Group-constrained manifold learning: Application to AD risk assessment.

R. Guerrero^{a,1,*}, C. Ledig^a, A. Schmidt-Richberg^a, D. Rueckert^a, The Alzheimer's Disease Neuroimaging Initiative (ADNI)²

^a*Biomedical Image Analysis Group, Department of Computing,
Imperial College London, UK*

Abstract

The exploration of high-dimensional medical imaging datasets often requires the use of dimensionality reduction techniques. Generally, individual instances in longitudinal studies are not independent as several measurements of the same subjects, acquired at several time points, are included. Most non-linear dimensionality reduction techniques, that rely on defining a neighborhood graph to uncover local geometry, neglect potentially valuable grouping information and thus can lead to a poor estimation of the underlying manifold. Incorporating a-priori grouping knowledge of the data can lead to an improved low-dimensional representation. In this work, an approach to leverage such a-priori information is proposed, with the main interest in generating longitudinal biomarkers to assess a subject's risk of developing Alzheimers disease (AD), and hence estimate time-to-conversion (TTC) and the mini mental state examination (MMSE) of subjects that develop AD. This is achieved via a novel formulation of the Laplacian Eigenmaps and Isomap manifold learning algorithms, that allows the incorporation of group-constraints. The potential of the proposed formulation was demonstrated on a swiss-roll toy example that clearly illustrates the problems faced by tradi-

*Corresponding author: Ricardo Guerrero

Email address: `reg09@imperial.ac.uk` (R. Guerrero)

¹This project was partially funded by the Innovate UK (formerly Technology Strategy Board - TSB).

²Data used in the preparation of this article were obtained from the ADNI database (<http://adni.loni.usc.edu>). 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.

tional techniques when additional structure is observed in the data. A real data experiment showed that the proposed constrained methods yielded improvements of $\sim 16\%$ and $\sim 6\%$ over the original unconstrained methods in k NN estimation of the TTC and MMSE on the ADNI database.

Keywords: Alzheimer’s disease, manifold learning, longitudinal, neurodegeneration.

1. Introduction

Non-linear dimensionality reduction techniques, such as Laplacian Eigenmaps [1] or isometric mapping (Isomap) [2], have become an important step in the exploration of high-dimensional medical imaging datasets. Individual measurements in longitudinal datasets are not necessarily independent from each other, as they exhibit a certain degree of structure and correlation within them. For example, in longitudinal studies multiple images of the same subject are acquired at several time points, and therefore are very similar to each other and thus highly correlated. Most classical non-linear dimensionality reduction techniques rely on finding similarities in the data to uncover the local manifold geometry. However, these approaches neglect potentially valuable grouping (e.g. by subject) information. Treating images that are inherently correlated as independent can lead to a poor estimation of the underlying manifold. It has been previously shown that exploiting a-priori knowledge of the underlying data grouping can lead to improved low-dimensional embeddings. In spatio-temporal Isomap [3] the original weights in the graph of local neighbors are empirically altered to emphasize similarity between temporally related points. Back constrained Gaussian Process latent variable models [4] include temporal coherence constraints to ensure the smoothness of the mapping between spaces. In temporal Laplacian Eigenmaps [5] the temporal relationships are modeled as part of the objective function. However, these techniques focus on obtaining temporal consistency within time series, where it can generally be assumed that each series contains complete realizations of a cyclical process, e.g. cardiac or breathing processes. Such techniques are not well suited for population studies where the acquired time series usually represents only a small time window of a longer process, e.g. the development and progression of Alzheimer’s disease (AD).

AD is the most common form of dementia, usually associated with the elderly population, that had a world wide prevalence of around 26.6 million

reported cases in 2006. Predictions suggest that this figure will increase to approximately 100 million by the year 2050 [6]. Preliminary evidence suggests that intervention to slowdown the progression of the disease is likely to be most effective in early stages of the disease [7]. 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 deceleration of the disease progression, may have deteriorated. Therefore, early diagnosis has the potential to have a significant impact 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.

Recently, the assessment and prediction of a subject’s current and future risk of developing AD has received a substantial amount of attention. Many of the proposed frameworks rely on modeling cognitive scores, e.g. the mini mental state examination (MMSE), as surrogate measures of disease progression [8, 9, 10, 11]. However, as noted by Jack et al. [12], although cognitive markers are especially relevant for the advanced stages of the disease, they are less sensitive at earlier stages. In this regard, and following the now generally accepted hypothetical model from Jack et al. [12], structural biomarkers such as those derived from magnetic resonance (MR) imaging offer higher sensitivities at earlier stages of the disease. Wolz et al. [13] explore the use of Laplacian Eigenmaps to derive longitudinal features for AD from MR images. In their work, two approaches were explored: independent embedding of the baseline and follow-up scans, and embedding of the difference images. Given that the images were considered independent during embedding and only two time points per subject were used, the underlying trajectory of each subject might be poorly estimated. Furthermore, it is not obvious how a manifold based on difference images could be learned for a time series with more than two time points. In [14], Zhang et al. proposed a longitudinal feature selection method which can select brain regions across multiple time points for MR and positron emission tomography (PET) images. A sparse linear regression model was trained for each time point that included an extra group regularization to group the weights corresponding to the same brain region across multiple time points, which allowed the selection of brain regions based on multiple time points. 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.

In the work presented here, group-constraints are introduced to the standard Laplacian Eigenmaps and Isomap that allow the incorporation of a-priori knowledge of image grouping. The proposed method is employed to calculate a low-dimensional embedding of T1-weighted MR images acquired from patients with memory complaints at multiple time points. It is shown that the incorporation of group-constraints allows an improved estimation of the time-to-conversion (TTC) of the subjects embedded in the manifold. The proposed approach also offers the capability to visualize a subjects transition within the manifold across time as the disease progresses. The derived features show a natural ordering across the visualized dimensions, with regards to TTC and disease progress.

