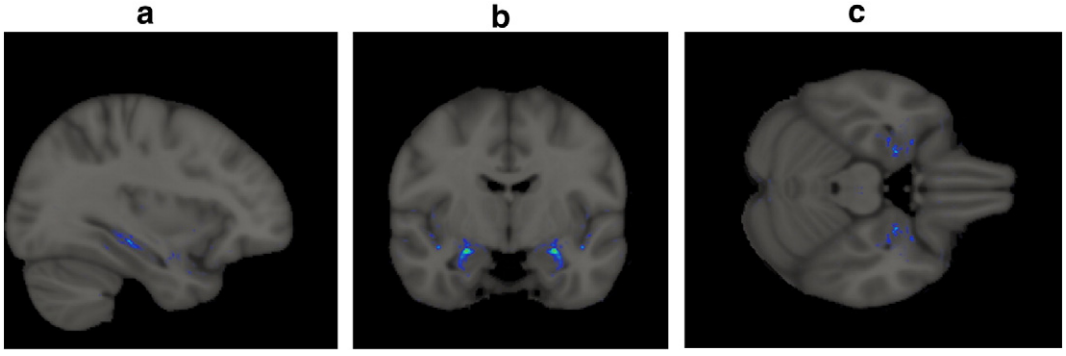

Fig. 1. (a) Sagittal,(b) coronal and (c) axial orthogonal views of MMSE probabilistic variable selection mask (best seen in color).

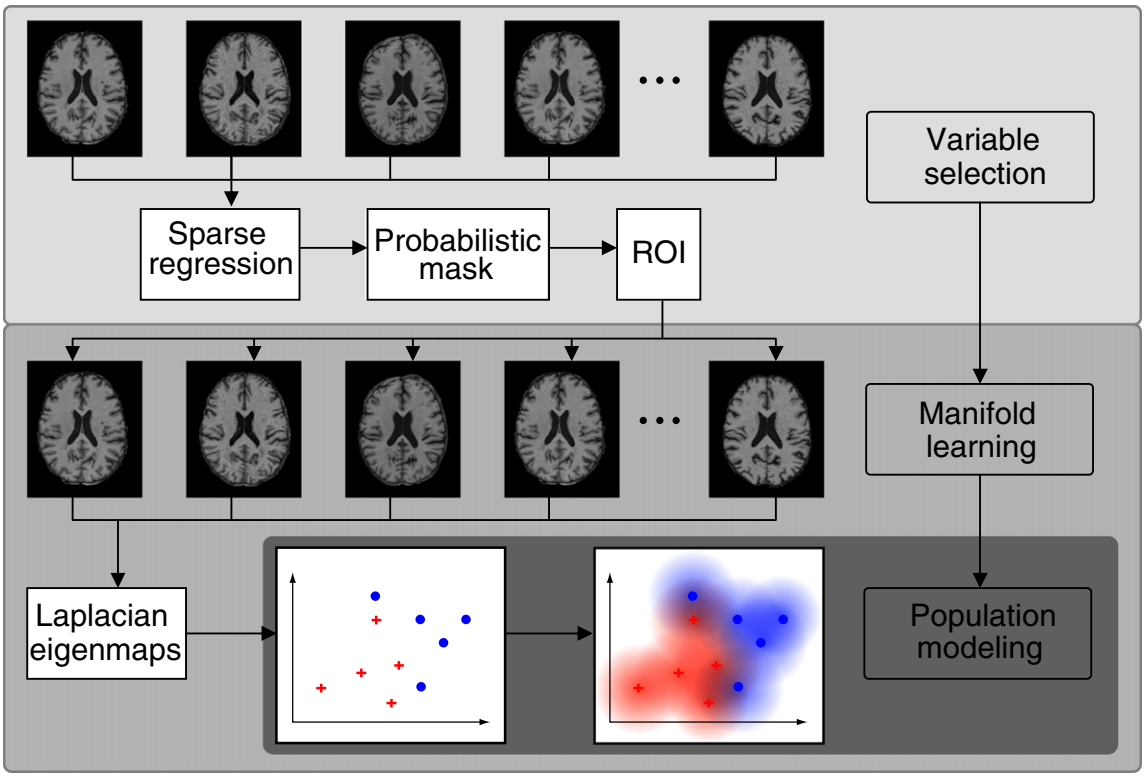

Fig. 2. Diagram showing the three main stages of the method: variable selection, manifold learning and population modeling (best seen in color).

where $\hat{\Sigma}_{smp}$ is the estimated covariance from all available samples. The structure **F** of the bandwidth **H** is approximated using the covariance matrix of the samples as $\mathbf{F} = \hat{\Sigma}_{smp}$ (Wand and Jones, 1994).

If the number of samples N is large and is made available to the population density estimation procedure described above, then the Gaussian mixture model defined by $\hat{P}_{KDE}(y)$ can be unnecessarily complex and over fitted to the data. Hence a model compression step can be used to reduce the model components (Kristan et al., 2011) from N to M, where $M < N$, as long as the compressed distribution $\widehat{P}_{KDE}(y)$ is within a certain Helliger distance (Pollard, 2002). This means that if

