ANOMALY DETECTION IN SINGLE SUBJECT VS GROUP USING MANIFOLD LEARNING

Florian Tilquin⋆*, Sylvain Faisan*⋆*, Fabrice Heitz*⋆*, Vincent Noblet*⋆*, Frédéric Blanc*⋆ *& Izzie Namer*⋆
⋆ ICube UMR 7357, Université de Strasbourg, CNRS, Strasbourg, France

ABSTRACT

This paper compares several linear and non-linear multivariate models for the detection of abnormal patterns in neuroimaging data, when comparing a single subject to a normal control group. The proposed methods learn the manifold spanned by the normal controls using non-linear dimension reduction techniques. The image of a subject is projected on the control group manifold either via a standard projection or through an embedding/reconstruction scheme. A comparison of the reconstruction with the subject's original neuroimaging data allows for the detection of abnormal patterns by way of statistical tests on the residuals. The different abnormality detection methods are assessed on synthetic data and real (MRI) neuroimaging data. The importance of non-linear modeling of the manifold in the reduced-dimension subspace is highlighted, as well as robustness to large abnormalities.

***Index Terms*—** Anomaly detection, subject vs group comparison, manifold learning, PCA, Isomap, LLE.

1. INTRODUCTION

1.1. Context

We consider the detection of abnormal patterns in neuroimaging data, in the context of comparing a single subject to a normal control group. Standard approaches for anomaly detection [1] are related to the one-class classification problem, in which one tries to detect outliers (corresponding here to "abnormal" subjects) with respect to a learned distribution of normal controls. These approaches will make a global statement about the subject class (i.e. pathological or not) but do not provide a spatial localization of abnormal patterns within the subject's image data. On the other hand, the approaches developed for localizing subject-specific abnormalities [2] generally resort to univariate voxel-wise or ROI-based statistical tests and often rely on Gaussian distribution assumptions [3].

In this paper we present and compare different methods for the detection and localization of subject-specific abnormal patterns within the framework of subject-versus-group comparison. The proposed methods rely on global (multivariate) non-linear models of normal image data, which allow for the representation of complex spatial patterns with non Gaussian distributions. The manifold of normal image patterns is learned from a control group with the help of non-linear dimension reduction techniques. Identifying abnormalities is mathematically associated with finding the projection of a subject onto the manifold in which the control group lies. The proposed projection paradigm is described in section 2, along with the five different models tested in this paper. In section 3, experiments on synthetic and real data underline the benefit of using non-linear multivariate representations, compared to standard univariate or multivariate linear approaches.

2. METHODS

2.1. The projection onto the manifold paradigm

We use the strategy developed in [4, 5] to provide localized anomaly detection, when a normal control database is available. In essence the paradigm consists in "projecting" any new subject image $\mathbf{Y}$ onto a learned manifold representing the normal controls. The "projection" $\mu(\mathbf{Y})$ will correspond to the image closest to that of the tested subject $\mathbf{Y}$, while belonging to the normal controls manifold. The residual $\mathbf{Y} - \mu(\mathbf{Y})$, which is representative of any abnormalities present in $\mathbf{Y}$, is computed and converted into a *p*-value map (or a z-score map) that can be thresholded for detection purposes. The p-value conversion requires at first the estimation of the probability density function (*pdf*) of the residual value at each component (voxel) under $\mathcal{H}_0$ (when the subject belongs to the control group). To learn the geometric structure of the control group while coping with the huge dimensionality of the data, dimension reduction algorithms are used. They transform high dimensional data points into lower dimensional embeddings while "untangling" the data geometry. We will also need a reconstruction operator ρ, that reconstructs a high dimensional sample from a low dimensional one. As a result, we have $\mu = \rho \circ \pi$, where π is the embedding (dimension reduction) operator.

Specifically, the proposed approaches work as follows from a learning set $\mathbf{X}$ of controls: first, the projection operator μ is learned from a subset of $\mathbf{X}$. Then, given μ, another subset of $\mathbf{X}$ allows to compute the *pdf*s of the residual values at each voxel under $\mathcal{H}_0$. For the sake of simplicity, the *pdf*s of the residuals are supposed to be independent and to follow a zero mean Gaussian *pdf* with a standard deviation that varies across the voxels. Consequently, a z-score map related to the testing of a new subject $\mathbf{Y}$ is obtained by dividing each component of $\mathbf{Y} - \mu(\mathbf{Y})$ with the related standard deviation.