The main objective of the methods proposed here is to derive a low-dimensional space, whose coordinates can be used as a longitudinally meaningful biomarker for AD risk assessment. There are three main steps involved in the proposed framework: (1) Feature selection for ROI learning using sparse regression (see Section 2.2), (2) feature “normalization” using local binary patterns (LBP) (see Section 2.3) and (3) group constrained manifold learning to derive a low-dimensional embedding of the data (see Section 2.4 and 2.5). Figure 1 shows a diagram illustrating these main steps, where steps (2) and (3) use a different set of images than step (1) in order to avoid a “double dipping” of the data and separate feature selection from the rest of the analysis. A description of the data used and details into the framework proposed here are given in Section 2, experimental results are provided in Section 3, and finally Section 4 offers a discussion on the implications, strengths and weaknesses of the proposed framework.

Figure 1: Illustration of the proposed framework.

2. Materials and Methods

2.1. Data

Data used in the work presented here was obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database. 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 non-profit organizations, as a large-scale multi-site study public-private partnership. The primary goal of ADNI has been to test whether serial MR imaging, 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. 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. To date, ADNI in its three studies (ADNI-1, -GO and -2) has recruited over 1500 adults, aged between 55 and 90 years, to participate in the study. Participants consist of cognitive normal subjects (CN), subjects with significant memory concerns (SMC), MCI, early MCI (eMCI), and early AD.

Here, a validation data set consisting of subjects enrolled in either ADNI-1, ADNI-GO or ADNI-2 (as of 19/02/2015) was considered. Of particular interest in this work is the subset of subjects that have at least four different time point samples and that convert to AD at some point during the study. In total, there were 5679 images belonging to 1063 subjects that had at least four time points. Out of these, 216 subjects (1309 samples) converted to AD (including 10 subjects that start as CN) during the study. Evaluations will focus on the subset 216 of AD converters, while manifold learning will take into account the set of 1063 subjects. As conversion to AD of these subjects could have occurred at any point between the last visit before being diagnosed as AD and the first visit after conversion, TTC was approximated as the mid point between these visits. Additionally, only subjects that had MMSE scores at every time point were considered. A separate set of 292 subjects (called training set) from ADNI-1 were used for feature selection to learn a probabilistic feature relevance mask. Here, the relevance of features in MNI space and from different alignment level of the image to the MNI

template is encoded. The mask was thresholded at $\tau = 0.1$ yielding a region of interest (ROI) with a size of 660, 922, 856, 580 and 190 voxels from the linear, 20, 10, 5 and 2.5 mm deformation levels, respectively. See Section 2.2 for further details.

2.1.1. Image preprocessing

All images used in this work were brain extracted using pyramidal intracranial masking [15]. Considering that images need to be aligned with each other for meaningful comparison, the question of which image alignment level is best needs to be addressed. Disease specific characteristics might manifest at different levels of image alignment, i.e. affinely aligned images of patients with AD might reveal discriminative information in the ventricles, while non-rigidly aligned images might reveal discriminative information of smaller structures such as the hippocampus. Structural differences can therefore be encoded in the registration deformation fields or in intensity differences between subjects [16, 17]. To this end, all images were registered to the MNI152 template using free-form-deformations [18] with different control point spacings to allow for different detail levels of deformation to the template: affine, 20mm, 10mm, 5mm and 2.5mm.

There are two types of registration (alignment) being considered in the work presented here: Longitudinal and cross-sectional. Longitudinal (intra-subject) registrations between baseline and follow-up images were performed using affine registration. This means that pathological changes such as atrophy will be encoded as intensity differences between baseline and follow-up images. Cross-sectional (inter-subject) registrations between baseline images and the template were performed using non-rigid free-form deformations (FFD) registration with different control point spacing and therefore different level of detail. The purpose of cross-sectional registration is to eliminate inter-subject variability. The aim of feature selection is thus to choose the optimal level of deformation that best compensates for inter-subject variability.

2.2. Multilevel feature selection

High-dimensional data such as 3D brain MR images contain a large amount of variables (voxels) that are not equally (or at all) responsive to changes in anatomy due to the presence and development of pathology. Elastic net [19] performs feature selection, while also encouraging the grouping of correlated features and can select any number of features. This is achieved by

adding L_1 and L_2 regularization terms, each weighted by parameters $\lambda_1 > 0$ and $\lambda_2 > 0$, respectively, to ordinary least square regression,

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \|\mathbf{I} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 \right\}. \quad (1)$$

The L_1 and L_2 terms encourage sparsity and grouping of features, respectively. In medical images, correlated features generally arise from neighboring regions, thus, feature selection can be viewed as learning a ROI [20]. When $\lambda_2 \rightarrow \infty$, the optimization problem has for each independent variable $\mathbf{x}_i$ a closed-form solution that can be written as:

$$\hat{\beta}_i = \left(|\mathbf{I}^T \mathbf{x}_i| - \frac{\lambda_1}{2} \right)_+ \operatorname{sign}(\mathbf{I}^T \mathbf{x}_i), \quad (2)$$

where $(\cdot)_+$ indicates the positive part, $\mathbf{I}$ is the dependent variable and $\mathbf{x}_i$ has zero mean and unit variance. Stability selection [21] is used additionally to introduce robustness against sampling errors. A probabilistic mask that indicates the likelihood of a feature to be selected is then obtained. Thresholding the probabilities at τ yields a binary ROI in the same way as in [20], where learning such data driven ROI was shown to improve performance over the use of an anatomically predefined ROI. In this work, independent features are MR image intensities, while the MMSE score acts as the dependent variable. In contrast to [20], each of the N subjects has associated R images that have been created by aligning (registering) the original scan to the MNI152 template at R different levels. As disease specific imaging characteristics might manifest at different alignment levels, this allows the selection of the most descriptive variables (voxels) in an unsupervised formulation [22, 23]. A matrix $\mathbf{X}$ is built, each column represents a spatial location in MNI space at a certain level of alignment to MNI space. Each row $\mathbf{x}_n$ in matrix $\mathbf{X}$ represents a subject at a specific time point $n \in N$ and is formed by concatenating R vectorized images at multiple levels $\mathbf{x}_n = [\mathbf{x}_{n1}, \dots, \mathbf{x}_{nr}, \dots, \mathbf{x}_{nR}]$. Elastic net regression then selects a subset of D features from $\mathbf{X}$, that correspond to column indices of $\mathbf{X}$. Finally, a mask that encodes selected features in MNI space and from different alignment levels of the image to the MNI template is obtained. Figure 2 illustrates the described multilevel feature selection framework.