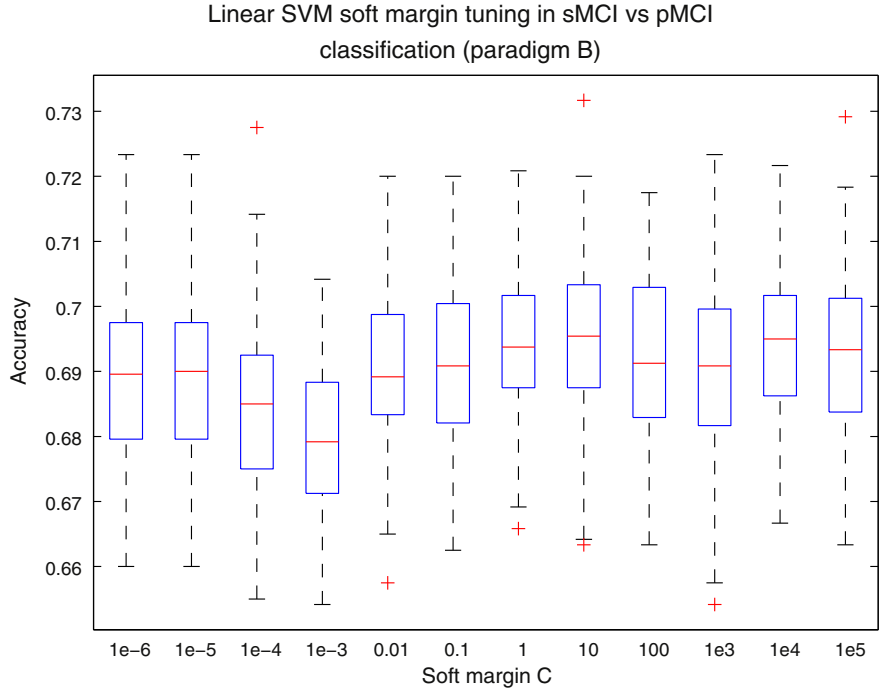

Fig. 3. Boxplot of results from a grid search of the soft margin parameter C in linear SVM. Instances are an average of the accuracies obtained across 50 manifolds (with 1–50 dimensions). The middle red line, box, whiskers and crosses represent the median, the 75th percentile, the extremes and the outliers, respectively. 100 runs done for every value of C.

$K \in N$ sample points are close to each other, then their corresponding Gaussians in the mixture model can be combined into a single Gaussian with a weight $\hat{\alpha}_i = \sum_{j=1}^{K} \alpha_j$.

MRI disease-state-score

We propose to model different stages in the disease trajectory using the probabilistic distribution of different classes that can be estimated from different class populations (Population distribution modeling section) and from the samples' coordinates in the low-dimensional manifold (Manifold learning section). We then construct a probabilistic scoring function that determines the class likelihood in the low-dimensional space, and hence, model the disease trajectory as a continuous variable. Thus

$$f(y) = P_B(y) - P_A(y)$$
$$f(y) = \sum_{v}^{N_B} \alpha_{Bi} \phi_{\Sigma_{Bi}}(y) - \sum_{i=1}^{N_A} \alpha_{Ai} \phi_{\Sigma_{Ai}}(y), \quad (15)$$

where P_A and P_B are the estimated probability distributions of classes A and B, respectively, α_{*i} and Σ_{*i} are the weights and covariance associated with the ith element in the Gaussian mixture model and N_A and N_B are the number of components in each model.

The difference between class probability functions (mixture of Gaussians) evaluated at a test point y (the unseen or test subject embedded in the manifold), can be written as the logarithm of their division. Normalizing the difference (logarithmic division) and rewriting this using a sigmoid (logistic) function we obtain the following scoring function:

$$S(y) = \left(\frac{2}{1 + \frac{P_A(y)}{P_B(y)}} \right) - 1. \quad (16)$$

Here $S(y)$ ranges from -1 to 1, and the sign represents the class while the absolute value indicates the class likelihood probability. The continuous nature of the proposed metric provides a richer variable that can be used to define "heat" maps in the manifold associated with a particular class, e.g. AD, CN, sMCI or pMCI. This could be used to define high certainty regions in the manifold where predictions can be made with a high degree of confidence. Additionally, the "heat" maps provide a color-coded visualization tool of a patient's current "state" for clinicians. Restricting classification/prediction to high confidence areas can be used for patient enrollment in clinical trials, e.g. it might be of particular interest to find subjects that with a certain (high) degree of confidence will convert from MCI to AD in a certain amount of time.

Results

Using sparse regression (elastic net) as described in the Elastic net section, we obtain a probabilistic mask of the relevance of each variable or voxel in the image. This mask relates the importance of each voxel in a regression that models the MMSE score. MMSE scores were used instead of the disease label since they offer a more continuous representation of disease progression. Three orthogonal 2D views of the probabilistic mask obtained from the elastic net algorithm can be seen in Fig. 1. In this image it can be observed that the variables with higher probability cluster around the hippocampus, which is a well known marker of AD. Thresholding this mask at a certain probability of a voxel being "picked" by the sparse regression, produces an ROI. In our experiments we found empirically that a 10% threshold produces the best results, which yields a mask of 1331 voxels. This parameter can be also tuned using cross validation.

Using the obtained mask to define the ROI in unseen labeled and unlabeled images we learn a low-dimensional representation of these ROIs using Laplacian eigenmaps (see Laplacian eigenmaps section) in a similar way as Belkin and Niyogi (2004), with cross-correlation as a similarity metric between subjects' ROIs. Then finally, the population distribution modeling was carried out directly on the learned subspace using the methodology described in the Population distribution modeling section. An overview diagram of the methods main steps is shown in Fig. 2.

In order to measure the different aspects of the proposed methodology, different experiments were designed. Although the proposed MRI Disease State Score (MRI-DSS) metric allows for a continuous disease modeling, experiments based on classification tasks are presented in order to allow easy comparison to previous work. In the following sections we report the classification performance for the clinically relevant class separations of the ADNI and ADNI-GO datasets. Additionally, the value of performing variable selection as well as manifold learning is illustrated by showing an overall improved classification accuracy. We also show accurate MMSE score prediction using the proposed MRI-DSS.

Table 4
Classification results on the manifold built using the learned ROI (Learned mask SVM and Learned mask MRI-DSS) and on a manifold built from the hippocampal mask used in Wolz et al. (2011) (Hippocampal mask SVM and Hippocampal mask MRI-DSS). In all cases results for classifiers A and B are presented separated by "/". Best results shown in bold numbers.

