The Relevance Voxel Machine (RVoxM): A Self-tuning Bayesian Model for Informative Image-based Prediction

Mert R. Sabuncu Koen Van Leemput
for the Alzheimer Disease Neuroimaging Initiative (ADNI)

***Abstract*—This paper presents the Relevance Voxel Machine (RVoxM), a dedicated Bayesian model for making predictions based on medical imaging data. In contrast to the generic machine learning algorithms that have often been used for this purpose, the method is designed to utilize a small number of spatially clustered sets of voxels that are particularly suited for clinical interpretation. RVoxM automatically tunes all its free parameters during the training phase, and offers the additional advantage of producing probabilistic prediction outcomes. We demonstrate RVoxM as a regression model by predicting age from volumetric gray matter segmentations, and as a classification model by distinguishing patients with Alzheimer's disease from healthy controls using surface-based cortical thickness data. Our results indicate that RVoxM yields biologically meaningful models, while providing state-of-the-art predictive accuracy.**

I. Introduction

Medical image-based prediction aims at estimating a clinically or experimentally relevant quantity directly from individual medical scans. In a typical scenario, the properties of prediction models are learned from so-called training data – a set of images for which the quantity of interest is known. The trained models can then be applied to make predictions on new cases. In so-called image-based *regression* problems, the quantity to be estimated is continuously valued, such as a score evaluating a subject's brain maturity. In other cases, the aim is to predict a discrete value indicating one of several conditions, such as a clinical diagnosis, which is a *classification* problem.

Image-based prediction models derive their predictive power from considering all image voxels simultaneously, distilling high prediction accuracy from many voxel-level measurements that are each only weakly predictive when considered individually. This approach is fundamentally different from more traditional ways of relating image content to the biomedical context, such as voxel- and deformation-based morphometry [1]–[3], cortical thickness analysis [4], or voxel-level fMRI analysis [5], in which maps of affected anatomical areas are generated by merely considering each location separately. Unlike such "mapping" approaches, prediction methods explore patterns of association *between* voxels, offering powerful new analysis tools in such applications as "mind reading" [6], [7], studying neural information processing [8]–[10], image-based clinical diagnosis [11]–[16], and examining global patterns associated with healthy development, aging, pathology or other factors of interest [17], [18].

The principal difficulty in obtaining good regression and classification models for medical imaging data is the enormous number of voxels in images, which leads to two interconnected modeling challenges. First, since the number of training images is typically several orders of magnitude smaller than the number of voxels, the complexity of voxel-based prediction models needs to be strictly controlled in order to avoid so-called "over-fitting" to the training data, where small modeling errors in the training images are obtained at the expense of poor prediction performance on new cases. Second, the aim is often not only to predict well, but also to obtain insight into the anatomical or functional variations that are driving the predictions – in some applications, such as task-related fMRI, this is even the primary result. Interpreting complex patterns of association between thousands of image voxels is a highly challenging task, especially when the results indicate that *all* the voxels are of importance simultaneously; when the voxels contributing most to the predictions appear randomly scattered throughout the image area; or when the relevant inter-location relationships are non-linear in nature [19]–[22].

Various approaches for restricting model complexity in medical image-based prediction have been proposed in the literature, often with the adverse effect of reducing biological interpretability. Some methods only allow a selected few voxels to contribute to the prediction, either by performing

M.R. Sabuncu is with the Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, Charlestown, MA 02129, USA; and the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

K. Van Leemput is with the Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School; the Department of Informatics and Mathematical Modeling, Technical University of Denmark, Denmark; and the Departments of Information and Computer Science and of Biomedical Engineering and Computational Science, Aalto University, Finland

Support for this research was provided by the NIH NCRR (P41-RR14075), the NIBIB (R01EB006758, R01EB013565, 1K25EB013649-01), the NINDS (R01NS052585), the Academy of Finland (133611), the Finnish Funding Agency for Technology and Innovation (ComBrain), and the Harvard Catalyst, Harvard Clinical and Translational Science Center (NIH grant 1KL2RR025757-01 and financial contributions from Harvard and its affiliations) and an NIH K25 grant (1K25EB013649-01, NIH NIBIB).

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database(ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators is available at http://tinyurl.com/ADNI-main.

voxel-wise tests a priori [7] or by pruning less predictive voxels as part of a larger modeling process [23]–[25]. Others aim at using spatially connected patches of voxels as "features" instead of individual voxels themselves, either by averaging over a priori defined anatomical structures [7], [20], [26], [27], or by trying to cluster neighboring voxels in such a way that good prediction performance is obtained [22]. Yet others rely directly on more off-the-shelf techniques, for instance using only a few of all the available training subjects in sparse kernel-based machine learning methods [14], or reducing feature dimensionality by spatial smoothing, image downsampling, or principal component analysis [9], [28], [29].

In order to obtain prediction models that are expressly more biologically informative, some authors have started to exploit the specific spatial, functional, or temporal structure of their imaging data as a basis for regularization [30]–[37]. Building on this idea, we propose in this paper the Relevance Voxel Machine (RVoxM), a novel Bayesian method for image-based prediction that combines excellent prediction accuracy with intuitive interpretability of its results. RVoxM considers a family of probabilistic models to express that (1) not all image locations may be equally relevant for making predictions about a specific experimental or clinical condition, and (2) image areas that are somehow biologically connected may be more similar in their relevance for prediction than completely unrelated ones. It then assesses which model within this family explains the training data best, using the fact that simple models that sufficiently explain the data without unnecessary complexity are automatically preferred in Bayesian analysis [38]. As we shall see, this technique yields models that are *sparse* – only a small subset of voxels is actually used to compute predictions – as well as *spatially smooth* – in our experiments we used spatial proximity as a measure of biological connectivity. Such models are easier to interpret than speckles of isolated voxels scattered throughout the image area, and at the same time have an adequately reduced number of degrees of freedom to avoid over-fitting to the training data.

Compared to many existing image-based prediction methods, our Bayesian approach has several advantages:

- **Simultaneous regularization, feature selection, and biological consistency:** Rather than computing discriminative features in a separate pre-processing step [22], [25], [28], or using post-processing to analyze which subset of voxels contributes most to the predictions [9], [34], [37], the proposed method automatically determines which voxels are relevant – and uses only these voxels to make predictions – in a single consistent modeling framework. In line with anatomical expectations, the obtained maps of "relevance voxels" have spatial structure, facilitating biological interpretation and contributing to the regularization of the method.
- **Self-tuning:** The proposed method automatically tunes all the parameters of the prediction model, allowing the model to adapt to whatever degree of spatial sparseness and smoothness is indicated by the training data. In contrast, other image-based prediction methods rely on regularization parameters that need to be determined externally, either by manual selection [30], [33], [34] or using cross-validation [22], [26], [28], [31], [35], [37]. As illustrated in [39], the latter can be extremely challenging when several regularization parameters need to be determined simultaneously.
- **Probabilistic predictions:** In contrast to the decision machines [40] widely used in biomedical image classification, which aim to minimize the risk of misclassification, the method we propose computes posterior probabilities of class membership, from which optimal class assignments can subsequently be derived. The ability to obtain probabilistic predictions rather than "hard" decisions is important for building real-world diagnostic systems, in which image-based evidence may need to be combined with other sources of information to obtain a reliable diagnosis, and the risk of making false positive diagnoses needs to be weighed differently than that of false negative ones [41].

We originally presented RVoxM in a short conference paper that only dealt with the regression problem [42]. The current manuscript extends the theory to encapsulate classification, contains more details on theoretical derivations, and includes more extensive experimental results.

A reference Matlab implementation of the method is freely available from the authors.

II. RVoxM for Regression

For regression problems, the aim is to predict a real-valued target variable $t \in \mathbb{R}$ from an image $\mathbf{x} = (x_1, \ldots, x_{M-1}, 1)^{\mathrm{T}}$, where $x_i \in \mathbb{R}$ denotes a voxel-level measurement at the voxel indexed by i, and $M-1$ is the total number of voxels. For notational convenience in the remainder, an extra element with value 1 is also included to account for constant offsets in our predictions.

We use a standard linear regression model for t, defined by the Gaussian conditional distribution

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}; \mathbf{w}), \beta^{-1}) \tag{1}$$

with variance β^{-1} and mean

$$y(\mathbf{x}; \mathbf{w}) = \sum_{i=1}^{M-1} x_i w_i + w_M = \mathbf{w}^{\mathrm{T}}\mathbf{x}, \tag{2}$$

where $\mathbf{w} = (w_1 \cdots w_M)^{\mathrm{T}} \in \mathbb{R}^M$ denotes a vector of unknown, adjustable "weights" encoding the strength of each voxel's contribution to the prediction of t. In order to complete the model, we also define a prior on these weights that expresses our prior anatomical expectations that not all locations in the image may be equally predictive, and that biologically related areas may be more similarly predictive than completely unrelated ones. In particular, we use a prior of the form

$$p(\mathbf{w}|\boldsymbol{\alpha}, \lambda) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^{M} \alpha_i w_i^2 - \frac{\lambda}{2}\|\boldsymbol{\Gamma}\mathbf{w}\|^2 \right), \tag{3}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)^{\mathrm{T}}$ and λ are so-called hyper-parameters, and $\boldsymbol{\Gamma}$ is a matrix chosen so that elements of the vector $\boldsymbol{\Gamma}\mathbf{w}$ evaluate to large values when biologically

connected voxels happen to have very different weights in $\mathbf{w}$, effectively making such configurations *a priori* less likely.

The role of the hyper-parameters in eq. (3) is to express a wide range of regression models that can each be tried and tested, ranging from very complex models with many degrees of freedom (when all the hyper-parameters are set to small values) to heavily regularized ones with limited expressive power (when all the hyper-parameters are large). It is these hyper-parameters that are automatically learned from training data, as will be explained in section II-A, allowing the data to select the appropriate form of the model. When a large value for λ is selected, the model encodes a preference for configurations of $\mathbf{w}$ in which biologically connected voxels have similar weights. At the same time, setting some of the hyper-parameters α_i to very large values (infinity in theory) will clamp the values for the weights w_i in the corresponding voxels to zero, effectively "switching off" the contribution of these voxels to the prediction and removing them from the model. This ability for the data to determine which inputs should influence the predictions is similar to the Automatic Relevance Determination (ARD) mechanism [43] used in the Relevance Vector Machine (RVM) [44]; in fact, for $\lambda = 0$ our model reduces to an RVM with the voxel-wise intensities stacked as basis functions.