Figure 2: Overview of the feature selection framework.

2.3. Local binary patterns

MR images acquired with different acquisition protocols have different intensity appearance. Thus, their intensities cannot be easily combined within a single framework without first mapping the image intensities into a common space. Many intensity normalization approaches (such as mean and variance or histogram matching) work very well for images in the absence of pathologies. However, in the presence of pathology the basic assumption made by these techniques does not necessarily hold (the mean, variance or quartiles can be shifted) and intensity normalization based on such techniques can potentially reduce disease related signal changes.

In this work we avoided the need for intensity normalization and local binary patterns (LBPs) [24] around the 26-connected neighborhood of each selected voxel (according to Section 2.2) were extracted and encoded as binary vectors instead. Figure 3 shows an example of a LBP extracted around the 8-connected neighborhood of a voxel. LBPs transform the MR intensity features to an augmented binary space in $\mathbb{R}^{26D}$ where the basic texture around each voxel is encoded as the sign of the difference between the selected voxel and its neighborhood as follows:

$$v_{ij} = s(x_{s_i}, x_{n_j}), \quad s(x_{s_i}, x_{n_j}) = \begin{cases} 1, & \text{if } x_{n_j} \geq x_{s_i} \\ 0, & \text{if } x_{n_j} < x_{s_i} \end{cases} \quad (3)$$

where $v_{ij} \in \mathbf{V}$ are the selected voxels (according to Section 2.2) in the augmented $\mathbb{R}^{26D}$ space, x_s represent the subset of selected voxels and x_n are those voxels in the 26-connected neighborhood of x_s , with $i = 1, \dots, N$ and $j = 1, \dots, 26$.

Thus, given images $I \in \mathcal{M}$ and $J \in \mathcal{N}$, where $\mathcal{M}, \mathcal{N} \in \mathbb{R}^D$ represent different intensity spaces, the objective of LBPs is to define a transformation T such that $T(I), T(J) \in \mathcal{L}$. Specifically with the definition of LBPs used in this work, we have $\mathcal{L} \in \mathbb{R}^{26D}$. In this new space $\mathcal{L}$, images that originally had distinct intensity characteristics can be combined, assuming that the original acquisition protocols are reasonably similar (e.g. both are T1-weighted).

Figure 3: 8-connected neighborhood LBP extraction.

2.4. Manifold learning

Manifold learning refers to a set of machine learning techniques that aim at finding a low-dimensional representation of high-dimensional data while preserving its intrinsic geometry. Deriving longitudinally meaningful biomarkers for AD based on manifold learning poses several challenges. Brain images reside in very high-dimensional space, although many dimensions (features or voxels) are not necessarily related to the disease being modeled. Relevant variable selection can be used to reduce the amount of features to those that are related to a specific variable. However, this usually results in a still relatively high dimensional space and hence the need for additional dimensionality reduction. In the following we will briefly introduce two popular techniques, Laplacian Eigenmaps and Isomap, based on which we illustrate the effects of the proposed group-constraints. A more comprehensive overview over manifold learning approaches can be found in [25].

Given a set of selected relevant features, the high-dimensional feature space $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N\} \in \mathbb{R}^{N \times D}$ is given by N vectors of length $D = 26D$ that represent LBP extracted around the most relevant voxels from a set of R images at different levels of alignment. Here $\mathbf{v}_i = \{v_1, v_2, \dots, v_D\}$ are LBPs extracted around the D (~ 3150) most relevant voxels from the i th

subject's set of R images and N refers to the total number of instances (all subjects and all time points here). The aim is to learn the underlying low-dimensional manifold in $\mathbb{R}^d$ ($d \ll \mathbf{D}$) that best represents the population through a mapping function f such that:

$$f : \mathbf{V} \rightarrow \mathbf{Y} , \quad \mathbf{y}_i = f(\mathbf{v}_i) \forall i . \quad (4)$$

2.4.1. Laplacian Eigenmaps

Laplacian Eigenmaps transforms data points to a low-dimensional space while preserving the local geometric properties of the manifold [1]. Local geometry can be determined using any similarity or distance metric. For example due to binary nature of the features used here, Hamming distance might be considered a good choice, however, experiments showed little difference in the use of different metrics, such as Hamming, correlation or sum of squared difference (SSD), to name a few. Therefore, local geometry was determined by converting pairwise SSD distances to a similarity matrix $\mathbf{W}$ using a Gaussian heat kernel as proposed in [1]. From $\mathbf{W}$, the k -neighborhoods of data points are used to construct a sparse neighborhood matrix $\mathbf{W}^*$. Laplacian eigenmaps seeks to place points $\mathbf{v}_i$ and $\mathbf{v}_j$ close together in $\mathbb{R}^d$ if they are close in the original $\mathbb{R}^{\mathbf{D}}$ space (large similarity $w_{i,j}^*$). This is achieved by means of minimizing the cost function:

$$\phi(\mathbf{Y}) = \operatorname{argmin} \sum_{i,j} \|\mathbf{y}_i - \mathbf{y}_j\|^2 w_{i,j}^* \quad (5)$$

under the constraint that $\mathbf{y}^T \mathbf{L} \mathbf{y} = 1$, which can be formulated as generalized eigenproblem $\mathbf{L} \nu = \mu \mathbf{M} \nu$, where $\mathbf{L} = \mathbf{M} - \mathbf{W}^*$ is the graph Laplacian and $\mathbf{M}$ is a degree matrix. Here ν and μ are the eigenvectors and eigenvalues, where the d eigenvectors corresponding to the smallest (non-zero) eigenvalues represent the new coordinate system.