	AD vs. CN			pMCI vs. sMCI		
	ACC	SEN	SPE	ACC	SEN	SPE
Learned mask SVM	84/**86**	84/86	**85**/**85**	69/**71**	**77**/75	60/**67**
Learned mask MRI-DSS	81/81	80/83	82/82	66/67	72/71	59/64
Hippocampal mask SVM	81/81	79/83	82/79	60/66	67/70	53/61
Hippocampal mask MRI-DSS	76/78	82/87	71/69	58/61	53/60	63/62
No manifold learning SVM	84/75	**91**/76	77/73	61/62	61/69	61/55
Elastic net	81/81	85/85	76/77	63/65	64/64	62/64
Elastic net + β stb. sel.	81/81	86/86	76/77	60/65	64/66	58/64

	pMCI vs. CN		
	ACC	SEN	SPE
Learned mask SVM	**82**/81	81/**86**	**83**/76
Learned mask MRI-DSS	77/78	72/85	82/70
Hippocampal mask SVM	76/75	75/71	77/79
Hippocampal mask MRI-DSS	70/69	63/55	77/82
No manifold learning SVM	68/66	77/77	59/55
Elastic net	79/74	81/76	76/71
Elastic net + β stb. sel.	79/74	82/77	76/71

Table 3
Framework parameter setting summary. A parameter setting of automatic means that it was set without any user input, explored means that the best result in a range of settings is reported and set means it was empirically chosen.

Stage	Variable	Parameter	Value	Setting	Data
Variable selection	λ	L_1 weight	1000	Set	315 ADNI subjects
	τ	Threshold	10%	Set	
Manifold learning	d	Dimensionality	1–50	Explored	511 ADNI subjects
	k	Nearest neighbors	20	Set	
Population modeling	G	Kernel bandwidth	–	Automatic	
SVM	C	Soft margin	1	Set	
	κ	Kernel	Linear	Set	

Table 5
p-Values from McNemar's χ^2 tests between classifier paradigms A and B.

	AD vs. CN	pMCI vs. sMCI	pMCI vs. CN
Learned mask SVM	$p < 0.001$	$p = 0.051$	$p = 0.172$
Learned mask MRI-DSS	$p = 0.488$	$p = 0.015$	$p < 0.001$
Hippocampal mask SVM	$p = 0.852$	$p < 0.001$	$p = 0.474$
Hippocampal mask MRI-DSS	$p = 0.557$	$p = 0.059$	$p < 0.001$
No manifold learning SVM	$p < 0.001$	$p = 0.051$	$p = 0.027$
Elastic net	$p = 0.233$	$p = 0.001$	$p = 0.011$
Elastic net + β stb. sel.	$p = 0.915$	$p < 0.001$	$p < 0.001$

ADNI classification

Classification tasks were carried out in the manifold subspace learned from the selected variables according to the Relevant variable selection section and the hippocampal mask from Wolz et al. (2012), as well as outside the manifold to show the added value of this step. Additionally, the regression model obtained from the variable selection step was also used for classification. Three types of classifiers were explored in this work, an SVM (Cortes and Vapnik, 1995), a threshold on the proposed probabilistic distribution (MRI-DSS) and a thresholded elastic net regression.

SVM uses training data to find an optimal separating hyperplane between two subject classes in an n-dimensional feature space. Using this n-dimensional hyperplane, test subjects are classified according to their relative position in the manifold. In this work we used SVM with a linear kernel function, soft-margin constant $C = 1$ and quadratic programing optimization. Fine tuning the soft-margin constant provides slightly better results, however results are generally robust for a very large range of values ($1e^{-6} < C < 1e^{5}$). Fig. 3 shows a boxplot of a grid search of C for sMCI vs. pMCI classification using a type B classifier. Here, instances are an average of the accuracies obtained in manifolds with 1–50 dimensions, while the red line, box, whiskers and crosses represent the median, the 75th percentile, the extremes and the outliers, respectively, of 100 runs were done for every value of C. Preliminary experiments showed that using non-linear kernels provided little to no improvement, while adding more tuning parameters to the framework. Results for type A classifiers as well as other classification tasks show similar robustness to the setting of C.

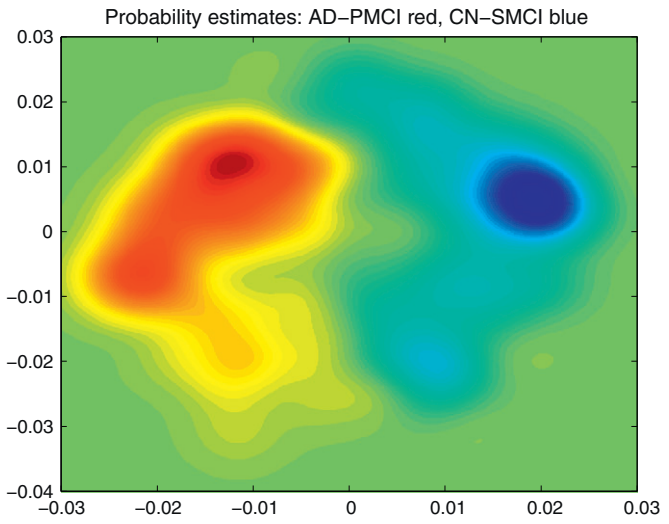

Fig. 5. 2D view of manifold's probability estimations for the ADNI dataset (best seen in color).

The probabilistic distribution threshold was obtained by combining the estimated distribution from both classes and normalizing values to form a sigmoid shaped MRI-DSS function. Values range from −1 to 1 and the absolute value indicates class likelihood probability. Thresholding this scoring function, Eq. (15) at zero allows us to binarize the scores in order to obtain a classification.

We used both methods to measure classification accuracy (ACC), sensitivity (SEN) and specificity (SPE). These metrics are defined using the true positive (TP), true negative (TN), false positive (FP) and false negative (FN) rates. These in turn represent the correctly identified,