2.2. Proposed multivariate models

All the proposed projection methods use multivariate models except the first one GLM, which is the reference univariate model. For comparison purposes, we also consider PCA as a standard linear projection model. The third model (RNGPA) uses linear dimension reduction but non-linear robust modeling in the subspace, while the last three models are based on non-linear dimension reduction.

2.2.1. General Linear Model (GLM) [3]

The General Linear Model (GLM), popularized by the software Statistical Parametric Mapping [6] (SPM), is a standard statistical tool for medical images group analysis. It has also been used in a degenerate case to compare a subject to a group [3]. The classical GLM is univariate. In subject versus group analysis, it amounts to using a constant projection operator μ that simply returns for each voxel the mean $\mathbf{m}$ of the control group [3].

2.2.2. Principal Component Analysis (PCA) [7, 8]

PCA is a linear dimension reduction technique that provides an analytic solution for both the dimension reduction and the reconstruction operators. Dimension reduction of a test subject $\mathbf{Y}$ is performed as: $\pi(\mathbf{Y}) = \mathbf{y} = \mathbf{W}^T(\mathbf{Y} - \mathbf{m})$, where $\mathbf{m}$ is the mean vector of the normal control training set, and $\mathbf{W}$ is a projection matrix that is composed of the eigenvectors associated to the highest eigenvalues, obtained with the Karhunen-Loeve transformation of the training set. The reconstruction $\rho(\pi(\mathbf{Y}))$ is derived as: $\rho(\mathbf{y}) = \mathbf{W}\mathbf{y} + \mathbf{m}$.

2.2.3. Robust non-Gaussian probabilistic PCA (RNGPCA) [5]

This method is based on probabilistic PCA, which provides the link between a complete multivariate probabilistic model, dimension reduction (by PCA) and maximum likelihood density estimation. The generative model writes: $\mathbf{Y} = \mathbf{W}\mathbf{y} + \mathbf{m} + \epsilon$. $\mathbf{W}$ and $\mathbf{m}$ are learned, as previously, from a set of normal controls $\mathbf{X}$ with standard PCA. The *pdf* of subspace variable $\mathbf{y}$ is modeled as a mixture of Gaussian kernels whose bandwidth is estimated with cross-validation (this parameter was set manually in [5]). In addition, a non-Gaussian noise model, robust to outliers, is specified for ϵ, yielding a completely non-linear statistical model. The dimension reduction $\mathbf{y} = \pi(\mathbf{Y})$ of a new subject $\mathbf{Y}$ is performed via robust Maximum a Posteriori (MAP) estimation using a mean shift and a semi-quadratic algorithm. The reconstruction $\rho(\pi(\mathbf{Y}))$ is performed with the generative linear model (see [5] for details).

2.2.4. Isometric Mapping (Isomap) [9]

Isomap is a manifold learning technique that aims to best preserve the geodesic distances between the original samples. The computation of geodesic distances is achieved with the Dijkstra's shortest path algorithm over a weighted graph computed from the samples. Preserving the distances in the low-dimensional space is achieved by means of multidimensional scaling [10]. Isomap does not naturally allow for neither an extension of the embedding to new points, nor a reconstruction of an embedded point. We address the first issue (corresponding to the specification of π) with the Nyström extension [11, 12], and the second one (specification of ρ) by treating it as a supervised regression problem with the Nadaraya-Watson kernel regression [13] (the bandwith of the kernel is estimated using a cross-validation strategy).