For the remainder of this paper, we will use a matrix $\mathbf{\Gamma}$ that simply encourages local spatial smoothness of $\mathbf{w}$ as a proxy for biological connectivity. In particular, we will use a sparse matrix with as many rows as there are pairs of neighboring voxels in the image; for a pair $\{i, j\}$, the corresponding row has zero entries everywhere expect for the $i^{\text{th}}$and j^{th} column, which have entries -1 and 1, respectively. To simplify notation in subsequent sections, we re-write eq. (3) in the form

$$p(\mathbf{w}|\boldsymbol{\alpha}, \lambda) = \frac{|\mathbf{P}|^{1/2}}{(2\pi)^{M/2}} \exp(-\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{P}\mathbf{w}), \tag{4}$$

which shows that the prior is a zero-mean Gaussian with inverse covariance $\mathbf{P} = \mathrm{diag}(\boldsymbol{\alpha}) + \lambda\mathbf{L}$, where $\mathbf{L} = \mathbf{\Gamma}^{\mathrm{T}}\mathbf{\Gamma}$ is also known as the *Laplacian matrix* in graph theory.

While our choice of $\mathbf{\Gamma}$ here simply penalizes spatial gradients in $\mathbf{w}$, it is worth noting that more advanced measures of biological connectivity can easily be integrated into the model as well – each with its own hyper-parameter that is automatically determined. Examples of such measures might include left-right symmetry relationships, as well as voxel-to-voxel connectivity strengths derived from functional image studies or based on detailed anatomical segmentations.

A. Training

Given a set of N training images $\mathbf{x}_n, n = 1, \ldots, N$ with corresponding target values $t_n, n = 1, \ldots, N$, we can determine the appropriate form of the regression function by estimating the hyper-parameters that maximize the so-called *marginal likelood function*, which expresses how probable the observed training data is for different settings of the hyper-parameters. Figure 1 shows the graphical model which depicts the dependency relationship between the variables. Collecting all the training images in the $N \times M$ "design" matrix $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]^{\mathrm{T}}$, and the corresponding target values in the vector $\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$, the marginal likelihood function is obtained by integrating out the weight parameters, yielding [41]

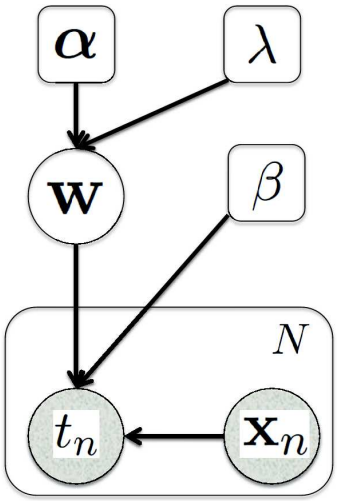

Fig. 1. Graphical representation of the regression model with N training subjects. Random variables are in circles and parameters are in squares. Shaded variables are observed. The plate indicates replication of N times.

$$\begin{aligned} p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \lambda, \beta) &= \int_{\mathbf{w}} \left(\prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}, \beta) \right) p(\mathbf{w}|\boldsymbol{\alpha}, \lambda) \mathrm{d}\mathbf{w} \\ &= \mathcal{N}(\mathbf{t}|\mathbf{0}, \mathbf{C}), \end{aligned} \tag{5}$$

where

$$\mathbf{C} = \beta^{-1}\mathbf{I} + \mathbf{X}\mathbf{P}^{-1}\mathbf{X}^{\mathrm{T}}. \tag{6}$$

Our goal is now to maximize eq. (5) with respect to the hyper-parameters $\boldsymbol{\alpha}$, λ, and β – known in the literature as the *evidence procedure* [45], *type-II maximum likelihood estimation* [46], or *restricted maximum likelihood estimation* [47].

We follow a heuristic optimization strategy similar to the one proposed in [45], which has also been used to train RVM regression models [44]. In particular, we maximize $\ln p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \lambda, \beta)$ – which is equivalent to maximizing $p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \lambda, \beta)$ but computationally more convenient – by observing that its derivatives to the hyper-parameters are given by the following expressions (see Appendix A for detailed derivations):

$$\begin{aligned} &\frac{\partial \ln p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \lambda, \beta)}{\partial \alpha_i} \\ &= \frac{1}{2\alpha_i}\left(1 - \alpha_i\Sigma_{ii} - \lambda\left(\mathbf{P}^{-1}\mathbf{L}\right)_{ii} - \alpha_i\mu_i^2\right) \end{aligned} \tag{7}$$

$$\begin{aligned} &\frac{\partial \ln p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \lambda, \beta)}{\partial \beta} \\ &= \frac{1}{2}\left(\frac{N}{\beta} - \mathrm{trace}\left(\mathbf{X}\mathbf{\Sigma}\mathbf{X}^{\mathrm{T}}\right) - \|\mathbf{t} - \mathbf{X}\boldsymbol{\mu}\|^2\right) \end{aligned} \tag{8}$$

$$\begin{aligned} &\frac{\partial \ln p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \lambda, \beta)}{\partial \lambda} \\ &= -\frac{1}{2}\left(\mathrm{trace}\Big(\left(\mathbf{\Sigma} - \mathbf{P}^{-1}\right)\mathbf{L}\Big) + \boldsymbol{\mu}^{\mathrm{T}}\mathbf{L}\boldsymbol{\mu}\right), \end{aligned} \tag{9}$$

where Σ_{ii} is the i^{th} diagonal component of the matrix

$$\boldsymbol{\Sigma} = (\beta \mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{P})^{-1} \tag{10}$$

and μ_i the i^{th} component of the vector

$$\boldsymbol{\mu} = \beta \boldsymbol{\Sigma}\mathbf{X}^{\mathrm{T}}\mathbf{t}. \tag{11}$$

Because their derivatives are zero at a maximum of the objective function, one strategy of optimizing for $\boldsymbol{\alpha}$ and β is to equate eq. (7) and (8) to zero and re-arranging, yielding the following re-estimation equations:

$$\alpha_i^{\text{new}} = \frac{1 - \alpha_i \boldsymbol{\Sigma}_{ii} - \lambda \left(\mathbf{P}^{-1}\mathbf{L}\right)_{ii}}{\mu_i^2} \tag{12}$$

and

$$\beta^{\text{new}} = \frac{N - \text{trace}(\beta \mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^{\mathrm{T}})}{\|\mathbf{t} - \mathbf{X}\boldsymbol{\mu}\|^2}. \tag{13}$$

In Appendix B, we show that both eq. (12) and (13) are guaranteed to yield non-negative α_i^{new} and β^{new}.

For the hyper-parameter λ, we use a standard gradient-ascent approach, yielding the following update equation:

$$\lambda^{\text{new}} \quad = \quad \lambda - \kappa \left(\text{trace}\Big(\left(\boldsymbol{\Sigma} - \mathbf{P}^{-1}\right) \mathbf{L}\Big) + \boldsymbol{\mu}^{\mathrm{T}}\mathbf{L}\boldsymbol{\mu}\right) \tag{14}$$

where κ is an appropriate step-size.

Training now proceeds by choosing initial values for the hyper-parameters $\boldsymbol{\alpha}$, β, and λ, and then iteratively re-computing $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ (eq. (10) and (11)) and the hyper-parameters (eq. (12), (13), and (14)), each in turn, until convergence. In our implementation, we initialize with $\alpha_i = 1$, $\forall i$, $\lambda = 1$, and $\beta = 10/\text{variance}(\mathbf{t})$. We monitor the value of the objective function $\ln p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \beta, \lambda)$ at each iteration, and terminate when the change over the previous iteration is below a certain tolerance. Section IV provides detailed pseudo-code for this algorithm, optimized for the computational and memory requirements of image-sized problems.

Although we have no theoretical guarantees that the proposed update equations for the hyper-parameters improve the objective function at each iteration, our experiments indicate that this is indeed the case.

B. Prediction

Once we have learned suitable hyper-parameters $\boldsymbol{\alpha}^*$, λ^*, and β^* from the training data, we can make predictions about the target variable t for a new input image $\mathbf{x}$ by evaluating the predictive distribution

$$\begin{aligned} & p(t|\mathbf{x}, \mathbf{X}, \mathbf{t}, \boldsymbol{\alpha}^*, \lambda^*, \beta^*) \\ & \quad = \quad \int_{\mathbf{w}} p(t|\mathbf{x}, \mathbf{w}, \beta^*) p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \boldsymbol{\alpha}^*, \lambda^*) \mathrm{d}\mathbf{w}, \end{aligned} \tag{15}$$

where $p(t|\mathbf{x}, \mathbf{w}, \beta^*)$ is given by eq. (1) and

$$\begin{aligned} & p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \boldsymbol{\alpha}^*, \lambda^*) \\ & \quad = \quad \frac{\left(\prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}, \beta^*)\right) p(\mathbf{w}|\boldsymbol{\alpha}^*, \lambda^*)}{p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}^*, \lambda^*, \beta^*)} \\ & \quad = \quad \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*) \end{aligned} \tag{16}$$

is the posterior distribution over the voxel weights. $\boldsymbol{\Sigma}^*$ and $\boldsymbol{\mu}^*$ are defined by eq. (10) and (11), in which $\boldsymbol{\alpha}$, λ, and β have been set to their optimized values.

It can be shown that the predictive distribution of eq. (15) is a Gaussian with mean

$$\boldsymbol{\mu}^{*\mathrm{T}}\mathbf{x} \tag{17}$$

and variance $1/\beta^* + \mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^*\mathbf{x}$ [41]. In practice, we will therefore use eq. (17) for making predictions, i.e., the linear regression model of eq. (2) where the voxel weights $\mathbf{w}$ have been set to $\boldsymbol{\mu}^*$. As we shall see in section V, most of these weights will typically be zero with the remaining voxels appearing in spatially clustered patches, immediately highlighting which image areas are driving the predictions.