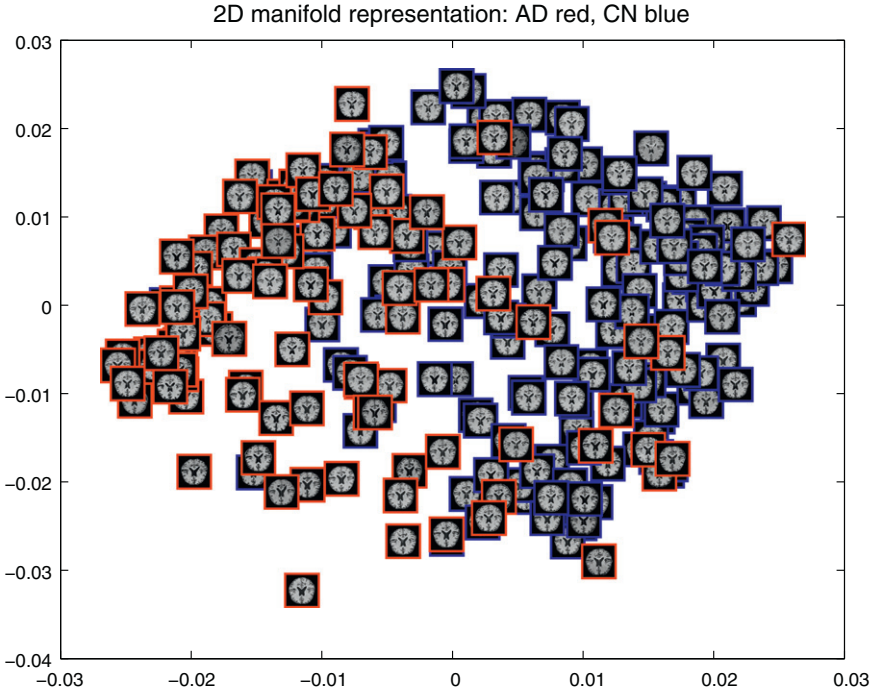

Fig. 4. 2D view of estimated manifold the ADNI dataset (best seen in color).

Table 6
Classification results using selected variable mask from ADNI to learn manifold of ADNI-GO at baseline.

	eMCI vs. CN			
	ACC	SEN	SPE	p-Value (MANOVA/Cramer)
SVM	61	76	46	0.0003/0.001
MRI-DSS	61	66	56	0.0002/0.004

correctly rejected, incorrectly identified and incorrectly rejected instances, respectively. The ACC, SEN and SPE are formulated as:

$$\begin{aligned} \mathrm{ACC} &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}, \\ \mathrm{SEN} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \\ \mathrm{SPE} &= \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}. \end{aligned} \tag{17}$$

The results for the comparisons AD vs. CN, pMCI vs. sMCI and pMCI vs. CN, using the ADNI dataset (see Table 1) are presented in Table 4. Results for the eMCI vs. CN, using the ADNI-GO dataset (see Table 2) are shown in Table 6. In all experiments we used a leave 10% out testing strategy, and the results presented reflect the average over 1000 runs.

Considering a disease progression that follows a trajectory from CN to MCI to AD, and assuming that sMCI subjects tend to be more like CN like, while at the same time pMCI subjects tend to be more AD like, grouping them together in order to boost the classifier training data can be justified. By doing so we can train a classifier or probability distribution as a class that includes CN and sMCI, and another group that includes pMCI and AD. From this point these two classification paradigms will be referred to as classifier A and classifier B, respectively.

The parameters for the Laplacian eigenmaps algorithm were set empirically based on previous experience (Guerrero et al., 2011; Wolz et al., 2011, 2012). The similarity measure of choice was cross-correlation, since these provided more robust results in previous experiments, with the added benefit that we do not need to convert distances to similarities, hence, avoiding the use of the heat kernel and eliminating the choice of the associated bandwidth parameter. The nearest neighbors used to build the similarity graph were set to 20, although similar results are obtained for values between 10 and 25. Finally, for every case the dimensionality of the manifold was explored from 1 to 50 dimensions, the best values are reported. We also found that the results are robust against the choice of these parameters, with stable SVM classification results in manifold dimensionalities from 10 to 30, and for the case of MRI-DSS the best performing dimensionalities are consistently in the 1–10 range, this is due to the relatively low number of samples used to learn the higher dimension probabilistic models. Table 3 gives a summary of the parameters involved in the proposed framework along with their setting as well as the data used to set them.

Table 4 shows classification results on the manifold learned from the selected variables, which in the table are referred to as learned mask SVM and learned mask MRI-DSS. It can be seen that the results are comparable to the state-of-the-art. In general, results indicate that SVM performs on average better than MRI-DSS, however, it must be noted that the proposed metric tries to model a more complex variable (the whole population distribution) with the added benefit of providing good visualization capabilities of the results that can potentially be used to show progression from a "low-risk" zone to a "mild" or "high-risk" zone in the manifold. In order to assess the value of doing variable selection, as opposed to using a predefined structural mask, we repeated the experiments using the same structural mask used in Wolz et al. (2011), which defines a ROI of around 30,000 voxels around the hippocampus. The results of classification on the manifold learned based on this ROI and in the same manner as before are also presented in Table 4 (Hippocampal mask SVM and Hippocampal mask MRI-DSS). It can be seen that in every case using the learned mask provides more accurate results.

Table 7
Classification results using hippocampal mask to learn manifold of ADNI-GO at baseline.

	eMCI vs. CN			
	ACC	SEN	SPE	p-Value (MANOVA/Cramer)
SVM	57	59	54	0.0041/0.004
MRI-DSS	57	54	59	0.0019/<0.0001

Table 8
Classification using selected variable mask from ADNI to learn manifold of ADNI-GO at baseline.

	eMCI vs. CN			
	ACC	SEN	SPE	p-value (MANOVA/Cramer)
SVM	65	61	69	<0.0001/<0.0001
MRI-DSS	61	50	72	<0.0001/<0.0001

Another important part of the proposed methodology is the use of variable selection and manifold learning. We have evaluated the importance of this by performing a comparison of classifiers trained and tested on the subjects without manifold learning as well as using the sparse regression model used in the variable selection step to directly classify the data. The regression coefficients $\boldsymbol{\beta}$ were estimated with and without the re-sampling technique described in the Training re-sampling section. Results for this experiments are presented in Table 4. It can be observed that for every case learning classifiers on the manifold space outperform classifiers learned in their original space as well as using the variable selection regression model for classification. Note that only a classifier like SVM that is able to deal with relatively high-dimensional data can be used for comparison.

Furthermore, we can notice that classifier paradigm B on average produces a slightly higher accuracy than paradigm A in the AD vs. CN and pMCI vs. sMCI classification tasks. We believe that this is due to the added training samples, which should provide a more robust classifier. However, this trend seems to reverse for the pMCI vs. CN classification task. Following the recommendations given by Salzberg and Note (1997), statistical significance between classifier paradigms A and B was calculated using McNemar's χ^2 test. This revealed mixed results on the statistical significance between classifier paradigms (see Table 5). We also note that the testing data belongs only to the specified groups, regardless of classifier paradigm.