2.4.2. Isomap

Isomap is a non-linear embedding technique that builds upon multidimensional scaling (MDS) [2]. This technique aims at preserving pairwise geodesic (or curvilinear) distances between data points, where geodesic distance refers to the distance between two points measured along the manifold. To resolve this problem, Isomap first calculates a similarity matrix $\mathbf{W}$ based

on distances (e.g. SSD) measured directly between the high-dimensional data points $\mathbf{V}$, and from $\mathbf{W}$ it determines a sparse k -neighborhood graph $\mathbf{W}^*$. Then, geodesic distances between all data points in $\mathbf{V}$ are estimated by finding all pairwise shortest paths in the neighborhood graph $\mathbf{W}^*$, thus forming a complete pairwise geodesic distance matrix $\mathbf{W}_G^*$. The low-dimensional representations $y_i \in \mathbf{Y}$ of the high-dimensional data points $v_i \in \mathbf{V}$ are computed by minimizing the following cost function:

$$E = \|\mathbf{W}_G^* - \mathbf{W}_Y\|_{L^2} . \quad (6)$$

Here $\mathbf{W}_Y$ refers to a matrix of pairwise distances in the low-dimensional space $\mathbf{Y}$. This equation can be globally minimized by setting the coordinates y_i to the top d eigenvectors of matrix $\mathbf{W}_G^*$.

2.5. Group-constrained manifold learning

Traditional manifold learning techniques neglect valuable grouping information in feature space. This can lead to a poor or suboptimal estimation of the underlying manifold. In the work presented here, this limitation is addressed by incorporating group constraints into the formulation of dimensionality reduction techniques. Specifically, two popular methods, the Laplacian Eigenmaps and Isomap algorithms, are modified and the benefit of the proposed constraints investigated. The proposed constraints can be equally integrated into many other manifold learning techniques.

The graph Laplacian $\mathbf{L}$ in Laplacian Eigenmaps acts as a design matrix that can be modified and adjusted, via changing the similarity matrix $\mathbf{W}$ to incorporate prior knowledge. In [1] $\mathbf{W}$ is sparsified using the data points' k -neighborhoods to form $\mathbf{W}^*$, which subsequently is used to calculate $\mathbf{L}$. In [22] $\mathbf{L}$ is calculated from a modified neighborhood matrix $\mathbf{W}^*$, where it is additionally used to drive the alignment of two different (albeit similar) manifolds. In a similar way, in Isomap [2] the similarity matrix $\mathbf{W}$ and the k -neighborhood graph $\mathbf{W}^*$ are used to estimate geodesic distances and can also be used as design matrices. Here, the focus lies in modeling longitudinal image data, specifically, the modeling of several image instances of a subject acquired at different time points. Such data is inherently grouped (subject wise) and hence additional (and potentially unwanted) structure is introduced. Since the standard formulation of neither Laplacian Eigenmaps nor Isomap allows the incorporation of grouping information, an additional grouping variable $\mathbf{g}$ is introduced to both of these frameworks. Data points

that belong to the same entity or group are associated with $\mathbf{g}$, e.g. g_i is categorical and all instances in the same category belong to the same group. In order to avoid nearest neighbors being biased towards any group g_i , a constraint is added so that at most one member per group $g_i \in \mathbf{g}$, for $i=1,2,\dots,N$, is admitted into the neighborhood (within group connections are not permitted). This allows only a single time point per subject to form part of the local neighborhood, such that the space where the manifold resides is sampled as evenly as possible. Henceforth, this modified sparse neighborhood matrix is referred to as $\mathbf{W}^{**}$. From Equations 5 and 6, the sparse weight matrix $\mathbf{W}^*$ is replaced with $\mathbf{W}^{**}$, in which every element is defined as:

$$w_{ij}^{**} = \begin{cases} w_{ij}, & \text{if } \mathbf{v}_j \in \mathcal{N}_i \\ 0, & \text{if } \mathbf{v}_j \notin \mathcal{N}_i \end{cases} \quad (7)$$

where $\mathcal{N}_i$ denotes the set of unique group $\mathbf{g}$ nearest neighbors of point $\mathbf{v}_i$ and w_{ij} the similarity between points $\mathbf{v}_i$ and $\mathbf{v}_j$.

Figure 4 illustrates the differences between the sparse unconstrained neighborhood matrix $\mathbf{W}^*$ as obtained in [1, 2] and the proposed group-constrained $\mathbf{W}^{**}$.

The neighborhoods shown in Figure 4 are chosen to be symmetric (undirected graph), and as such each point has at least k and at most $2k$ neighbors. Figure 5 gives an illustrative example of both connectivity graphs. It is easy to see that the local neighborhood of instances in a group would be poorly defined using only similarity or distance (dotted circles). However, by adding the proposed group-constraints to the local neighborhood definition (solid lines) the underlying geometry of the data is better sampled.

Figure 4: L_2 pairwise distance matrix $\mathbf{W}$, sparse unconstrained (symmetric) neighborhood $\mathbf{W}^*$ and the proposed group-constrained sparse (symmetric) neighborhood $\mathbf{W}^{**}$, for 7 subjects with 5-9 time points each.