III. RVoxM for Classification

In image-based binary classification, the aim is to predict a binary variable $b \in \{0, 1\}$ from an individual image $\mathbf{x}$. As in regression, we define the linear model

$$y(\mathbf{x}; \mathbf{w}) = \mathbf{w}^{\mathrm{T}}\mathbf{x}, \tag{18}$$

but transform the output by a logistic sigmoid function

$$\sigma(y) = \frac{1}{1 + \exp(-y)} \tag{19}$$

to map it into the interval $[0, 1]$. We can then use $\sigma(y(\mathbf{x}; \mathbf{w}))$ to represent the probability that $b = 1$ (Bernoulli distribution):

$$p(b|\mathbf{x}, \mathbf{w}) = \sigma(y(\mathbf{x}; \mathbf{w}))^b (1 - \sigma(y(\mathbf{x}; \mathbf{w})))^{1-b}, \tag{20}$$

and complete the model by using the same prior on $\mathbf{w}$ as in the regression case (eq. (3)). Note that, unlike in the regression model, there is no hyper-parameter β for the noise variance here.

Training the classification model entails estimating the hyper-parameters that maximize the marginal likelihood function

$$p(\mathbf{b}|\mathbf{X}, \boldsymbol{\alpha}, \lambda) = \int_{\mathbf{w}} p(\mathbf{b}|\mathbf{X}, \mathbf{w}) p(\mathbf{w}|\boldsymbol{\alpha}, \lambda) \mathrm{d}\mathbf{w}, \tag{21}$$

where

$$p(\mathbf{b}|\mathbf{X}, \mathbf{w}) = \prod_{n=1}^{N} p(b_n|\mathbf{x}_n, \mathbf{w})$$

and $\mathbf{b} = (b_1, \ldots, b_N)^{\mathrm{T}}$ contains the known, binary outcomes for all the training images $\mathbf{x}_n, n = 1, \ldots, N$. In contrast to the regression case, the integration over $\mathbf{w}$ cannot be evaluated analytically, and we need to resort to approximations. In Appendix C, we show that around a current hyper-parameter estimate $\{\tilde{\boldsymbol{\alpha}}, \tilde{\lambda}\}$, we can map the classification problem to a regression one:

$$\ln p(\mathbf{b}|\boldsymbol{\alpha}, \lambda) \simeq \ln \mathcal{N}(\tilde{\mathbf{t}}|\mathbf{0}, \tilde{\mathbf{C}}) + \text{const}, \tag{22}$$

where we have defined a covariance matrix

$$\tilde{\mathbf{C}} = \tilde{\mathbf{B}}^{-1} + \mathbf{X}\mathbf{P}^{-1}\mathbf{X}^{\mathrm{T}}$$

and local regression "target variables"

$$\tilde{\mathbf{t}} = \mathbf{X}\tilde{\mathbf{w}}_{\text{MP}} + \tilde{\mathbf{B}}^{-1}(\mathbf{b} - \tilde{\boldsymbol{\sigma}})$$

with the inverse variances of subject-specific regression "noise" $\tilde{\beta}_n = \tilde{\sigma}_n(1-\tilde{\sigma}_n)$ collected in the diagonal matrix $\tilde{\mathbf{B}} = \mathrm{diag}(\tilde{\beta}_1, \ldots, \tilde{\beta}_N)$, and $\tilde{\sigma}_n = \sigma(\mathbf{x}_n^{\mathrm{T}}\tilde{\mathbf{w}}_{\mathrm{MP}})$ and $\tilde{\boldsymbol{\sigma}} = (\tilde{\sigma}_1 \ldots, \tilde{\sigma}_N)^{\mathrm{T}}$. In these equations, $\tilde{\mathbf{w}}_{\mathrm{MP}}$ are the "most probable" voxel weights given the hyper-parameters $\{\tilde{\boldsymbol{\alpha}}, \tilde{\lambda}\}$:

$$\tilde{\mathbf{w}}_{\mathrm{MP}} = \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathbf{X}, \mathbf{b}, \tilde{\boldsymbol{\alpha}}, \tilde{\lambda}), \tag{23}$$

which involves solving a concave optimization problem. As detailed in Appendix C, we use Newton's method to perform this optimization.

Using the local mapping of eq. (22), learning the hyper-parameters now proceeds by iteratively using the regression update equations (12) and (14), but with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ defined as

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\mathbf{X}^{\mathrm{T}}\tilde{\mathbf{B}}\tilde{\mathbf{t}}. \tag{24}$$

and

$$\boldsymbol{\Sigma} = (\mathbf{X}^{\mathrm{T}}\tilde{\mathbf{B}}\mathbf{X} + \mathbf{P})^{-1}. \tag{25}$$

Once we have learned the hyper-parameters $\boldsymbol{\alpha}^*$ and λ^* this way, we can make predictions about the target variable b for a new input image $\mathbf{x}$ by evaluating the predictive distribution

$$\begin{aligned} & p(b=1|\mathbf{x}, \mathbf{X}, \mathbf{t}, \boldsymbol{\alpha}^*, \lambda^*) \\ &= \int_{\mathbf{w}} p(b=1|\mathbf{x}, \mathbf{w}) p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \boldsymbol{\alpha}^*, \lambda^*) \mathrm{d}\mathbf{w} \\ &\simeq \sigma(\tau \boldsymbol{\mu}^{*\mathrm{T}}\mathbf{x}), \end{aligned} \tag{26}$$

where $\tau = (1 + \pi\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^*\mathbf{x}/8)^{-1/2}$, and $\boldsymbol{\mu}^*$ and $\boldsymbol{\Sigma}^*$ are defined by eq. (24) and (25) in which the hyper-parameters have been set to their optimized values. The approximation in eq. (26) is based on the so-called Laplace approximation and on the similarity between the logistic sigmoid function and the probit function; see [41], pp. 217–220, for details. Eq. (26) can be thresholded at 0.5 to obtain a discrete prediction.

IV. Implementation

In most applications where RVoxM will be useful, the number of voxels to consider (M) is so large (e.g., 10^4 or 10^5) that a naive implementation of the proposed training update equations is computationally prohibitive – computing $\boldsymbol{\Sigma}$ alone already involves inverting a dense $M \times M$ matrix, which can take $\mathcal{O}(M^3)$ time.

One approach to alleviate the computational burden is to exploit the sparsity of the matrix $\mathbf{P}$ and use Woodbury's matrix identity [48] to compute $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\Sigma} = \mathbf{P}^{-1} - \mathbf{Z}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{Z}, \tag{27}$$

where $\mathbf{Z} = \mathbf{X}\mathbf{P}^{-1}$. Since $\mathbf{P}$ is sparse (with a the number of nonzero entries in each row being independent of M), the complexity of computing $\mathbf{P}^{-1}$ and subsequently $\mathbf{Z}$ is $\mathcal{O}(M^2)$ and $\mathcal{O}(MN)$, respectively. A naive computation of $\mathbf{C}$ using eq. (6) is $\mathcal{O}(M^2N)$. Yet re-writing $\mathbf{C}$ as

$$\mathbf{C} = \beta^{-1}\mathbf{I} + \mathbf{Z}\mathbf{X}^{\mathrm{T}}, \tag{28}$$

reduces the complexity of computing $\mathbf{C}$ to $\mathcal{O}(N^2M)$. Inverting $\mathbf{C}$ is $\mathcal{O}(N^3)$. Putting all this together, we can compute $\boldsymbol{\Sigma}$ using eq. (27) in $\mathcal{O}(M^2 + MN + N^2M + N^3)$ time. Since the number of available training subjects (N) is typically in the hundreds at best, in practice this means a reduction in computation time from $\mathcal{O}(M^3)$ to $\mathcal{O}(M^2)$.

Since M is so large, even an $\mathcal{O}(M^2)$ complexity is still a heavy computational burden. In practice, however, many of the α_i's tend to grow very large, effectively switching off the contribution of the corresponding voxels. We therefore resort to the type of greedy algorithm originally used for RVM training [44], whereby once a voxel has been switched off (i.e., its α_i has become larger than some threshold – in our implementation 10^{12} – it gets permanently discarded from the remaining computations. This provides a significant acceleration of the learning algorithm, as gradually more and more voxels are pruned from the model. To see how voxels can be removed from the computations, consider that $\mathbf{P}_{ii} \to \infty$ if $\alpha_i \to \infty$, and, as a result, the i'th row and i'th column of $\mathbf{P}^{-1}$ and $\boldsymbol{\Sigma}$ become zero vectors and $\mu_i \to 0$. Consequently, the update equations for the hyper-parameters are unaffected by simply deleting the i'th column from $\mathbf{X}$, and both the i'th column and the i'th row from $\boldsymbol{\Sigma}$ and $\mathbf{P}^{-1}$.

Finally, rather than manipulating the dense $M \times M$ matrices $\boldsymbol{\Sigma}$ and $\mathbf{P}^{-1}$ in their entirety, it is possible to compute their relevant contributions only one row at a time, avoiding the need to explicitly store such prohibitively large matrices.

Algorithm IV.1 provides pseudo-code for a RVoxM training procedure that has been optimized along the lines described above. For the classification case, a subroutine that optimizes for $\mathbf{w}_{\mathrm{MP}}$ is given in Algorithm IV.2.

V. Experiments

In order to illustrate the ability of RVoxM to yield informative models that predict well, we here present two experiments using T1-weighted structural magnetic resonance imaging (MRI) scans. The first experiment aims at predicting a subject's age from a volumetric gray matter segmentation (i.e., a regression scenario), whereas the second one focuses on discriminating Alzheimer's patients from healthy controls using surface-based cortical thickness measurements (illustrating a classification application).

A. Predicting Age

Both the structure and function of the human brain undergo significant changes over a person's life-time, and these changes can be detected using neuroimaging [49], [50]. Image-based prediction methods for estimating an individual's age from a brain MRI scan have attracted recent attention [17], [29], [51] since they provide a novel perspective for studying healthy development and aging patterns, while characterizing pathologic deviations in disease. In the current experiment, we employed the publicly available cross-sectional OASIS dataset [52], which consists of 436 subjects aged 18 to 96. For each subject, 3 or 4 individual T1-weighted MRI scans acquired in single scan sessions were averaged to obtain a single high-quality image. The subjects are all right-handed and include both men and women. 100 of the subjects over the age of 60 have been clinically diagnosed with very mild to moderate Alzheimer's disease (AD).