Fig. 4 shows a 2D visualization of the subjects using the best two eigenvectors of the learned subspace manifold, based on the learned ROI, and Fig. 5 shows the probability distribution mixture of classes.

ADNI-GO classification

Experiments were carried out using the learned ROI's from the ADNI database to classify the ADNI-GO database performing manifold learning and population modeling, see Table 2. Preprocessing was carried out in the same fashion as with ADNI. The results are presented in Tables 6, 7, and 8. Table 6 presents the results obtained using the same variable mask as for the ADNI experiments. The p-values indicate the probability that the manifold coordinates from both classes belong to the same distribution. Two permutation tests were used to assess this, MANOVA and the Cramer test (Baringhaus and Franz, 2004). The former assumes a normal distribution of the data, while the latter does not make such an assumption. As can be seen in Fig. 6, the normality assumption of the data distribution does not necessarily hold, nonetheless, results from both tests are presented. Table 7 presents the results of using the hippocampal mask used in Wolz et al. (2011), again to show the added value of the variable selection step. The results shown in Table 8 use a ROI obtained from the variable selection procedure with a threshold of 1% on the probabilistic soft mask. This thresholding yielded 17,428 voxels, which includes more varied areas

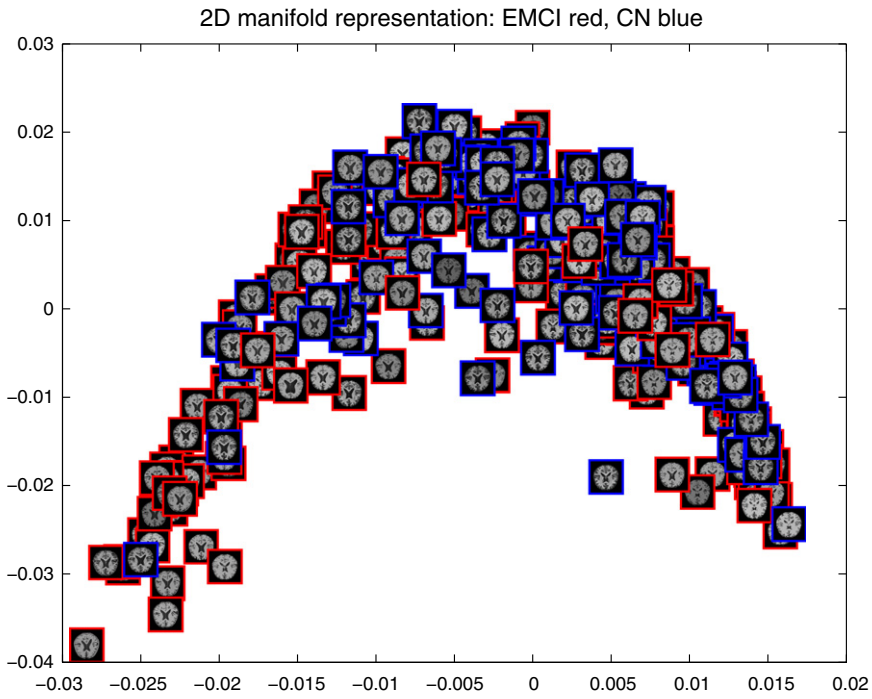

Fig. 6. 2D view of estimated manifold for the ADNI-GO dataset (best seen in color).

of the brain other than the hippocampus and its vicinity. The improvement in the results is hypothesized to be due to the subtle contributions of areas of the brain other than the hippocampus. Fig. 6 shows the population in the manifold and Fig. 7 shows the class probability distributions. As expected, we see that the classes' probability distributions pose more challenging questions, hence, accounting for the relatively low classification accuracy.

MMSE prediction

An experiment was carried out to estimate MMSE scores from the manifold. Using a linear regression model built from the MRI-DSS obtained from the learned low-dimensional population distributions yielded an average error of 1.5 points. From Table 1 we can see that in ADNI class mean MMSE values are separated by 2.2 points for AD-MCI, 3.72 points for MCI-CN, and a smaller separation of 0.65 points exists between pMCI-sMCI mean MMSE values. Furthermore, in ADNI-GO (Table 2) a separation between CN-eMCI mean MMSE scores of 0.7 points can be noted. When originally proposed, the MMSE (Folstein et al., 1975) was shown to have test–re-test mean variation of 1.1 points when the same tester did both examinations within a 24 hour period on the same patients, while a slightly higher mean variation of 1.3 was observed when one tester did the first test and another tester did the second test. The prediction accuracy of the presented method is comparable to the variability of the test itself.

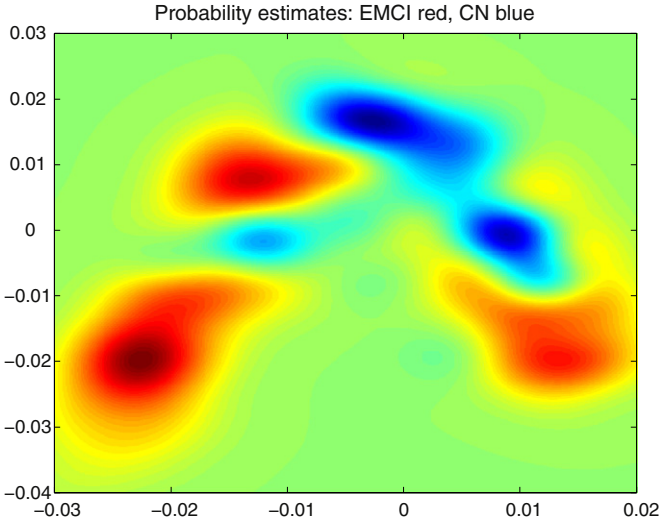

Fig. 7. 2D view of manifold's probability estimations for the ADNI-GO dataset (best seen in color).