Figure 5: Schematic illustration of a spatial connectivity graph, where dotted indicate groups of variables, solid lines show the original unconstrained ($\mathbf{W}^*$) and the proposed group-constrained ($\mathbf{W}^{**}$) neighborhoods, on the left and right respectively.

3. Results

The main objective of this work is to find low-dimensional representations of MR images in the presence of structure in the data, e.g. correlation between instances. To this end, a synthetic data experiment is proposed to illustrate the pitfalls of not taking into account the structure in the data. On the other hand, when working with real data low-dimensional representation should be useful to derive clinically relevant information and visualization. In the experiments detailed below some clinically relevant variables are estimated and visualized in such low-dimensional spaces. [Such estimated variables include the TTC, MMSE, functional activities questionnaire \(FAQ\), clinical dementia rating sum of squared boxes \(CDRSB\) and Alzheimer’s disease assessment scale cognitive subscale \(ADAS13\).](#) Estimated values are found by averaging values within the k -neighborhood of an unknown instance. The following Sections detail both the qualitative synthetic experiments as well as a real data quantitative estimation analysis.

3.1. Synthetic data: Swiss-roll

One of the problems with population analysis of longitudinal medical images is that, in general, longitudinal intra-subject variability is much smaller than inter-subject variability. Neighborhoods definition, used to find local structure in the data, generally relies on finding similarities (or distances) between data points in an unconstrained fashion. Neighborhoods found in this scenario would generally be a poor estimation of local geometry as they would be biased to include several instances of the same subject. Here, the classical 3D swiss-roll toy example is used to illustrate this effect. Figure 6 (a) shows a swiss-roll where the Y-axis is in a much larger scale than

the X- and Z-axes, with color indicating change along the desired variable (embedded in the 3D space). Figure 6 (b) reveals additional information about the data, where grouping of instances is indicated with one color per group. In this example the Y-axis simulates inter-subject variability (with the much larger scale indicating a relatively poor sampling rate) and the X- and Z-axes simulate intra-subject variability. Furthermore, the experiment is designed such that all intra-group distances are smaller than inter-group distances for illustrative purposes. The data consists of 100 different groups with 50 randomly distributed points along the X and Y surface of the manifold each (5000 data points). Additionally, the data is corrupted by noise on the X and Y axis, with the Z axis not corrupted as this axis represents different groups (noise generally does not make two groups more similar). Under this scenario, a qualitative experiment was carried out to compare standard Laplacian Eigenmaps (LE) [1], standard Isomap (IS) [2], the proposed group-constrained Laplacian eigenmaps (cLE) and group-constrained Isomap (cIS). Although only the aforementioned methods were compared, it is easy to see that the constraints proposed for cLE and cIS can be integrated to any method that defines local structure based on similarity of nearest neighbors. Figures 7 (a) and (b) show the results of using LE and cLE, respectively. For both cases the standard deviation of the heat kernel was set to $\sigma = 1$, while the number of nearest neighbors k for LE were set to 500 (with very similar results for values between 51-2500) and for the cLE was set to 5 (with very similar results for values between 3-50), both of which represent about 10% of all possible nearest neighbors. The relatively large amount of nearest neighbors used in LE stems from the fact that intra-group changes are much smaller than inter-group changes, hence in order to have a fully connected graph the number of nearest neighbors must be larger than 50 in order to ensure connection between groups. Similarly, Figures 7 (c) and (d) show the results of using IS and cIS, respectively. It is easy to see that the unconstrained algorithms fail to “unfold” the swiss-roll, and furthermore, it can also easily be deduced that in the proposed toy experiment, they will never be able to do so as inter-group distances are always larger than intra-group distances. One can be tempted to assume that a simple feature standardization might suffice to counteract the effect of the different scale in the Y-axis. This approach would certainly be applicable in the toy example illustrated here. However, in real data the effect of an observable larger inter-subject variability in comparison to temporal intra-subject variability persists even after instances have been registered to a common space (which

can be seen as feature standardization).

Figure 6: Swiss-roll colored according to (a) the desired variable and (b) data groups, where each color represents a different group.

3.2. Real data: Longitudinal AD biomarker learning

As mentioned in Section 2.3, LBPs around each voxel in the learned ROI (see Section 2.2) were extracted from images at their corresponding alignment levels and used to learn low-dimensional representations of them. Five different low-dimensional embeddings were calculated and compared: Principal component analysis (**PCA**) in order to demonstrate the non-linear nature of the manifold, **LE**, **cLE**, **IS** and **cIS**. Parameters for the non-linear low-dimensional embeddings were empirically set according to Table 1, while the parameter σ of the heat kernel in LE and cLE was set to 1. Additionally, in order to demonstrate the beneficial effect of normalizing features through LBPs, methods were also compared using mean and variance matching (MVM) normalization.

Table 1: Non-linear embedding neighborhood parameter k setting using local binary patterns (LBP) and mean and variance matching (MVM) normalization.

Features	Normalization	
	LBP	MVM
LE	400	200
cLE	250	50
IS	200	300
cIS	350	200

Figure 7: Swiss-roll mapping (best seen in color) using (a) LE, (b) cLE, (c) IS and (d) cIS, colored according to the desired variable (not groups). Figures (a) and (c) appear to have a single color as all group instances are clustered on top of each other.

3.2.1. Visualization:

Using the chosen parameters, Figure 8 and Figure 9 (marker color indicates instance TTC in months or MMSE, respectively) show the first two dimensions of the learned manifolds using cLE, LE, cIS, IS and PCA. It can be observed that there is a more coherent ordering (with respect to TTC and MMSE) of the data with the proposed cLE and cIS embeddings. The temporal trajectory of an exemplar subject is shown by arrows, where it can be seen that in general cLE provides a more visually pleasing movement of the subject across the low-dimensional mapping of the population. Figure 10 displays the general direction of subjects movements depending on their location in the manifold, where it can be appreciated that there is a more coherent direction of movement in both cLE and cIS. Figure 11 displays every subjects trajectory as a line in the 2D space and here as well it can be seen that both

cLE and cIS provide a more organized representation movement across the manifold.