Algorithm IV.1: RVOXM_TRAINING($\mathbf{L}, \mathbf{X}, \mathbf{t}$ **or** $\mathbf{b}$)

comment: Input $\mathbf{t}$ for regression, $\mathbf{b}$ for classification

comment: $\mathbf{A}_{i.}$ and $\mathbf{A}_{.i}$ denote the i'th row and column of a matrix $\mathbf{A}$, respectively

comment: $\mathbf{e}_{i.}$ denotes a vector of all zeros except the i'th entry, which is a one

$\text{cost\_tol} \leftarrow 10^{-5}, \alpha_{max} = 10^{12}$
$\text{cost} \leftarrow \infty,\ \text{prev\_cost} \leftarrow \infty$
$\text{RelevantVoxels} \leftarrow \{1, \ldots, M\}$
$\lambda \leftarrow 1,\ \boldsymbol{\alpha} \leftarrow \mathbf{1}$
if *regression*
  then $\beta \leftarrow 10/\text{variance}(\mathbf{t})$
$\text{iter} \leftarrow 0$
repeat
 $\text{iter} \leftarrow \text{iter} + 1$
 if *classification*
  then $\begin{cases} \mathbf{w}_{\text{MP}} \leftarrow \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathbf{X}, \mathbf{b}, \boldsymbol{\alpha}, \lambda) & \textbf{comment:} \text{Call Algorithm IV.2.} \\ \mathbf{B} \leftarrow \text{diag}(\beta_1, \ldots, \beta_N) & \textbf{comment:} \beta_n = \sigma_n(1-\sigma_n) \text{ and } \sigma_n = \sigma(\mathbf{X}_{n.}\mathbf{w}) \\ \mathbf{t} \leftarrow \mathbf{X}\mathbf{w}_{\text{MP}} + \mathbf{B}^{-1}(\mathbf{b} - \boldsymbol{\sigma}) & \textbf{comment:} \boldsymbol{\sigma} = (\sigma_1 \ldots, \sigma_2)^{\text{T}} \end{cases}$
  else $\mathbf{B} \leftarrow \beta\mathbf{I}$
 $\mathbf{P} \leftarrow \text{diag}(\boldsymbol{\alpha}) + \lambda\mathbf{L}$
 comment: $\mathbf{P}$ and $\mathbf{B}$ are stored as sparse matrices
 $\mathbf{Z} \leftarrow \mathbf{X}\mathbf{P}^{-1}$ **comment:** Compute one row at a time by solving $\mathbf{P}\mathbf{Z}_{i.}^{\text{T}} = \mathbf{X}_{i.}^{\text{T}}$
 $\mathbf{C} \leftarrow \mathbf{B}^{-1} + \mathbf{Z}\mathbf{X}^{\text{T}}$
 Compute and save $\mathbf{C}^{-1}$
 $\text{DeleteVoxels} \leftarrow \{\}$
 $\text{a} \leftarrow 0$
 for each $i \in \text{RelevantVoxels}$
  do $\begin{cases} \text{Compute } (\mathbf{P}^{-1})_{i.} & \textbf{comment:} \text{Compute by solving } \mathbf{P}(\mathbf{P}^{-1})_{i.} = \mathbf{e}_i \\ \boldsymbol{\Sigma}_{i.} \leftarrow (\mathbf{P}^{-1})_{i.} - \mathbf{Z}_{i.}\mathbf{C}^{-1}\mathbf{Z} \\ \text{a} \leftarrow \text{a} + \left(\boldsymbol{\Sigma}_{i.} - (\mathbf{P}^{-1})_{i.}\right)\mathbf{L}_{.i} \\ \mu_i \leftarrow \boldsymbol{\Sigma}_{i.}\mathbf{X}^{\text{T}}\mathbf{B}\mathbf{t} \\ \alpha_i \leftarrow \frac{1-\alpha_i\boldsymbol{\Sigma}_{ii}-\lambda(\mathbf{P}^{-1})_{i.}\mathbf{L}_{.i}}{\mu_i^2} \\ \textbf{if } (\alpha_i > \alpha_{max}) \\ \quad \textbf{then } \text{DeleteVoxels} \leftarrow \text{DeleteVoxels} + \{i\} \end{cases}$
 $\text{prev\_cost} \leftarrow \text{cost}$
 $\text{cost} \leftarrow -\frac{1}{2}\ln|\mathbf{C}| - \frac{1}{2}\mathbf{t}^{\text{T}}\mathbf{C}^{-1}\mathbf{t}$
 if *regression*
  then $\beta \leftarrow \frac{N - \text{trace}(\beta\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^{\text{T}})}{\|\mathbf{t} - \mathbf{X}\boldsymbol{\mu}\|^2}$
 $\kappa \leftarrow \frac{1}{\sqrt{\text{iter}}}$
 $\lambda \leftarrow \lambda - \kappa\left(\text{a} + \boldsymbol{\mu}^{\text{T}}\mathbf{L}\boldsymbol{\mu}\right)$
 $\text{RelevantVoxels} \leftarrow \text{RelevantVoxels} - \text{DeleteVoxels}$
 $\mathbf{X} \leftarrow \mathbf{X}_{-\text{DeleteVoxels}}, \quad \mathbf{L} \leftarrow \mathbf{L}_{-\text{DeleteVoxels}}, \quad \boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha}_{-\text{DeleteVoxels}}$
 comment: Columns and rows corresponding to deleted voxels are removed.
until $(|(\text{prev\_cost} - \text{cost})/(\text{cost})| < \text{cost\_tol})$

Algorithm IV.2: OPTIMIZE_ $\mathbf{w}(\mathbf{X}, \mathbf{b}, \mathbf{w}, \mathbf{P}, \text{RelevantVoxels})$

$\text{tol} \leftarrow 0.01$
$\mathbf{Z} \leftarrow \mathbf{X}\mathbf{P}^{-1}$
repeat
 $\mathbf{w}^{\text{old}} \leftarrow \mathbf{w}$
 $\mathbf{B} \leftarrow \text{diag}(\beta_1, \ldots, \beta_N)$ **comment:** $\beta_n = \sigma_n(1-\sigma_n)$ and $\sigma_n = \sigma(\mathbf{X}_{n.}\mathbf{w})$
 $\mathbf{C} \leftarrow \mathbf{B}^{-1} + \mathbf{Z}\mathbf{X}^{\text{T}}$
 Compute and save $\mathbf{C}^{-1}$
 $\mathbf{s} \leftarrow \mathbf{X}\boldsymbol{\sigma}$
 for each $i \in \text{RelevantVoxels}$
  do $\begin{cases} \text{Compute } (\mathbf{P}^{-1})_{i.} \\ \boldsymbol{\Sigma}_{i.} \leftarrow (\mathbf{P}^{-1})_{i.} - \mathbf{Z}_{i.}\mathbf{C}^{-1}\mathbf{Z} \\ w_i \leftarrow w_i + \boldsymbol{\Sigma}_{i.}(\mathbf{X}\mathbf{b} - \mathbf{s} - \mathbf{P}\mathbf{w}) \end{cases}$
until $\|\mathbf{w}^{\text{old}} - \mathbf{w}\| < \text{tol}$

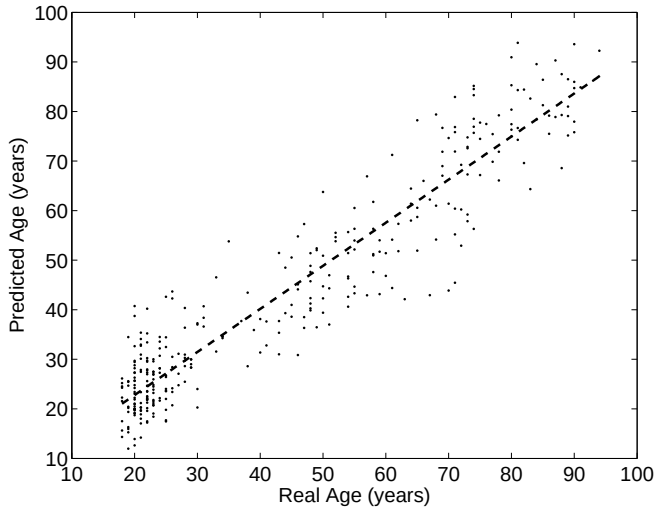

Fig. 2. RVoxM based predicted age versus real age in a cohort of 336 cognitively healthy subjects.

We processed all the MRI scans with SPM8[1], using default settings, to obtain spatially aligned gray matter maps for each subject. Briefly, the SPM software performs a simultaneous registration and segmentation of each MRI volume [53], aligning the image non-linearly with a standard template while at the same time computing each voxel's probability of belonging to different tissue types, such as gray or white matter. The resulting gray matter probability maps are then spatially transferred over to the coordinates of the template and modulated by the Jacobian of the non-linear transformations, yielding the so-called *gray matter density maps* commonly analyzed in voxel-based morphometry (VBM) studies [1].

Unsmoothed gray matter density values were used as the voxel-level measurements x_i in the present experiment. The average gray matter density volume computed on the training data was thresholded at 50% to obtain a mask of voxels that went into the analysis. On the analyzed data, there were an approximate total of 75k voxels in the mask. We employed a 6-connectivity neighborhood to define the Laplacian matrix.

To assess generalization accuracy and stability, we performed 5-fold cross-validation on the data from the cognitively normal and healthy subjects ($N = 336$, 43.7 ± 23.8 years, 62.5% female). In each cross-validation session, four fifths of the data were used for training an RVoxM. This model was then applied to the remaining fifth for testing. Each training session took about 100 CPU hours with our Matlab implementation, on a Xeon 5472 3.0GHz CPU.

Figure 2 shows the predicted age versus the real age for each subject. Note that each subject was treated as a test subject in only one of the 5 cross-validation sessions; the figure shows the predictions pooled across the sessions. The correlation between the real vs. the predicted age is 0.94, and the root mean square error (RMSE) is less than 7.9 years. It is interesting to note that the deviation from the best fit line seems to increase for older subjects who are beyond middle-age. This is likely driven by latent pathology, as recent studies have estimated that up to 30% of cognitively normal elderly subjects are actually at the pre-clinical stages of Alzheimer's disease [54].

Figure 3 illustrates the "relevance voxels" – those voxels that have non-zero contribution in the final prediction model – across the five training sessions. It can be appreciated that most voxels have a zero contribution (i.e., the model is sparse), and that the relevance voxels occur in clusters, providing clear clues as to what parts of the gray matter are driving the age prediction process. Furthermore, the relevance voxels exhibit an overall consistent pattern across the five training sessions, as can be deduced from the yellow regions in the bottom row of Figure 3, thus providing evidence that these patterns are likely to be associated with the underlying biology and can be interpreted. The relevance patterns include peri-sylvian areas (e.g., Heschl's gyrus) as well as deep structures (e.g., thalamus), and are in broad agreement with published aging-associated morphology maps (e.g., [55]).