Discussion

Recently the task of predicting conversion to AD has received a lot of attention, particularly for subjects in the MCI group. Several approaches that seek to classify the data in order to carry out this prediction task have been proposed in the literature. The proposed method learns a ROI using elastic net regression with a richer response variable (MMSE scores) rather than what could be considered over-simplistic class labels that do not fully explain the disease stage. In a database such as ADNI the MMSE scores should be highly correlated with the class labels since MMSE scores form part of the inclusion and diagnostic criteria of the study. Another important point to note is that the proposed MRI-DSS metric parameterizes the class likelihood as a continuous score. This could potentially be used to define areas of high or low diagnostic certainty. An added benefit of the proposed MRI-DSS is the intuitive visualization of the probability maps in lower dimensional spaces (1–3 dimensions). Classification results reported for other methods are shown in Table 9. A direct comparison between methods is difficult due to differences in the datasets, preprocessing steps, validation techniques, etc. However, some observations can be made about the advantages and disadvantages of the different methods. Here we focus the discussion studies that report results on the sMCI/pMCI classification task, as this is arguably the most clinically relevant one.

Cho et al. (2012) classified subjects based on cortical thickness features using the same samples as Cuingnet et al. (2011), obtaining an accuracy of 71% but with relatively low sensitivity of 63%. Chupin et al. (2009) obtain a classification accuracy of 64% use hippocampal volumes as features, they also have a low sensitivity of 60%. Coupé et al. (2012)

Table 9
Previous work results on classification of sMCI vs. pMCI.

Article	Feature(s)	Method	Conversion period
Cho et al. (2012)	Cortex	Cortical thickness	0–18 months
Chupin et al. (2009)	Hip. and amygdalae	Atlas based	0–18 months
Coupé et al. (2012)	Hip. and entorhinal cortex	Atlas based (LNOCV)	0–48 months
	–	Atlas based (LOOCV)	–
Cuingnet et al. (2011)	Hippocampus	Atlas based	0–18 months
	Whole brain	VBM (GM)	–
	Cortex	Cortical thickness	–
Davatzikos et al. (2011)	Whole brain	VBM	0–36 months
Eskildsen et al. (2013)	Cortex	TBM, Cortical ROIs	0–48 months
Koikkalainen et al. (2011)	Whole brain	TBM, combination of classifiers	0–36 months
Misra et al. (2009)	Whole brain	VBM, ROIs	0–36 months
Querbes et al. (2009)	Cortex	Cortical thickness	0–24 months
Westman et al. (2011)	Cortical and subcortical	Thickness and volume	0–12 months
Wolz et al. (2011)	Hippocampus	Atlas based	0–48 months
	Whole brain	TBM	–
	Hip. and amygdalae ROI	Manifold learning	–
	Cortex	Cortical thickness	–
	Combination	Combination	–
Zhang et al. (2012)	Whole brain	Whole brain ROIs	0–24 months

Article	N (sMCI, pMCI)	ACC (SEN/SPE) %
Cho et al. (2012)	131, 72	71 (63/76)
Chupin et al. (2009)	134, 76	64 (60/65)
Coupé et al. (2012)	238, 167	74 (73/74)
	–	71 (70/71)
Cuingnet et al. (2011)	134, 76	67 (62/69)
	–	71 (57/78)
	–	70 (32/91)
Davatzikos et al. (2011)	170, 69	56 (95/38)
Eskildsen et al. (2013)	227, 161	68 (68/69)
Koikkalainen et al. (2011)	215, 164	72 (77/71)
Misra et al. (2009)	76, 27	82 (—/—)
Querbes et al. (2009)	50, 72	73 (75/69)
Westman et al. (2011)	256, 62	59 (74/56)
Wolz et al. (2011)	238, 167	65 (63/67)
	–	64 (65/62)
	–	65 (64/66)
	–	56 (63/45)
	–	68 (67/69)
Zhang et al. (2012)	50, 38	78 (79/78)

use patch based segmentation to segment the hippocampus while at the same time scoring them as AD-like or CN-like, to our knowledge presented the best results using all the available images from the ADNI cohort, with a reported accuracy of 74%, although they have a more complex preprocessing pipeline. Cuingnet et al. (2011) evaluated various structural methods, for which the obtained accuracies range from 57–71% with relatively low sensitivities. Davatzikos et al. (2011) used voxel-based morphometry (VBM) to classify subjects, they achieved an accuracy of 56% with a high sensitivity of 95% but at the cost of a very low specificity of 38%. Eskildsen et al. (2013) used tensor based morphometry (TBM) along with cortical ROI, also, subjects with similar time to conversion were pooled together and tested independently achieving high accuracies (~79%), however when the features selected were used on the whole dataset the accuracy obtained was of 68%. Koikkalainen et al. (2011) used TBM with a combination of classifiers to achieve an accuracy of 72%, however it is suggested in Coupé et al. (2012) and Eskildsen et al. (2013) that this high accuracy might be biased, since the ROI used is obtained using the training and testing dataset. Misra et al. (2009) use VBM to find discriminative ROI in the images, the highest accuracy reported is of 82%, however the low number of subjects included in the study makes it hard to compare to other methods. Querbes et al. (2009) used cortical thickness features within ROI to achieve an accuracy of 73%, however, as in Koikkalainen et al. (2011), the ROIs were learned from both training and testing dataset. Westman et al. (2011) used predefined cortical ROI and subcortical structure volumes to predict conversion, achieving and accuracy of 59%. Wolz et al. (2011) used a combination of methods and features to obtain precision accuracies between 64 and 68%. Zhang et al. (2012) use longitudinal data to learn ROI within the whole brain, the highest accuracy reported is of 78%, however, as in Misra et al. (2009) and Querbes et al. (2009), the small amount of subjects used in this study makes it hard to compare with other methods.

As it can be seen, the proposed method offers comparable classification and prediction results to other state-of-the-art techniques. One of the main strengths of the proposed method is the ability to model the entire population. This provides good visualization properties in the learned manifold, which can also be used to define “hot” spots where there is a high degree of confidence in the classification/prediction made. However, as with any other method it has some disadvantages, mainly the fact that the manifold and distribution have to be relearned every time a new subject is added to the cohort. A potential pitfall of the proposed methodology is that it requires the tuning of several parameters, some of which can be automatically found or set according to empirical knowledge. However, the dimensionality parameter in the work presented here was explored within a given range, potentially limiting the generalizability of the reported finds.