2.2.5. Locally Linear Embedding (LLE)-based approach

LLE [14], like Isomap, provides nonlinear dimensionality reduction. It yields a neighborhood preserving mapping, the preserved local properties being the weights that best reconstruct each data point from its neighbors. In our paradigm, using the LLE mapping would require to address the problem of extending the embedding to new points and of reconstructing embedded points. However, we propose simply to define the projection operator μ from the LLE weights: given a test sample $\mathbf{Y}$ and its K-nearest neighbors, $\mu(\mathbf{Y})$ is defined as the best (linear) reconstruction of $\mathbf{Y}$ from its neighbors in the normal group $\mathbf{X}$. The weights are computed as in [14]: they are constrained to sum to one and a L^2 loss function is used. Optimization is achieved by solving a linear system of equations [15]. In order to penalize large weights, a regularization term is added in the system of equations [15]. One can easily show that this leads to a ridge regression problem:

$$\begin{aligned} \mu(\mathbf{Y}) =& \mathbf{X}.\arg\min_{w} \left(\|\mathbf{X}\mathbf{w} - \mathbf{Y}\|_2^2 + \lambda\|\mathbf{w}\|_2^2\right) \\ & \text{s.t. } \sum_i w_i = 1, w_i \neq 0 \Leftrightarrow i \in \mathcal{V}(\mathbf{Y}) \end{aligned} \quad (1)$$

where λ is the regularization parameter and $\mathcal{V}(\mathbf{Y})$ is the set containing the indexes of the neighbors of $\mathbf{Y}$ among the normal group $\mathbf{X}$, computed with an L^2 euclidean distance.

2.2.6. Robust LLE-based approach (RLLE)

In LLE, the L^2 distances are heavily affected by abnormal voxels and can prevent from finding correct neighbors. Moreover, the reconstruction weights are estimated using a L^2 loss function, and thus abnormal components have a large influence on the reconstruction (ideally, they should have no influence on the results). In order to obtain a method that is robust to abnormal components, we define an iterative algorithm for robust locally linear projection (RLLE) which works as follows: given a test sample $\mathbf{Y}$, we first perform a robust projection onto its L^2 neighbors by minimizing the L^1 reconstruction error (instead of L^2). Then, we repeat the following step until convergence: we use the neighbors of the current projection to obtain a new projection of $\mathbf{Y}$ (by minimizing the L^1 error reconstruction). Convergence is achieved when

the L^2 neighbors of the new projection are identical to those of the current projection. At iteration k, the new projection $\mu(\mathbf{Y})_{k+1}$ is derived from the previous one $\mu(\mathbf{Y})_k$ as follows:

$$\begin{aligned}\mu(\mathbf{Y})_{k+1} =& \mathbf{X}.\arg\min_{w}\left(\|\mathbf{X}\mathbf{w}-\mathbf{Y}\|_1+\lambda\|\mathbf{w}\|_2^2\right)\\ &\text{s.t. } \sum_i w_i = 1, w_i \neq 0 \Leftrightarrow i \in \mathcal{V}(\mu(\mathbf{Y})_k)\end{aligned} \quad (2)$$

where λ is the regularization parameter and $\mathcal{V}(\mu(\mathbf{Y})_\mathbf{k})$ the set containing the indexes of the k-th projection of $\mathbf{Y}$ $(\mu(\mathbf{Y})_\mathbf{k})$ neighbors among the normal group $\mathbf{X}$.

Optimization is achieved using sequential quadratic programming. Convexity of the criterion and of the constraints ensures convergence to a global minimum (it may be not unique since the criterion is not strictly convex).

3. EXPERIMENTS

3.1. Comparison of methods on synthetic data

We first create a synthetic dataset in which we control the location of abnormalities. As it can be quite cumbersome to sample directly from a high-dimensional space, we embed a half-sphere (to avoid any closed manifold issue) of dimension d into a space of dimension D with $D >> d$. To this end, an orthonormal transformation that maps the low- to the high-dimensional space is randomly generated. The embedding mapping is thus linear, but the data in the latent space is distributed non linearly. Control data are generated as follows: we first sample N points on the half-sphere. The orthonormal transformation is then applied to each point, producing N points of dimension D. Finally, a zero-mean Gaussian noise of variance σ^2 is added independently on each component of the points. σ is specified so that the L^2 norm of most of the noisy samples are between 0.95 and 1.05, independently from the value of D (the L^2 norm of the noise-free samples is 1), thus "preserving relatively well the topology of the manifold". The abnormal data are generated as the control ones except that anomalies are introduced in $\alpha\%$ of the components. The components that present the highest variance in the control data are selected to be abnormal. Anomalies are introduced as a multiple of the noise standard deviation: 4σ is added to selected components.