In addition to RVoxM, we also tested two other methods as benchmarks. The first method, referred to as "RVM", was specifically proposed recently for estimating age from structural MRI [29]. It uses a principal component analysis (PCA) to achieve a dimensionality-reduced representation of the image data, and subsequently applies a linear RVM algorithm in the resulting feature space. We used the optimal implementation settings that were described in [29] and a public implementation of RVM[2]. The second benchmark ("RVoxM-NoReg") was an implementation of RVoxM with no spatial regularization, i.e., with the hyper-parameter λ intentionally clamped to zero. A comparison with the latter benchmark gives us an insight into the effect of spatial regularization on the results.

Figure 4 plots the average RMSE for all three algorithms (top), as well as the average difference between the individual-level prediction errors (square of predicted age minus true age) obtained by RVoxM and the other two methods (bottom). Overall, RVoxM yields the best accuracy with a RMSE less than 7.9 years – the difference between RVoxM's performance and the other two benchmarks is statistically significant (paired t-test, $P < 0.05$). RVoxM also attains the highest correlation (r-value) between the subjects' real age and predicted age among all three methods: 0.94 for RVoxM vs. 0.9 and 0.93 for RVM and RVoxM-NoReg, respectively. We note that [29] reported a slightly better correlation value for RVM ($r = 0.92$), which is probably due to the increased sample size (410 training subjects instead of the 268 training subjects used here).

Finally, we also examined the deviation of the predicted "brain age" from the real age particularly in elderly subjects. Recent work on a young cohort attributed such a deviation observed in fMRI data to the nonlinear development trajectory of the brain [17]. Moreover, neuroimaging studies on dementia and Alzheimer's have suggested that these diseases might *accelerate* atrophy in the brain [56]. As such, we hypothesized that the mini mental state examination (MMSE) score, a cognitive assessment that predicts dementia, may explain some

[1]http://www.fil.ion.ucl.ac.uk/spm/software/spm8/

[2]http://www.vectoranomaly.com/downloads/downloads.htm

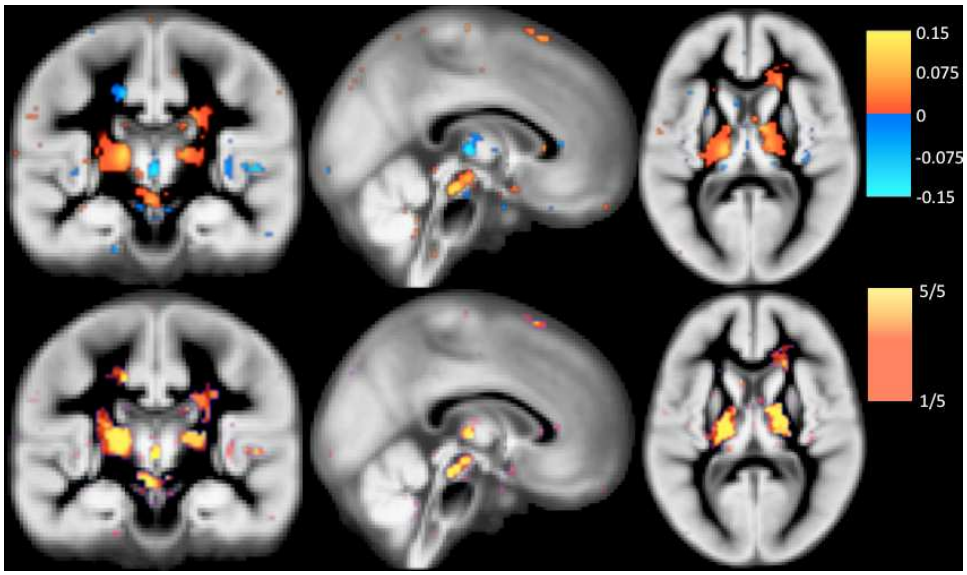

Fig. 3. Relevance voxels for predicting age, overlaid on the average gray matter density image across all subjects. Top row: The $\boldsymbol{\mu}^*$ map (eq. (11) in which the hyper-parameters have been set to their optimized values) averaged across 5 cross-validation sessions. Voxels with zero weight are transparent. Bottom row: The frequency at which a voxel was selected as being relevant (i.e., receiving a non-zero weight in a training session) across the 5 cross-validation sessions.

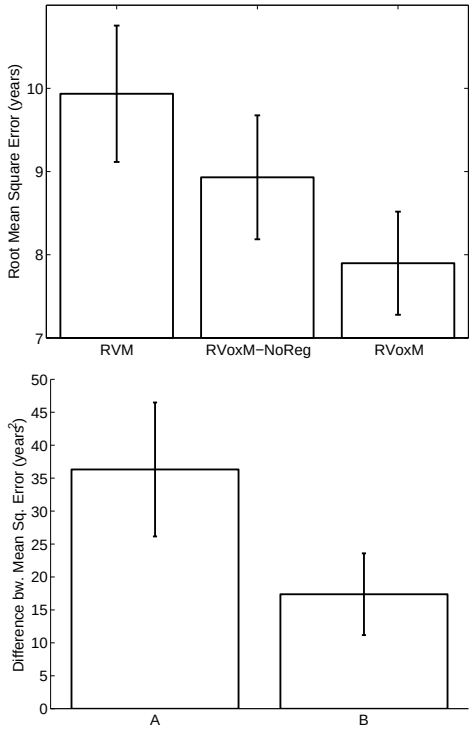

Fig. 4. Top: average root mean square error for the three age regression models. Bottom: average difference between subject-level prediction errors, measured as the square of real age minus predicted age. (A) Error of RVM minus error of RVoxM. (B) Error of RVoxM-NoReg minus error of RVoxM. Error bars show the standard error of the mean.

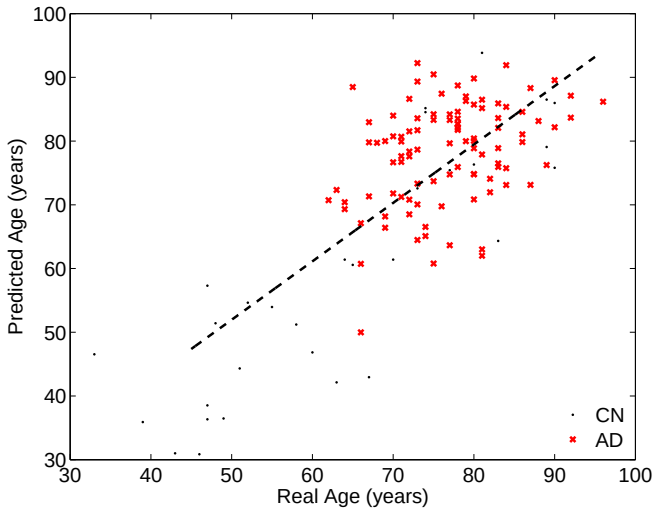

Fig. 5. RVoxM based predicted age versus real age in a cohort of 30 cognitively healthy subjects and 100 AD patients.

of the non-linear behavior in the predicted "brain age" of elderly subjects. To test this hypothesis, we used the RVoxM from the first of the 5-fold cross-validation experiment, which was trained on 268 cognitively healthy subjects. We applied this RVoxM model to predict the "brain age" of 100 AD patients and 30 cognitively healthy elderly subjects from the test group (see Figure 5). Note that none of these subjects were used to train the RVoxM and we excluded 33 young healthy subjects, for which we did not have an MMSE score. We then conducted a linear regression analysis, where the predicted age was treated as the outcome variable and real age, MMSE and sex were the independent variables. Both the real age (coefficient: 0.84, P-val $< 10^{-22}$) and the MMSE score (coefficient: -0.77, P-val$< 10^{-4}$) were independently associated with the predicted age, but the subject's sex was not. This suggests that pathological processes that are reflected as cognitive decline might explain some of the deviation in the predicted brain age.

B. Predicting Alzheimer's Diagnosis

Here we demonstrate RVoxM as a classifier for discriminating healthy controls from AD patients based on their brain MRI scans. Instead of working with volumetric MRI data, we implemented RVoxM on a 2D surface model of the cerebral cortex, further demonstrating the versatility of the proposed algorithm. We applied RVoxM to the publicly available Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset[3], which consisted of 818 subjects at the time of writing. At recruitment, 229 subjects were categorized as cognitively healthy; 396 subjects as amnestic Mild Cognitive Impairment (MCI) – a transitionary, clinically defined pre-AD stage; and 193 subjects as AD. All subjects were clinically followed up every six months, starting from a baseline clinical assessment. Each follow-up visit typically included a cognitive assessment and a structural MRI scan. In the present experiment, we only analyzed baseline MRI scans. We processed all MRI scans with the FreeSurfer software suite [57], [58], computing subject-specific models of the cortical surface as well as thickness measurements across the entire cortical mantle [4]. Subject-level thickness measurements were then transferred to a common coordinate system, represented as a icosohedron-based triangulation of the sphere, via a surface-based nonlinear registration procedure [59], and analyzed by RVoxM. We utilized the so-called *fsaverage6* representation, consisting of approximately 82,000 vertices across the two hemispheres with an inter-vertex distance of approximately 2 mm. We emphasize that we did not smooth these cortical thickness maps for any of our experiments. The matrix $\mathbf{L}$ for the spatial regularization was obtained by using the neighborhood structure of the triangulated mesh.

Our analysis used MRI scans from 150 AD patients (75.1 ± 7.4 years, 47% female), and an age and sex-matched control group (N=150; cognitively normal (CN); 76.1 ± 5.8 years; 47% female)[4]. As in the age prediction experiment, we conducted a five-fold cross-validation, where each clinical group was first divided into five subgroups. During each fold, one AD and one CN subgroup were set aside as the test set and the RVoxM classification algorithm was trained on the remaining subjects, which took around 110 CPU hours (Matlab implementation, Xeon 5472 3.0GHz CPU). The obtained classification model was then tested on the corresponding test group. The presented results are combined across all five training/test sessions.