Figure 8: 2D low-dimensional representations. Data points colored with respect to TTC in months (best seen in color) according to color-bar. Arrows (blue MCI and red AD) indicate an exemplar subject’s “flow” across time.

3.2.2. Correlation analysis:

Feature ordering was empirically tested in the first two dimensions independently. Table 3.2.1 summarizes the results of calculating the correlation coefficient (CC) between the first (D1) or second (D2) dimension of the low-dimensional embeddings (cLE, LE, cIS, IS and PCA) and TTC, MMSE, FAQ, CDRSB or ADAS13 scores. It can be seen that in every case the cLE

Figure 9: 2D low-dimensional representations. Data points colored with respect to MMSE score (best seen in color) according to color-bar. Arrows (blue MCI and red AD) indicate an exemplar subject’s “flow” across time.

embedding has higher absolute CC, which can be interpreted as a more coherent ordering of features (with regards to TTC and MMSE, [FAQ](#), [CDRSB](#) or [ADAS13 scores](#)). Another important aspect to note is that both cLE and cIS have higher absolute CC than LE, IS or PCA, further strengthening the capabilities of the proposed constraints.

3.2.3. TTC and biomarker estimation:

The low-dimensional embeddings’ local neighborhood (k NN estimation) of each instance was used to estimate subject instance TTC, MMSE, FAQ, CDRSB and ADAS13. The reference (or silver standard) value of the TTC

Figure 10: Vector field of 2D low-dimensional representations. Arrows indicate the average direction in the regular grid and are colored with respect to TTC in months (best seen in color) according to color-bar.

for each subject was defined as the half-way time point between the last MCI diagnosis and the first AD. In a leave-one-out experiment, TTC, MMSE, FAQ, CDRSB and ADAS13 were estimated for all instances using differently sized k -neighborhoods in the low-dimensional embedding by finding the average TTC and MMSE in the defined neighborhood. The lowest TTC estimation mean absolute error, of ~ 12.7 months, was obtained with the proposed method (cLE) using neighborhoods of between 100-150 instances. This compares to an mean absolute error ~ 14.8 for LE, ~ 12.8 for cIS, ~ 15.5 for IS and ~ 15.5 for PCA. Figures 3.2.3 and 3.2.3 summarize these results, where on the left column LBP normalization was used, while on the right col-

Figure 11: All subject trajectories in the 2D low-dimensional representations. Lines join subsequent subject time points.

umn MVM normalization was employed. Comparing Figures 3.2.3 and 3.2.3 left and right column it can also be observed that in every non-linear embedding method the added constraints improve the overall TTC and biomarker estimation, and that all methods converged to ~ 17.25 months when the neighborhoods included most data points. Similarly, the lowest and second lowest MMSE estimation mean absolute errors ~ 2.45 and ~ 2.55 points were obtained with cLE and cIS, respectively. This compared to ~ 2.61 for LE, and ~ 2.7 for IS and PCA.

4. Discussion and Future Work

In work presented here, a novel formulation of the Laplacian Eigenmaps and Isomap algorithms that allows the incorporation of grouping constraints

Figure 12: Comparison of absolute mean error of TTC, MMSE and FAQ estimation using cLE, LE, cIS, IS and PCA (best seen in color). Estimation with LBP (left column) and MVM (right column) normalization.

Figure 13: Comparison of absolute mean error of CDRSB and ADAS13 estimation using cLE, LE, cIS, IS and PCA (best seen in color). Estimation with LBP (left column) and MVM (right column) normalization.

Table 2: Correlation coefficient between TTC, MMSE, FAQ, CDRSB or ADAS13 scores, and D1 or D2 of LE, cLE, IS, cIS and PCA. Highest overall result in bold.

Features		TTC	MMSE	FAQ	CDRSB	ADAS13
D1	LE	0.501	-0.428	0.441	0.462	0.410
	cLE	0.608	-0.517	0.545	0.569	0.500
	IS	0.460	0.442	0.451	0.479	0.435
	cIS	0.471	0.388	0.405	0.411	0.381
	PCA	-0.450	0.395	-0.395	-0.390	-0.391
D2	LE	0.129	-0.100	0.051	0.064	0.102
	cLE	0.520	-0.523	0.513	0.557	0.502
	IS	0.108	0.109	0.071	0.068	0.104
	cIS	0.500	0.460	0.475	0.503	0.440
	PCA	0.103	-0.093	0.039	0.043	0.089

was presented. It was shown that by taking into account structure in the data by adding grouping constraints, the learned low-dimensional spaces were not only more visually pleasing but were also better suited to for TTC and MMSE estimation. Furthermore and as was mentioned before, it is also worth noting that the grouping constraints added to the neighborhood matrix of Laplacian Eigenmaps and Isomap can be applied to any method that relies in finding nearest neighbors, such as: Locality preserving projection (LPP) [26], locally linear embedding (LLE) [27] or neighborhood preserving embedding (NPE) [28], to name a few.