We present results obtained with different parameter settings for data generation. We start from an easy case (EC) ($N = 10000$, $D = 100$, $d = 3$ and $\alpha = 30$) : the training set is composed of a high number of points ($N = 10000$) compared to the intrinsic dimension of the manifold ($d = 3$) and the high-dimensional space is of dimension $D = 100$. Finally, $\alpha = 30\%$ of components are selected to be abnormal for the creation of abnormal subjects. Then, we derive three more challenging settings from the EC case as follows: *(i)* we increase the intrinsic dimension ($d = 20$) from the EC setting, *(ii)* we reduce the number of samples ($N = 500$) from EC , and *(iii)* the number of created abnormal components is reduced to $\alpha = 5$ in EC. These three new settings are denoted respectively ECd, ECN, and $EC\alpha$. Results are expressed in terms of AUC (Area Under the ROC Curve). Instead of integrating the true positive rate for a false positive rate from 0 to 1, we integrate it only from 0 to 0.01 because the behaviour of the ROC curve is of no interest for large values of the false positive rate in a medical context. Since the modified AUC score is now at most equal to 0.01, it is multiplied by 100 so that it may vary from 0 to 1. Table 3.1 presents results obtained in terms of the modified AUC score for the different settings (EC,ECd,ECN, and $EC\alpha$). In all settings, the number of neighbors for Isomap, LLE and RLLE was set to 30, the number of dimension for all dimension reduction methods was set to 20 and the λ regularizing parameter for RLLE and LLE was set to 0.01.

	EC	ECd	ECN	$EC\alpha$
GLM	0.005 ±0.001	0.010 ±0.001	0.005 ±0.005	0.001 ±0.001
PCA	0.129 ±0.052	0.151 ±0.033	0.151 ±0.076	0.670 ±0.148
RNGPCA	**0.674** ±0.163	**0.370** ±0.157	**0.651** ±0.105	**0.917** ±0.019
ISO	0.386 ±0.027	0.028 ±0.003	0.248 ±0.046	0.841 ±0.087
LLE	0.252 ±0.014	0.081 ±0.011	0.185 ±0.041	0.653 ±0.080
RLLE	0.521 ±0.050	0.092 ±0.011	0.492 ±0.069	0.859 ±0.052

Table 1. Modified AUC scores (best possible score is 1, and worst is 0) for the four presented experiments and for each presented method.

We can first note that the univariate GLM approach provides very bad results for all settings. This is due to the fact that it does not properly model the distribution of the data. The RNGPCA method appears here as a clear winner, and while the method of embedding the data in a higher dimension space is linear (and thus perfectly fits the PCA model), the data distribution in the latent space is not, and is therefore nicely captured by RNGPCA. PCA globally gives poor results. Overall, RNGPCA, which relies on non-Gaussian and robust modeling, is a clear winner, whatever the parameter settings. The second best method is RLLE, which also relies on robust subspace modeling. The other nonlinear methods (ISO, LLE) give mixed results and are very sensitive on parameter settings. PCA-based methods are less affected by an increase of the intrinsic dimension d or by a decrease of the number of samples N. Indeed, the other methods rely on a neighborhood that may be strongly disturbed by the emptiness of the latent space. Finally, we can observe that for all methods, results obtained with the $EC\alpha$ setting are better than those obtained with the EC setting. This is especially true for PCA, ISO and LLE that are not robust to an excessive proportion of abnormal components. These results clearly show that robustness to abnormal components is a crucial feature of the best methods. They also show that a linear-projection method such as RNGPCA can compete with non-linear dimension reduction approaches provided that an appropriate non-linear model is used in the reduced subspace.

3.2. Application on a database of AD patients

We applied the previous two best methods (RNGPCA and RLLE) and the GLM approach, that is widely used in neu-

roimaging, to a real dataset comprising MRI samples of both healthy subjects and Alzheimer's disease (AD) patients. The healthy subjects serve as a control group over which AD patients are projected to localize the anomalies. We used three public databases OASIS [16], ADNI [17] and IXI (http://brain-development.org/ixi-dataset/), as well as an internal database. This extended dataset contains a total of 1507 structural MRI images (around 1000 of which are healthy subjects scans).