There are some unexplored avenues of research on this paper. In this work we use a 10 mm free form deformation (FFD) (Rueckert et al., 1999) grid, in order to remove coarse non-linear inter-subject anatomical variations, while aligning smaller structures. There is no guarantee that the selected level of deformation is optimal, or furthermore, there is no guarantee that there exists any optimal one. Future work could

include a multilevel variable selection step, using sparse regression to select 4D variables, where 3 dimensions would represent x y z voxel coordinates and the fourth dimension represents the level of deformation (i.e. from 20 to 5 mm FFD control point spacing). Another avenue to explore is the incorporation of longitudinal features, variable selection could be done also on longitudinal images in a similar fashion, and using longitudinal images in the fourth dimension.

Acknowledgment

This project is partially funded by CONACyT (grant 306758), SEP-DGRI, the Rabin Enzra trust and the 7th Framework Programme by the European Commission (http://cordis.europa.eu/ist/; EU-Grant-224328-PredictAD; Name: From Patient Data to Personalised Healthcare in Alzheimer's Disease).

References

Aisen, P.S., Petersen, R.C., Donohue, M.C., Gamst, A., Raman, R., Thomas, R.G., Walter, S., Trojanowski, J.Q., Shaw, L.M., Beckett, L.A., Clifford Jr., R.J., Jagust, W., Toga, A.W., Saykin, A.J., Morris, J.C., Green, R.C., Weiner, M.W., 2010. Clinical core of the Alzheimer's Disease Neuroimaging Initiative: progress and plans. Alzheimers Dement. 6 (3), 239–246.

Aljabar, P., Heckemann, R., Hammers, A., Hajnal, J., Rueckert, D., 2009. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. NeuroImage 46 (3), 726–738.

Anderson, W.N., Morley, T.D., 1985. Eigenvalues of the Laplacian of a graph. Linear Multilinear Algebra 18, 141–145.

Baringhaus, L., Franz, C., 2004. On a new multivariate two-sample test. J. Multivar. Anal. 88 (1), 190–206.

Belkin, M., Niyogi, P., 2002. Laplacian eigenmaps and spectral techniques for embedding and clustering. Advances in Neural Information Processing Systems, vol. 14, pp. 585–591.

Belkin, M., Niyogi, P., 2004. Semi-supervised learning on Riemannian manifolds. Mach. Learn. 56 (1–3), 209–239.

Brookmeyer, R., Johnson, E., Ziegler-Graham, K., Arrighi, H.M., 2007. Forecasting the global burden of Alzheimers disease. Alzheimers Dement. 3 (3), 186–191.

Cho, Y., Seong, J.K., Jeong, Y., Shin, S.Y., 2012. Individual subject classification for Alzheimer's disease based on incremental learning using a spatial frequency representation of cortical thickness data. NeuroImage 59 (3), 2217–2230.

Chupin, M., Gérardin, E., Cuingnet, R., Boutet, C., Lemieux, L., Lehéricy, S., Benali, H., Garnero, L., Colliot, O., 2009. Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI. Hippocampus 19 (6), 579–587.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning, pp. 273–297.

Coupé, P., Eskildsen, S.F., Manjn, J.V., Fonov, V.S., Collins, D.L., 2012. Simultaneous segmentation and grading of anatomical structures for patient's classification: application to Alzheimer's disease. NeuroImage 59 (4), 3736–3747.

Cox, T., Cox, M., 1994. Multidimensional Scaling. Chapman & Hall, London.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehricy, S., Habert, M.O., Chupin, M., Benali, H., Colliot, O., 2011. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. NeuroImage 56 (2), 766–781.

Davatzikos, C., Bhatt, P., Shaw, L.M., Batmanghelich, K.N., Trojanowski, J.Q., 2011. Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. Neurobiol. Aging 32 (12), 2322.e19–2322.e27.

Duong, T., Hazelton, M., 2003. Plug-in bandwidth matrices for bivariate kernel density estimation. J. Nonparametric Stat. 15 (1), 17–30.

Eskildsen, S.F., Coupé, P., García-Lorenzo, D., Fonov, V., Pruessner, J.C., Collins, D.L., 2013. Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning. NeuroImage 65, 511–521.

Fan, Y., Shen, D., Gur, R., Gur, R., Davatzikos, C., 2007. COMPARE: classification of morphological patterns using adaptive regional elements. IEEE Trans. Med. Imaging 26 (1), 93–105.

Fan, Y., Batmanghelich, N., Clark, C.M., Davatzikos, C., 2008. Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. NeuroImage 39 (4), 1731–1743.

Folstein, M.F., Folstein, S.E., McHugh, P.R., 1975. "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. J. Psychiatr. Res. 12 (3), 189–198.

Frisoni, G.B., Fox, N.C., Jack, C.R., Scheltens, P., Thompson, P.M., 2010. The clinical use of structural MRI in Alzheimer disease. Nat. Rev. Neurol. 6, 67–77.

Gerardin, E., Chtelat, G., Chupin, M., Cuingnet, R., Desgranges, B., Kim, H.S., Niethammer, M., Dubois, B., Lehricy, S., Garnero, L., Eustache, F., Colliot, O., 2009. Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. NeuroImage 47 (4), 1476–1486.

Gerber, S., Tasdizen, T., Fletcher, P.T., Joshi, S., Whitaker, R., 2010. Manifold modeling for brain population analysis. Med. Image Anal. 14 (5), 643–653.

Guerrero, R., Wolz, R., Rueckert, D., 2011. Laplacian eigenmaps manifold learning for landmark localization in brain mr images. Medical Image Computing and Computer-Assisted Intervention (MICCAI). Lecture Notes in Computer Science, vol. 6892, pp. 566–573.

Hamm, J., Ye, D.H., Verma, R., Davatzikos, C., 2010. GRAM: a framework for Geodesic Registration on Anatomical Manifolds. Med. Image Anal. 14 (5), 633–642.

Hampel, H., Frank, R., Broich, K., Teipel, S.J., Katz, R.G., Hardy, J., Herholz, K., Bokde, A.L.W., Jessen, F., Hoessler, Y.C., Sanhai, W.R., Zetterberg, H., Woodcock, J., Blennow, K., 2010. Biomarkers for Alzheimer's disease: academic, industry and regulatory perspectives. Nat. Rev. Drug Discov. 9, 560–574.