Choosing a larger dimensionality $d > 2$ in the LE, IS or PCA algorithms generally improves their performance, however, neither of these approaches performs as good as cLE or cIS regardless of how many dimensions are included. It is hypothesized that LE and IS perform poorly in the experiments due to the fact that local manifold geometry is not well estimated by the definition of neighborhoods that contain several points of the same subject, thus ignoring the structure with the data. Results also indicate that a more compact representation of the proposed cLE and cIS features over LE, IS or PCA is obtained, while displaying an improved ability in the temporal ordering of the features, without compromising in discriminative power. It is important that TTC and MMSE estimation results are put into the right context. For TTC estimation it needs to be considered that the time between study visits was on average 8.4 (range: [6; 48]) months and that the point of conversion for every subject was chosen to be the mid-way point between visits, and since conversion could have in reality occurred at any point between these two visits there is a minimum expected absolute error of

2.1 months. While for MMSE it needs to be considered that when originally proposed, it was shown that the test had test-re-test mean variation of 1.1 points when a the same tester did both examinations within a 24 hour period on the same patients, while a mean variation of 1.3 was observed when one tester did the first test and another the second [29]. Additionally, from the 1309 available samples 722 have a non-AD diagnosis and 587 were diagnosed as probable AD, that is, there are 23% more non-AD samples. This likely leads to biased results towards more pre-AD TTC estimations. However, defining neighborhoods based on a fixed distance (e.g. Euclidean) from an instance (instead of the number of instances) may alleviate or eliminate this bias. Mean classification values indicate, on average, a small advantage of the standard algorithms over the proposed methods. This is primarily due to the choice of parameter k . The proposed cLE seems to have an advantage in terms of TTC and MMSE estimation and visualization over the other compared methods, with cIS being a close second best method. Additionally, the proposed methods are able to estimate the intrinsic non-linear geometry of the data, while reducing the amount of bias towards the amount of samples per subject.

Age has been shown to be an important covariate of disease progression [11]. However, in the framework proposed here there is no straight forward way to incorporate subject age information. Aging related brain atrophy has been previously shown to be very similar to AD related atrophy. This poses the potential problem that two images (from different subjects) could be found to be similar, even though for example one could be from a healthy 90 year old and the other from a diseased 60 year old. Weak grouping constraints could be used to replace the proposed categorical grouping constraints. Under this scenario age could be used as a continuous weak grouping variable, that is, age could be used to only allow instances of subjects at a certain age range to the local neighborhood, with age difference acting as a edge weighing factor. The derived biomarkers contain several confounding aspects, e.g. MR imaging artifacts of apparent anatomical variations (e.g. for example due to hydration levels [30] or the menstrual cycle [31]) that affect the appearance of the MR images. Disease “progress” obtained directly from these biomarkers might be poorly estimated. Due to the unknown nature of the measurement’s confounding aspects, fitting a parametric disease progression model to these measurements could be used as an outlier removal. These models could be used to project the movement of a subject in the manifold and along the disease trajectory. Another interesting aspect that can

be addressed is the use of more than one type of biomarker, i.e. cognitive, volumetric, CSF measurements, etc. Specifically, models could be generated for several biomarkers and combined to give an aggregate estimate of the disease progress. In this regard, additional experiments to estimate current AD progress and disease progression rate [32, 33] found that the features (biomarkers) proposed here to provide a good performance as well as contributing complementary information to more traditional biomarkers, such as structural volumes and cognitive scores.

Acknowledgments

The research discussed here was developed in a joint research project with IXICO plc and funded by an Innovate UK (formerly UK Technology Strategy Board) grant. Data used in preparation of this work was obtained from the ADNI database (adni.loni.usc.edu). 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.

Author biographies

- **Ricardo Guerrero** works as a Research Associate at the Biomedical Image Analysis (BiomedIA) group, Imperial College London, UK where he received his PhD degree in Computer Science. His current research interests include pattern recognition and machine learning as well longitudinal modeling in medical imaging.
- **Christian Ledig** is a Research Associate in the Biomedical Image Analysis (BiomedIA) group at Imperial College London, UK where he received a PhD degree in Computing Research. His general research interests are image registration/segmentation, machine learning and classification.
- **Alexander Schmidt-Richberg** received his PhD at the Institute of Medical Informatics, University of Lübeck, Germany, in 2013. He worked as a Research Associate at the Biomedical Image Analysis (BiomedIA) group, Imperial College London, UK from 2013 to 2014. His interests are medical image processing, analysis and registration.

- **Daniel Rueckert** is Professor of Visual Information Processing and heads the Biomedical Image Analysis (BiomedIA) group at Imperial College London, UK. He received a Diploma in Computer Science (equiv to M.Sc.) from the Technical University Berlin and a PhD in Computer Science from Imperial College London.

References

- [1] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, in: *Advances in Neural Information Processing Systems 14*, Vol. 14, 2002, pp. 585–591.
- [2] J. B. Tenenbaum, V. Silva, J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, *Science* 290 (5500) (2000) 2319–2323.
- [3] O. C. Jenkins, M. J. Mataric, A spatio-temporal extension to Isomap nonlinear dimension reduction, in: *International conference on machine learning*, 2004, p. 56.
- [4] N. D. Lawrence, J. Quiñonero Candela, Local distance preservation in the GP-LVM through back constraints, in: *International conference on machine learning*, 2006, pp. 513–520.
- [5] M. Lewandowski, J. Martinez-del Rincon, D. Makris, J.-C. Nebel, Temporal Extension of Laplacian Eigenmaps for Unsupervised Dimensionality Reduction of Time Series, in: *International Conference on Pattern Recognition*, 2010, pp. 161–164.
- [6] R. Brookmeyer, E. Johnson, K. Ziegler-Graham, H. M. Arrighi, Forecasting the global burden of Alzheimers disease, *Alzheimer’s & Dementia* 3 (3) (2007) 186–191.
- [7] E. R. Siemers, K. L. Sundell, C. Carlson, M. Case, G. Sethuraman, H. Liu-Seifert, S. A. Dowsett, M. J. Pontecorvo, R. A. Dean, R. Demattos, Phase 3 solanezumab trials: Secondary outcomes in mild Alzheimer’s disease patients., *Alzheimer’s & dementia : the journal of the Alzheimer’s Association*.