For comparison, we also implemented the following four benchmark algorithms:

1) RVoxM-NoReg: Similar to the regression experiment, we implemented the RVoxM classifier with the spatial regularization intentionally switched off, i.e., with $\lambda = 0$.
2) SVM: We also applied a linear support vector machine (SVM) classifier, a demonstrated state-of-the-art AD classification method [39], to the cortical thickness maps. For this purpose, we used the popular SVM implementation provided by the freely available LIBSVM software package [60].
3) SVM-Seg: We applied the same linear SVM implementation to thickness measurements in 70 automatically segmented cortical subregions. In particular, we used FreeSurfer to parcellate the entire cortical mantle based on the underlying sulcal pattern [61], computed a list of average thickness measurements for each of the resulting subregions, and used these as the attribute vector for the SVM.
4) RVM-Seg: Finally, we also applied an implementation of the RVM binary classifier[5] with a linear kernel to the same thickness measurements of the 70 cortical ROIs used for SVM-Seg.

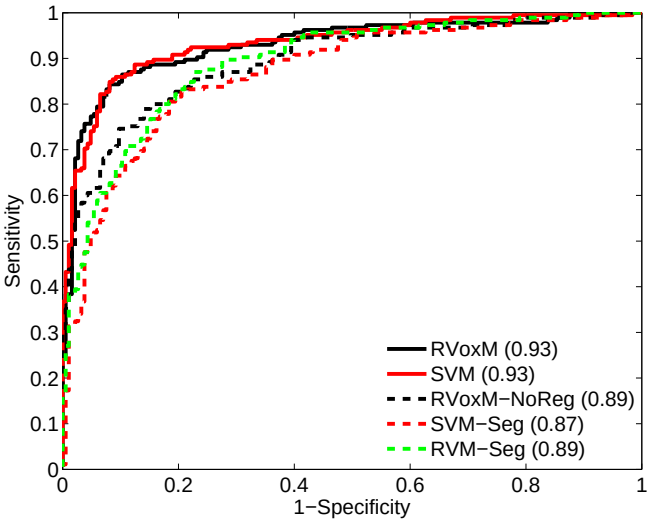

Fig. 6. The receiver operating characteristics curve for five classifiers discriminating between AD patients and controls. Area under the curve values are listed in parentheses in the legend. See text for a description of the methods.

Figure 6 shows the receiver operating characteristics (ROC) curve for RVoxM and the four benchmark algorithms. The ROC curves were generated by varying a threshold applied to the continuous prediction score that each of the algorithms generates (eq. (26) for RVoxM). For each threshold value, we computed the specificity and sensitivity values on each test group corresponding to each of the five folds. These specificity and sensitivity values were then averaged to obtain the presented ROC curves. Based on the area under the ROC curve (AUC), SVM (93%) and RVoxM (93%) perform the best for discriminating AD patients from healthy controls.

There is a clear difference between RVoxM and RVoxM-NoReg (AUC: 89%), which once again underscores the significance of incorporating the spatial smoothness term into the model. Although SVM and RVoxM have a similar classification performance, it is worth emphasizing that SVM uses all 82,000 mesh vertices simultaneously to make its predictions, complicating the interpretation of its underlying models.

Figure 7 illustrates the RVoxM "relevance vertices" that play a role in discriminating the two clinical groups based on thickness measurements. This figure shows the average $\boldsymbol{\mu}^*$

[3]For detailed information, visit http://www.adni-info.org/

[4]We selected the first 150 AD patients that were successfully processed with FreeSurfer.

[5]http://www.vectoranomaly.com/downloads/downloads.htm

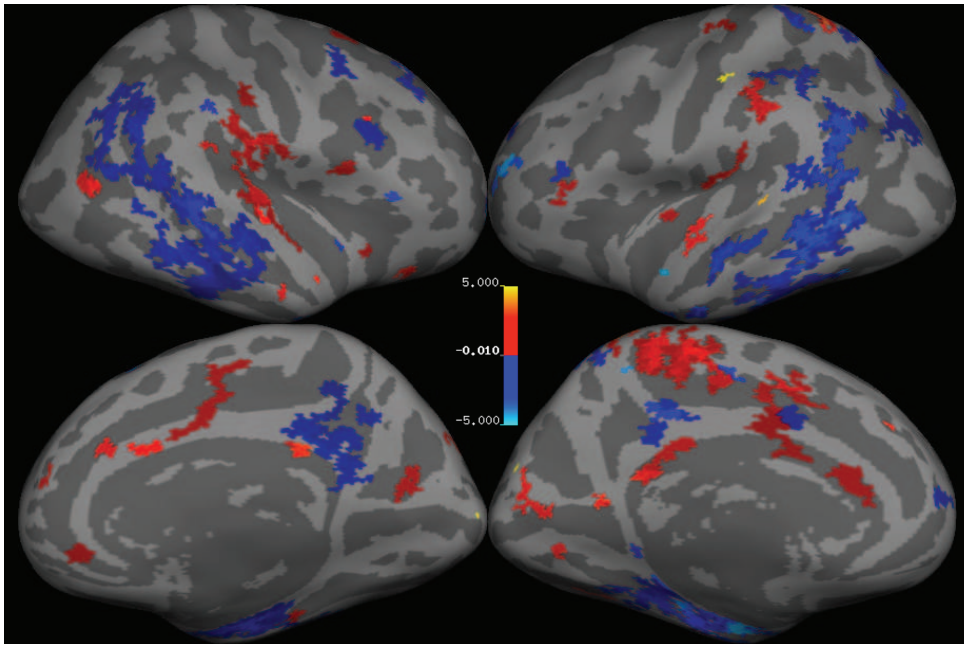

Fig. 7. Relevance "voxels" (or more accurately: vertices) for AD vs. control discrimination. The weights for the classification model, $\boldsymbol{\mu}^*$ (eq. (24) in which the hyper-parameters have been set to their optimized values), averaged across 5 cross-validation sessions, are illustrated on an inflated cortical representation, overlaid on the folding pattern of the FreeSurfer template subject. Blue regions have a negative contribution (assuming AD is positive) and red regions exhibit a positive weight. Voxels with zero contribution are transparent.

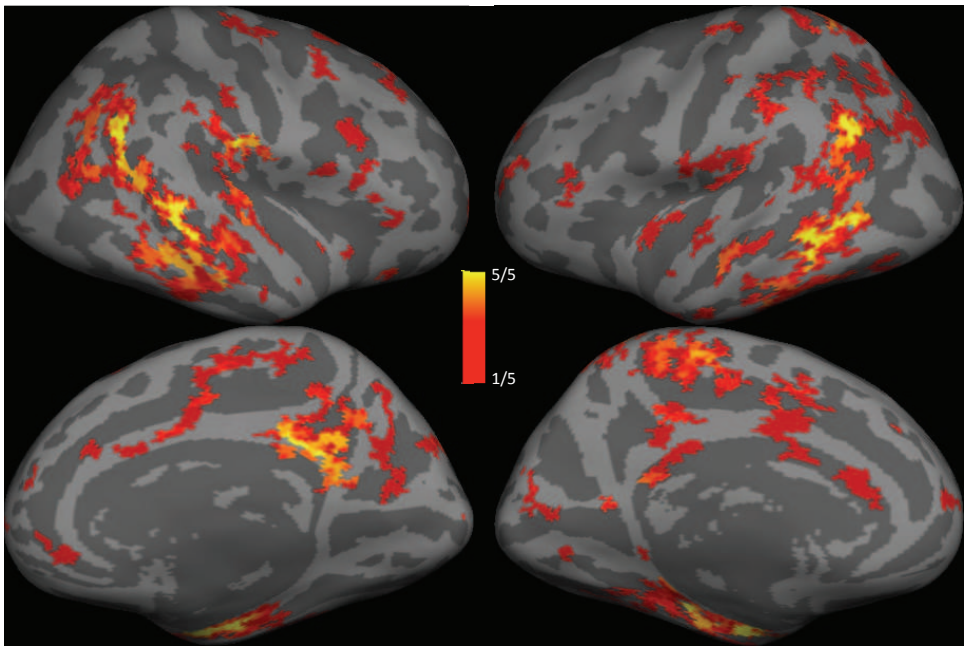

Fig. 8. The frequency at which each vertex had a non-zero contribution in the final model across the five cross-validation sessions. Transparent voxels never had a non-zero contribution.

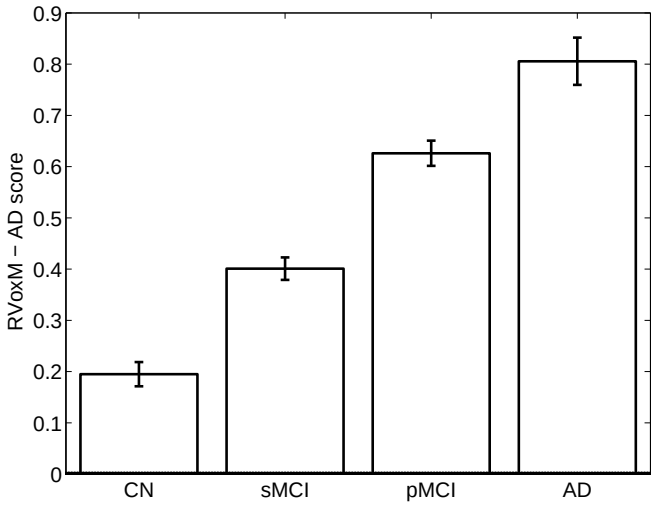

Fig. 9. Average RVoxM-based AD score for four groups in ADNI. Error bars indicate standard error of the mean.

values, where the average was taken across the five cross-validation sessions. Similar to the regression result, one can appreciate that a large number of vertices have a zero contribution (i.e., the model is sparse), and that the relevance vertices appear in spatial clusters. Figure 8 shows the frequency at which each relevance vertex was selected (i.e., had a non-zero contribution) across the five cross-validation sessions. There are certain regions that consistently contribute across the different training sessions (in particular those colored yellow). These vertices include the entorhinal cortex, superior temporal sulcus and posterior cingulate/precuneus, and overlap the so-called default network regions that are known to be targeted in AD [62]. However, there are other regions (mostly in shades of red) that are less consistent and are chosen only in one or two training sessions. We will discuss the possible causes of such model instabilities, as well as ways to mitigate their effect, in section VI.