Hastie, T., Stuetzle, W., 1989. Principal curves. J. Am. Stat. Assoc. 84, 502–516.

Jack, C.R., Knopman, D.S., Jagust, W.J., Shaw, L.M., Aisen, P.S., Weiner, M.W., Petersen, R.C., Trojanowski, J.Q., 2010. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. Lancet Neurol. 9 (1), 119–128.

Klein, S., Loog, M., van der Lijn, F., den Heijer, T., Hammers, A., de Bruijne, M., van der Lugt, A., Duin, R.P.W., Breteler, M.M.B., Niessen, W.J., 2010. Early diagnosis of dementia based on intersubject whole-brain dissimilarities. IEEE International Symposium on Biomedical Imaging (ISBI), pp. 249–252.

Koikkalainen, J., Ltjnen, J., Thurfjell, L., Rueckert, D., Waldemar, G., Soininen, H., 2011. Multi-template tensor-based morphometry: application to analysis of Alzheimer's disease. NeuroImage 56 (3), 1134–1144.

Kristan, M., Leonardis, A., Skocaj, D., 2011. Multivariate online kernel density estimation with Gaussian kernels. Pattern Recogn. 44 (10–11), 2630–2642.

Lerch, J.P., Pruessner, J., Zijdenbos, A.P., Collins, D.L., Teipel, S.J., Hampel, H., Evans, A.C., 2008. Automated cortical thickness measurements from MRI can accurately separate Alzheimer's patients from normal elderly controls. Neurobiol. Aging 29 (1), 23–30.

Leung, K., Barnes, J., Modat, M., Ridgway, G., Bartlett, J., Fox, N., Ourselin, S., 2011. Automated brain extraction using Multi-Atlas Propagation and Segmentation (MAPS). IEEE International Symposium on Biomedical Imaging (ISBI), pp. 2053–2056.

Misra, C., Fan, Y., Davatzikos, C., 2009. Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI. NeuroImage 44 (4), 1415–1422.

Nyl, L.G., Udupa, J.K., 1999. On standardizing the mr image intensity scale. Magn. Reson. Med. 42 (6), 1072–1081.

Petersen, R.C., Smith, G.E., Waring, S.C., Ivnik, R.J., Tangalos, E.G., Kokmen, E., 1999. Mild cognitive impairment: clinical characterization and outcome. Arch. Neurol. 56 (3), 303–308.

Pollard, D., 2002. A User's Guide to Measure Theoretic Probability. Cambridge University Press.

Querbes, O., Aubry, F., Pariente, J., Lotterie, J.A., Dmonet, J.F., Duret, V., Puel, M., Berry, I., Fort, J.C., Celsis, P., Initiative, T.A.D.N., 2009. Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve. Brain 132 (8), 2036–2047.

Ranginwala, N.A., Hynan, L.S., Weiner, M.F., White III, C.L., 2008. Clinical criteria for the diagnosis of Alzheimer disease: still good after all these years. Am. J. Geriatr. Psychiatry 16, 384–388.

Roweis, S.T., Saul, L.K., 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290, 2323–2326.

Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L.G., Leach, M.O., Hawkes, D.J., 1999. Nonrigid registration using free-form deformations: application to breast MR images. IEEE Trans. Med. Imaging 18 (8), 712–721.

Salzberg, S.L., Note, M., 1997. On comparing classifiers: pitfalls to avoid and a recommended approach. Data Min. Knowl. Disc. 1 (3), 317–328.

Scott, D., 1992. Multivariate Density Estimation: Theory, Practice, and Visualization. Wiley Series in Probability and StatisticsWiley (URL: http://books.google.co.uk/books?id=7crCUS_F2ocC).

Tenenbaum, J.B., Silva, V., Langford, J.C., 2000. A global geometric framework for nonlinear dimensionality reduction. Science 290 (5500), 2319–2323.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Methodol. 58 (1), 267–288.

Tikhonov, A.N., Arsenin, V.Y., 1977. Solutions of Ill-Posed Problems.

van der Maaten, L., Postma, E.O., van den Herik, H.J., 2009. Dimensionality reduction: a comparative review. Tech. Rep. TiCC-TR 2009-005Tilburg University.

Vemuri, P., Gunter, J.L., Senjem, M.L., Whitwell, J.L., Kantarci, K., Knopman, D.S., Boeve, B.F., Petersen, R.C., Jack, C.R., Jr., 2008. Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. NeuroImage 39 (3), 1186–1197.

Wand, M.P., Jones, M.C., 1994. Kernel Smoothing (Chapman & Hall/CRC Monographs on Statistics & Applied Probability), 1st edition. Chapman and Hall/CRC.

Westman, E., Simmons, A., Muehlboeck, J.S., Mecocci, P., Vellas, B., Tsolaki, M., Koszewska, I., Soininen, H., Weiner, M.W., Lovestone, S., Spenger, C., Wahlund, L.O., 2011. AddNeuroMed and ADNI: similar patterns of Alzheimer's atrophy and automated MRI classification accuracy in Europe and North America. NeuroImage 58 (3), 818–828.

Wolz, R., Aljabar, P., Hajnal, J.V., Hammers, A., Rueckert, D., 2010. LEAP: Learning Embeddings for Atlas Propagation. NeuroImage 49 (2), 1316–1325.

Wolz, R., Julkunen, V., Koikkalainen, J., Niskanen, E., Zhang, D.P., Rueckert, D., Soininen, H., Ltjnen, J., the Alzheimer's Disease Neuroimaging Initiative, 2011. Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease. PLoS ONE 6 (10), e25446.

Wolz, R., Aljabar, P., Hajnal, J.V., Ltjen, J., Rueckert, D., 2012. Nonlinear dimensionality reduction combining MR imaging with non-imaging information. Med. Image Anal. 16 (4), 819–830.

Zhang, D., Shen, D., the Alzheimer's Disease Neuroimaging Initiative, 2012. Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers. PLoS ONE 7 (3), e33182.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B 67, 301–320.