- [8] R. S. Doody, V. Pavlik, P. Massman, S. Rountree, E. Darby, W. Chan, Predicting progression of Alzheimer’s disease., *Alzheimer’s research & therapy* 2 (1) (2010) 2.
- [9] E. Yang, M. Farnum, V. Lobanov, T. Schultz, R. Verbeeck, N. Raghavan, M. N. Samtani, G. Novak, V. Narayan, A. DiBernardo, Quantifying the pathophysiological timeline of Alzheimer’s disease., *Journal of Alzheimer’s disease: JAD* 26 (4) (2011) 745–753.
- [10] I. Delor, J.-E. Charoin, R. Gieschke, S. Retout, P. Jacqmin, Modeling Alzheimer’s Disease Progression Using Disease Onset Time and Disease Trajectory Concepts Applied to CDR-SOB Scores From ADNI., *CPT: pharmacometrics & systems pharmacology* 2 (2013) e78.
- [11] K. Ito, B. Corrigan, Q. Zhao, J. French, R. Miller, H. Soares, E. Katz, T. Nicholas, B. Billing, R. Anziano, T. Fullerton, Disease progression model for cognitive deterioration from Alzheimer’s Disease Neuroimaging Initiative database., *Alzheimer’s & dementia: the journal of the Alzheimer’s Association* 7 (2) (2011) 151–60.
- [12] C. R. Jack, D. S. Knopman, W. J. Jagust, L. M. Shaw, P. S. Aisen, M. W. Weiner, R. C. Petersen, J. Q. Trojanowski, Hypothetical model of dynamic biomarkers of the Alzheimer’s pathological cascade, *The Lancet Neurology* 9 (1) (2010) 119–128.
- [13] R. Wolz, P. Aljabar, J. Hajnal, D. Rueckert, Manifold learning for biomarker discovery in MR imaging, in: *Machine Learning in Medical Imaging*, Vol. 6357 of *Lecture Notes in Computer Science*, 2010, pp. 116–123.
- [14] D. Zhang, D. Shen, the Alzheimer’s Disease Neuroimaging Initiative, Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers, *PLoS ONE* 7 (3) (2012) e33182.
- [15] R. A. Heckemann, C. Ledig, K. R. Gray, P. Aljabar, D. Rueckert, J. V. Hajnal, A. Hammers, Brain Extraction Using Label Propagation and Group Agreement: Pincram., *PloS one* 10 (7) (2015) e0129211.
- [16] J. Ashburner, K. J. Friston, Why voxel-based morphometry should be used., *NeuroImage* 14 (6) (2001) 1238–43.

- [17] F. L. Bookstein, Voxel-Based Morphometry Should Not Be Used with Imperfectly Registered Images, *NeuroImage* 14 (6) (2001) 1454–1462.
- [18] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, D. J. Hawkes, Nonrigid registration using free-form deformations: Application to breast MR images, *IEEE Transactions on Medical Imaging* 18 (8) (1999) 712–721.
- [19] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society, Series B* 67 (2005) 301–320.
- [20] R. Guerrero, R. Wolz, A. W. Rao, D. Rueckert, Manifold population modeling as a neuro-imaging biomarker: Application to ADNI and ADNI-GO., *NeuroImage* 94C (2014) 275–286.
- [21] N. Meinshausen, P. Bühlmann, Stability selection, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72 (4) (2010) 417–473.
- [22] R. Guerrero, C. Ledig, D. Rueckert, Manifold Alignment and Transfer Learning for Classification of Alzheimer’s Disease, in: *Machine Learning in Medical Imaging*, Vol. 8679, 2014, pp. 77–84.
- [23] R. Guerrero, C. Ledig, A. Schmidt-Richberg, D. Rueckert, Group-constrained Laplacian Eigenmaps: Longitudinal AD biomarker learning, in: *Machine Learning in Medical Imaging*, Vol. 9352, 2015, pp. 178–185.
- [24] T. Ojala, M. Pietikäinen, D. Harwood, A comparative study of texture measures with classification based on featured distributions, *Pattern Recognition* 29 (1) (1996) 51–59.
- [25] L. van der Maaten, E. Postma, J. van den Herik, Dimensionality reduction: A comparative review, *Tilburg University Technical Report*.
- [26] X. He, P. Niyogi, Locality preserving projections, in: *Neural information processing systems*, 2004.
- [27] S. T. Roweis, L. K. Saul, Nonlinear dimensionality reduction by locally linear embedding, *Science* 290 (2000) 2323–2326.

- [28] X. He, D. Cai, S. Yan, H.-J. Zhang, Neighborhood preserving embedding, in: IEEE International Conference on Computer Vision (ICCV), Vol. 2, 2005, pp. 1208–1213.
- [29] M. F. Folstein, S. E. Folstein, P. R. McHugh, “Mini-Mental State”: A practical method for grading the cognitive state of patients for the clinician, *Journal of Psychiatric Research* 12 (3) (1975) 189–198.
- [30] T. Duning, S. Kloska, O. Steinsträter, H. Kugel, W. Heindel, S. Knecht, Dehydration confounds the assessment of brain atrophy., *Neurology* 64 (3) (2005) 548–50.
- [31] G. Hagemann, T. Ugur, E. Schleussner, H.-J. Mentzel, C. Fitzek, O. W. Witte, C. Gaser, Changes in brain size during the menstrual cycle., *PloS one* 6 (2) (2011) e14655.
- [32] A. Schmidt-Richberg, C. Ledig, R. Guerrero, H. Molina-Abril, A. Frangi, D. Rueckert, Learning Biomarker Models for Progression Estimation of Alzheimer’s Disease., *PloS one* 11 (4) (2016) e0153040.
- [33] A. Schmidt-Richberg, R. Guerrero, C. Ledig, H. Molina-Abril, A. F. Frangi, D. Rueckert, the Alzheimers Disease Neuroimaging Initiative, Multi-stage biomarker models for progression estimation in alzheimer’s disease, in: International Conference on Information Processing in Medical Imaging (IPMI), 2015.