We also used the RVoxM model from the first of the 5-fold cross-validation training sessions to compute an "AD score" (eq. (26) for all remaining ADNI subjects, which we subdivided into four groups: (1) cognitively normal, healthy controls, who remained so throughout the study and who were not included in the training data set ($N = 109$); (2) subjects with MCI who were stable and did not progress to dementia (sMCI, $N = 221$); (3) subjects who had MCI at baseline but then progressed to AD (pMCI; $N = 159$); and (4) AD patients ($N = 73$). Figure 9 plots the average AD score for each of these groups, computed from their baseline MRI scans. We observe that, at baseline, the stable MCI group has an average AD score less than 0.5 and therefore appears more "control-like", whereas the progressive MCI group has a more "AD-like" average AD score that is greater than 0.5. These results suggest that an RVoxM based classification of someone's MRI scan might be informative for predicting future clinical decline. To test this hypothesis directly, we conducted a survival analysis with a Cox regression model [63] on all MCI subjects combined ($N = 380$), where the outcome of interest was time-to-diagnosis. Age, sex, education (years), APOE $\epsilon 4$ allele count, APOE $\epsilon 3$ allele count and the RVoxM-based AD score were entered as independent variables. The only variable that was associated with time-to-diagnosis was the RVoxM-based AD score (coefficient: 0.66, P-val $< 10^{-3}$). This results suggests that a baseline MRI scan contains predictive information about future clinical decline and this information is, to some extent, extracted by the RVoxM AD classifier.

VI. Discussion and Conclusion

In this paper, we presented the Relevance Voxel Machine (RVoxM), a novel Bayesian model for image-based prediction that is designed to yield intuitive and interpretable results. It allows the predictive influence of individual voxels to vary, and to be more similar in biologically related areas than in completely unrelated ones. Bayesian analysis is then used to select the appropriate form of the model based on annotated training data. As demonstrated in our experiments, RVoxM yields models that are *sparse* and *spatially smooth* when spatial proximity is used as a measure of biological connectivity. We believe that such models are easier to interpret than models that use all the image voxels simultaneously, or that base their predictions on a set of isolated voxels scattered throughout the image area. Importantly, our experiments also indicate that RVoxM automatically avoids over-fitting to the training data and produces excellent predictions on test data.

Compared to other prediction models used in medical image analysis, RVoxM offers the following advantages:

- Regularization, feature selection, and biological consistency within a single algorithm;
- Automatic tuning of all parameters, i.e., no free parameters to set manually or via cross-validation; *and*
- Probabilistic classification predictions, rather than binary decision outcomes.

Although we only applied RVoxM to structural gray matter morphometry in this paper, the method is general and can be extended to handle multiple tissue types at the same time; analyze functional or metabolic imaging modalities; or include non-imaging sources of information such as blood tests, cerebrospinal fluid (CSF) markers, and genetic or demographic data [28]. Furthermore, one can easily incorporate more advanced measures of biological connectivity than the simple spatial smoothness prior used in our experiments. Connectivity information based on symmetry (if the two hemispheres are expected to have similar contributions) or obtained from functional or diffusion imaging can be added by including extra terms "$\|\mathbf{\Gamma w}\|^2$" in eq. (3), with corresponding hyper-parameters that will then be automatically learned from the training data as well.

When discussing the properties of RVoxM, it is useful to consider the training-phase optimization of its hyper-parameters within an ideal Bayesian framework, which would not involve any optimization at all. For the sake of clarity, we will concentrate on the regression case only, although similar arguments apply to the classification case as well. Letting $\boldsymbol{\eta} = (\ln \alpha_1, \ldots, \ln \alpha_M, \ln \lambda, \ln \beta)^{\mathrm{T}}$ denote the collection of

log-transformed hyper-parameters[6], and assuming a uniform prior distribution on $\boldsymbol{\eta}$: $p(\boldsymbol{\eta}) \propto 1$, the true Bayesian predictive distribution over the target variable t for a new input image $\mathbf{x}$ is given by

$$p(t|\mathbf{x}, \mathbf{X}, \mathbf{t}) \quad = \quad \int_{\boldsymbol{\eta}} \int_{\mathbf{w}} p(t, \mathbf{w}, \boldsymbol{\eta}|\mathbf{x}, \mathbf{X}, \mathbf{t}) \mathrm{d}\mathbf{w} \mathrm{d}\boldsymbol{\eta},$$

which involves integrating over both $\mathbf{w}$ and $\boldsymbol{\eta}$. RVoxM effectively performs the integration over $\mathbf{w}$ analytically, while approximating the remaining integral over $\boldsymbol{\eta}$, assuming it is dominated by the optimal hyper-parameter value $\boldsymbol{\eta}^* = \arg\max_{\boldsymbol{\eta}} p(\boldsymbol{\eta}|\mathbf{X}, \mathbf{t})$:

$$\begin{aligned} p(t|\mathbf{x}, \mathbf{X}, \mathbf{t}) &= \int_{\boldsymbol{\eta}} p(t|\mathbf{x}, \mathbf{X}, \mathbf{t}, \boldsymbol{\eta}) p(\boldsymbol{\eta}|\mathbf{X}, \mathbf{t}) \mathrm{d}\boldsymbol{\eta} \\ &\simeq p(t|\mathbf{x}, \mathbf{X}, \mathbf{t}, \boldsymbol{\eta}^*). \end{aligned} \tag{29}$$

RVoxM first estimates $\boldsymbol{\eta}^*$ by maximizing $p(\boldsymbol{\eta}|\mathbf{X}, \mathbf{t}) \propto p(\mathbf{t}|\mathbf{X}, \boldsymbol{\eta})$ (optimization of eq. (5)), and then uses the resulting distribution $p(t|\mathbf{x}, \mathbf{X}, \mathbf{t}, \boldsymbol{\eta}^*)$ to make predictions (eq. (15)).

The approximation of eq. (29) has the disadvantage that it gives rise to a high-dimensional, non-convex optimization problem, putting RVoxM at risk of local optima and other convergence issues [64]. As demonstrated in [65], these problems can be avoided by approximating the integral over $\boldsymbol{\eta}$ using Monte Carlo sampling instead: given enough samples from the posterior $p(\boldsymbol{\eta}|\mathbf{X}, \mathbf{t})$, the resulting predictive distribution can be made arbitrarily close to the true one. Although theoretically superior to RVoxM, this approach will only be computationally feasible when a small subset of potentially relevant image voxels are somehow selected *a priori* [65], limiting its appeal in practical settings.

The ideal Bayesian prediction model that RVoxM approximates also helps explain why RVoxM tends to set many voxel weights to zero values, even though its prior (eq. (4)) may not seem to encourage such solutions. Writing

$$p(t|\mathbf{X}, \mathbf{t}, \mathbf{x}) \quad = \quad \int_{\mathbf{w}} p(t|\mathbf{X}, \mathbf{t}, \mathbf{x}, \mathbf{w}) p(\mathbf{w}|\mathbf{X}, \mathbf{t}) \mathrm{d}\mathbf{w}$$

reveals that the predictive distribution is obtained by adding contributions of all possible values of $\mathbf{w}$, each weighed by its posterior probability $p(\mathbf{w}|\mathbf{X}, \mathbf{t})$. Although the integral over $\mathbf{w}$ can not easily be approximated to obtain a practically useful algorithm [44], [66], the crucial insight is that the posterior $p(\mathbf{w}|\mathbf{X}, \mathbf{t}) \propto p(\mathbf{t}|\mathbf{X}, \mathbf{w}) p(\mathbf{w})$ will be high for $\mathbf{w}$'s with many zero entries, because the "true" prior

$$p(\mathbf{w}) = \int_{\boldsymbol{\eta}} p(\mathbf{w}|\boldsymbol{\eta}) p(\boldsymbol{\eta}) \mathrm{d}\mathbf{w}, \tag{30}$$

obtained by integrating out the hyper-parameters, encourages such solutions. Indeed, for the special case where the spatial smoothness hyper-parameter λ is clamped to zero but otherwise $p(\boldsymbol{\eta}) \propto 1$, eq. (30) evaluates to [44]:

$$p(\mathbf{w}) \propto \prod_i 1/|w_i|,$$

which is sharply peaked at zero for each voxel and therefore favors sparsity. This "true" prior can be compared to the so-called Laplace prior $p(\mathbf{w}) \propto \prod_i \exp(-|w_i|)$ often used to obtain sparsity in Bayesian models [67], or – taking the negative log – as the ℓ_1 norm $\sum_i |w_i|$ in the popular "lasso" regularized regression method [68].

RVoxM goes beyond merely inducing sparsity in the models by allowing non-zero values for the hyper-paramer λ, enforcing spatial consistency. This helps remedy the well-known problem with sparsity-only promoting methods that when several variables (i.e., voxels) have similar prediction power, only one tends to be picked with little regard as to which one [69], [70]. In order to avoid such overly sparse models, which hamper biological interpretation, a popular solution in regularized regression is the so-called "elastic net", which adds a ℓ_2 regularization term to the sparsity-inducing ℓ_1 regularizer of lasso [70]. In Bayesian approaches, proposed remedies include using hyper-parameters that optimize another objective function than the likelihood [69], or assuming voxels belong to a small set of clusters with common regularization [65], [71]. The way RVoxM addresses this issue is by expanding the family of candidate models that can be tried to explain the training data, relying on the fact that relatively simple models – with fewer degrees of freedom – tend to provide better explanations than overly complex ones [38]. By also allowing high values of λ, simple and therefore good models are no longer only those in which just a select few predictive voxels are in the model, but especially those in which neigboring, similarly predictive voxels are in the model *together*.

Because of the way it seeks sparse but spatially connected solutions, RVoxM is closely related to so-called "structured sparsity"-inducing methods, which aim at selecting problem-relevant *groups* of variables for inclusion in the model, rather than single variables individually [72]–[76]. In such methods, group-level sparsity is often obtained by variations on the so-called "group lasso", a generalization of lasso in which the ℓ_2 norm of each group, rather than the amplitude of individual variables, is penalized using the ℓ_1 norm [77], [78]. Perhaps most closely related to RVoxM is the so-called "smooth lasso" method, a variant of the elastic net in which the ℓ_1 norm for sparsity is preserved, but the ℓ_2 norm on the variables themselves is replaced by an ℓ_2 norm on their spatial derivatives [79].