3.2.1. Preprocessing

All images are registered using the ANTS [18] registration tool. ANTS provides for each image a deformation field warping this image from its original space to a common one, as well as to a template (defined in the common space). For the observations, we consider volumetric information that informs on how each anatomical structure of the brain has to be reduced or enlarged to match the structures composing the template. This is achieved, for each deformation field, by computing the determinant of the Jacobian (DJ) of the deformation field at each voxel (i.e. the differential matrix of the vector field). To make the DJ images invariant to the subjects head volume, each image is divided by its sum. Finally, to put both the volumes atrophies and hypertrophies on a common scale, we compute the voxel-wise log of the normalized DJ, as multiplying or dividing the volume of an inner structure by a factor α corresponds to adding or subtracting α to the log of DJ.

3.2.2. Abnormality detection

For each tested method (GLM, RNGPCA, RLLE), we computed individual z-score maps for each AD subject., We cannot verify that all detected anomalies are coherent with the medical history of the tested subjects but since all subjects suffer from AD, the intersection of the detected maps should be in accordance with the pathology. That is why we computed the voxel-wise proportion of z-scores greater than 3 (in magnitude) for a given voxel.

Figure 1 presents the detection results. We can first note that the 3 methods detect the parahippocampal regions (see bottom images of Figure 1). This is a consistent result because the volume of the hippocampus is a known biomarker of AD. Then, one can observe that the GLM results differ significantly from the two other ones, as it introduces quite a nearly full detection of the ventricles in a large number of cases ($> 20\%$), while only some of the edges of the ventricles are detected with multivariate methods. This difference is related to the fact that the size of the ventricles may vary strongly in the normal control group. Most of them have small ventricles but some of them have very large ventricles. In the case of the GLM approach, large ventricles of a AD subject are detected as abnormal because most of the normal controls have small ventricles. In the case of the RLLE approach, an AD subject (with large ventricles) may be well reconstructed because the method selects, as the neighborhood, the few controls that also have large ventricles. As a result, ventricles may be well reconstructed and may be not considered as abnormal. Similarly, with the RNGPCA approach, the distribution of the normal controls in the reduced subspace may indirectly account for the size of the ventricles. Therefore, the reduction of an AD subject (with large ventricles) will lead to a reduced subject whose reconstruction may have large ventricles. Two different conclusions can be drawn. One can conclude with the GLM approach that, in the average, AD subjects have larger ventricles than normal controls. With the other approaches we see that the ventricle size is not specific to AD. This example highlights the interest of using non-linear multivariate approaches for a fine analysis of abnormal patterns in neuroimaging data. Finally, RNGPCA and RLLE provide very similar results, with maybe a small advantage to the RLLE method that provides more detections in the expected areas.

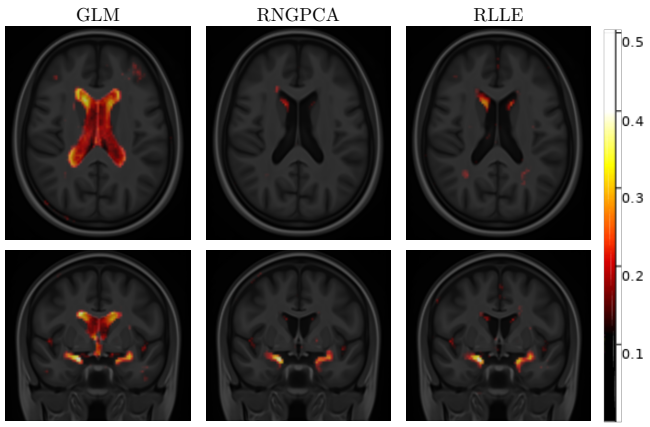

Fig. 1. Percentage of AD subjects in which a voxel had a z-score greater than 3 in magnitude. Only mean detections over 15% are represented. Top: axial slice, bottom: coronal slice.

4. CONCLUSION

In this paper, we have compared several linear and non-linear methods for performing anomaly detection in images in the single subject versus group setup. The proposed methods were based on manifold learning methods such as PCA, LLE or Isomap. We confronte these methods to the most widely used algorithm for anomaly detection in medical images: the GLM. We presented results over both a non-linear, synthetic dataset and a real MRI dataset with focus on Alzheimer's Disease. The two methods that clearly stand out (RNGPCA and RLLE) both rely on non-linear subspace modeling and implement robustness to large anomalies. The results obtained on the AD data set must still be analyzed in depth by a medical expert, based on the history of each patient.

5. REFERENCES

[1] Varun Chandola, Arindam Banerjee, and Vipin Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 15:1–15:58, 2009.

[2] Andrew R. Mayer, Edward J. Bedrick, Joseph M. Ling, Trent Toulouse, and Andrew Dodd, "Methods for identifying subject-specific abnormalities in neuroimaging data," *Human Brain Mapping*, vol. 35, pp. 5457–5470, 2014.

[3] Rik Henson, "Comparing a single patient versus a group of controls (and SPM)," `http://www.mrc-cbu.cam.ac.uk/personal/rik.henson/personal/Henson_Singlecase_06.pdf`, accessed: 2018-10-15.

[4] Torbjørn Vik, Fabrice Heitz, Izzie Namer, and Jean-Paul Armspach, "On the modeling, construction, and evaluation of a probabilistic atlas of brain perfusion," *Neuroimage*, vol. 24, no. 4, pp. 1088–1098, 2005.

[5] Torbjørn Vik, Fabrice Heitz, and Pierre Charbonnier, "Robust pose estimation and recognition using non-gaussian modeling of appearance subspaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, 2007.

[6] Karl J Friston, Andrew P Holmes, Keith J Worsley, J-P Poline, and Chris D Frith, "Statistical parametric maps in functional imaging: a general linear approach," *Human Brain Mapping*, vol. 2, no. 4, pp. 189–210, 1994.

[7] Ian T Jolliffe, "Principal component analysis and factor analysis," in *Principal component analysis*, pp. 115–128. Springer, 1986.

[8] Míriam López, Javier Ramírez, Juan Manuel Górriz, Ignacio Álvarez, Diego Salas-Gonzalez, Fermín Segovia, Rosa Chaves, Pablo Padilla, Manuel Gómez-Río, Alzheimer's Disease Neuroimaging Initiative, et al., "Principal component analysis-based techniques and supervised classification schemes for the early detection of Alzheimer's disease," vol. 74, no. 8, pp. 1260–1271, 2011.

[9] Joshua B Tenenbaum, Vin De Silva, and John C Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[10] Mark L Davison, *Multidimensional scaling*, Krueger, 1991.

[11] Yoshua Bengio, Jean-Françcois Paiement, Pascal Vincent, Olivier Delalleau, Nicolas L Roux, and Marie Ouimet, "Out-of-sample extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering," in *Advances in neural information processing systems (NIPS)*, 2004, pp. 177–184.

[12] Petros Drineas and Michael W Mahoney, "On the Nyström method for approximating a Gram matrix for improved kernel-based learning," *Journal of Machine Learning Research*, vol. 6, no. Dec, pp. 2153–2175, 2005.

[13] Elizbar A Nadaraya, "On estimating regression," *Theory of Probability & Its Applications*, vol. 9, no. 1, pp. 141–142, 1964.

[14] Sam T Roweis and Lawrence K Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[15] Lawrence K Saul and Sam T Roweis, "An introduction to locally linear embedding," `https://cs.nyu.edu/~roweis/lle/papers/lleintro.pdf`, accessed: 2018-10-15.

[16] Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner, "Open access series of imaging studies (OASIS): cross-sectional mri data in young, middle aged, nondemented, and demented older adults," *Journal of Cognitive Neuroscience*, vol. 19, no. 9, pp. 1498–1507, 2007.

[17] Clifford R Jack, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L Whitwell, Chadwick Ward, et al., "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods," *Journal of Magnetic Resonance Imaging*, vol. 27, no. 4, pp. 685–691, 2008.

[18] Brian B Avants, Nick Tustison, and Gang Song, "Advanced normalization tools (ANTS)," *Insight Journal*, vol. 2, pp. 1–29, 2009.