An issue we have not fully addressed in this work is quantifying how repeatable the relevant voxel set is when the RVoxM model is trained on different subjects drawn from the same population. Although the relevant voxel pattern was quite consistent across different training datasets in our regression experiment (see bottom row of Figure 3), there was an appreciable amount of variation in the classification case (see Figure 8). We believe such variations can be further decreased by making more relevant anatomical information available to the RVoxM model – e.g., by including a symmetry regularization term in the prior. The stability of the relevant patterns and the predictions can also be improved by using randomization experiments, in which different models are learned from resampled training data to obtain an average, ensemble prediction model [80], or to select only those vox-

[6]It is natural to work with log-transformed values here, as the hyper-parameters are all positive (scale) parameters.

els that appear frequently across the different models [81], [82]. Instability of informative feature sets has been studied extensively in the literature [83], [84], and can be attributed to three related factors: (1) the limited amount of training data and over-fitting to the quirks of these data, (2) the mismatch between the utilized model and the underlying true discriminative pattern, and (3) the local optima the numerical solver might get trapped in. All three factors apply to the case of RVoxM, and a detailed analysis of these effects will be carried out in future work, using techniques similar to the ones employed in [85]–[87].

One drawback of the presented training algorithm is its computational complexity, which under typical conditions is quadratic in the number of voxels. Our experiments demonstrate that, using standard Matlab code, we can train on a dataset of relatively high-resolution data from hundreds subjects in a matter of days. This computation time, we believe, is acceptable for such datasets that can take years to collect. It is worth emphasizing that a heavy computational burden is incurred only once for a given training dataset and that, after the model has been trained, making predictions on new images is very fast. Since RVoxM automatically tunes all its hyper-parameters within a single training session, there is no need for the repeated cross-validation training runs that are necessary in most other image-based prediction methods and that also take time. Furthermore, more advanced regularization terms can be added to the prior of RVoxM with minimal additional computational cost, whereas the number of training runs required to set the corresponding hyper-parameters using cross-validation would increase exponentially and quickly become impractical.

Although the reported computation times can be reduced significantly by using a non-Matlab based implementation that exploits the parallellization opportunities inherent in Algorithm IV.1; classification problems with more than two classes, as well as higher-resolution and much larger datasets, will still present a serious computational challenge to analyze with RVoxM. The training algorithm we have presented starts with all voxels included in the initial model, and gradually prunes the vast majority of the voxels as the iterations progress. Although this causes the algorithm to gradually speed up, the computational complexity of the first few iterations is still quadratic in the number of voxels. Similar to the dramatically accelerated training procedure for RVM models developed in [88], we are therefore investigating an alternative, "constructive" approach that starts with an empty model and sequentially *adds* voxels instead, while also modifying the weights of the voxels already in the model. In [88], this was accomplished by deriving an analytical expression for the optimal weight of a voxel, given the current weight of all other voxels; we are currently exploring if a similar approach is also possible for RVoxM.

VII. Acknowledgements

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimers Association; Alzheimers Drug Discovery Foundation; Amorfix Life Sciences Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health Rev March 26, 2012 (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129 and K01 AG030514.

Appendix A
Derivations for the Regression Model

We here derive various partial derivatives of $\ln p(\mathbf{t}|\mathbf{X},\boldsymbol{\alpha},\beta,\lambda)$ with respect to the hyper-parameters $\boldsymbol{\alpha}$, β, and λ.

Using the determinant identity

$$|\beta^{-1}\mathbf{I}+\mathbf{X}\mathbf{P}^{-1}\mathbf{X}^{\mathrm{T}}| = \frac{|\beta^{-1}\mathbf{I}||\beta\mathbf{X}^{\mathrm{T}}\mathbf{X}+\mathbf{P}|}{|\mathbf{P}|}$$

and Woodbury's inversion identity

$$\left(\beta^{-1}\mathbf{I}+\mathbf{X}\mathbf{P}^{-1}\mathbf{X}^{\mathrm{T}}\right)^{-1} = \beta\mathbf{I}-\beta\mathbf{X}\left(\beta\mathbf{X}^{\mathrm{T}}\mathbf{X}+\mathbf{P}\right)^{-1}\mathbf{X}^{\mathrm{T}}\beta,$$

we can write $\ln p(\mathbf{t}|\mathbf{X},\boldsymbol{\alpha},\beta,\lambda)$ as:

$$\begin{aligned}\ln p(\mathbf{t}|\mathbf{X},\boldsymbol{\alpha},\beta,\lambda) &= -\frac{N}{2}\ln(2\pi)-\frac{1}{2}\ln|\mathbf{C}|-\frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{t}\\ &= -\frac{N}{2}\ln(2\pi)+\frac{N}{2}\ln\beta-\frac{1}{2}\ln|\boldsymbol{\Sigma}^{-1}|+\\ &\quad \frac{1}{2}\ln|\mathbf{P}|-\\ &\quad \frac{1}{2}\mathbf{t}^{\mathrm{T}}\left(\beta\mathbf{I}-\beta\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^{\mathrm{T}}\beta\right)\mathbf{t}. \end{aligned} \quad (31)$$

Using $\boldsymbol{\Sigma}^{-1}=\beta\mathbf{X}^{\mathrm{T}}\mathbf{X}+\mathrm{diag}(\boldsymbol{\alpha})+\lambda\mathbf{L}$, we obtain

$$\begin{aligned}\frac{\partial\ln|\boldsymbol{\Sigma}^{-1}|}{\partial\alpha_i} &= \mathrm{trace}\left(\boldsymbol{\Sigma}\frac{\partial\mathrm{diag}(\boldsymbol{\alpha})}{\partial\alpha_i}\right)\\ &= \Sigma_{ii}, \end{aligned} \quad (32)$$

and similarly

$$\begin{aligned}\frac{\partial\ln|\boldsymbol{\Sigma}^{-1}|}{\partial\beta} &= \mathrm{trace}\left(\boldsymbol{\Sigma}\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\\ &= \mathrm{trace}\left(\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^{\mathrm{T}}\right) \end{aligned} \quad (33)$$

and

$$\frac{\partial \ln |\boldsymbol{\Sigma}^{-1}|}{\partial \lambda} = \mathrm{trace}(\boldsymbol{\Sigma}\mathbf{L}). \quad (34)$$

Using the same technique on $\mathbf{P} = \mathrm{diag}(\boldsymbol{\alpha}) + \lambda\mathbf{L}$, we have

$$\frac{\partial \ln |\mathbf{P}|}{\partial \alpha_i} = \left(\mathbf{P}^{-1}\right)_{ii} \quad (35)$$

and

$$\frac{\partial \ln |\mathbf{P}|}{\partial \lambda} = \mathrm{trace}(\mathbf{P}^{-1}\mathbf{L}) \quad (36)$$

Finally, we have

$$\begin{aligned}\frac{\partial \left(\mathbf{t}^{\mathrm{T}}\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^{\mathrm{T}}\mathbf{t}\right)}{\partial \alpha_i} &= \frac{\partial \left(\mathbf{t}^{\mathrm{T}}\mathbf{X}\left(\boldsymbol{\Sigma}^{-1}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{t}\right)}{\partial \alpha_i} \\ &= -\mathbf{t}^{\mathrm{T}}\mathbf{X}\boldsymbol{\Sigma}\frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \alpha_i}\boldsymbol{\Sigma}\mathbf{X}^{\mathrm{T}}\mathbf{t} \\ &= -\frac{1}{\beta}\boldsymbol{\mu}^{\mathrm{T}}\frac{\partial \mathrm{diag}(\boldsymbol{\alpha})}{\partial \alpha_i}\boldsymbol{\mu}\frac{1}{\beta} \\ &= -\frac{\mu_i^2}{\beta^2}, \end{aligned} \quad (37)$$

and similarly

$$\frac{\partial \left(\mathbf{t}^{\mathrm{T}}\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^{\mathrm{T}}\mathbf{t}\right)}{\partial \beta} = \frac{-1}{\beta^2}\boldsymbol{\mu}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\mu} \quad (38)$$

and

$$\frac{\partial \left(\mathbf{t}^{\mathrm{T}}\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^{\mathrm{T}}\mathbf{t}\right)}{\partial \lambda} = -\frac{1}{\beta^2}\boldsymbol{\mu}^{\mathrm{T}}\mathbf{L}\boldsymbol{\mu}. \quad (39)$$

To obtain eq. (7), we take the partial derivative of eq. (31) with respect to α_i and plug in the result of eq. (32), (35), and (37), yielding

$$\frac{\partial \ln p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \lambda, \beta)}{\partial \alpha_i} = -\frac{1}{2}\Sigma_{ii} + \frac{1}{2}\left(\mathbf{P}^{-1}\right)_{ii} - \frac{1}{2}\mu_i^2. \quad (40)$$

Rewriting $\mathbf{P}^{-1}$ using Woodbury's inversion identity as

$$\begin{aligned}\mathbf{P}^{-1} &= \mathrm{diag}(\frac{1}{\boldsymbol{\alpha}}) - \\ &\quad \mathrm{diag}(\frac{1}{\boldsymbol{\alpha}})\left(\mathbf{I} + \lambda\mathbf{L}\mathrm{diag}(\frac{1}{\boldsymbol{\alpha}})\right)^{-1}\lambda\mathbf{L}\mathrm{diag}(\frac{1}{\boldsymbol{\alpha}}) \\ &= \mathrm{diag}(\frac{1}{\boldsymbol{\alpha}}) - \lambda\mathbf{P}^{-1}\mathbf{L}\mathrm{diag}(\frac{1}{\boldsymbol{\alpha}}) \end{aligned} \quad (41)$$

and plugging this result into eq. (40), we finally obtain eq. (7).

Taking the partial derivative of eq. (31) with respect to β and plugging in eq. (33) and (38) yields

$$\begin{aligned}\frac{\partial \ln p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha}, \lambda, \beta)}{\partial \beta} &= \frac{N}{2\beta} - \frac{1}{2}\mathrm{trace}\left(\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^{\mathrm{T}}\right) - \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{t} + \\ &\quad \mathbf{t}^{\mathrm{T}}\mathbf{X}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\mu} \end{aligned} \quad (42)$$

which explains eq. (8). Similarly, we obtain eq. (9) by taking the partial derivative of eq. (31) with respect to λ and plugging in eq. (34), (36), and (39).

APPENDIX B
NON-NEGATIVE PROPERTIES OF THE UPDATE RULES

We here show that the update equations (12) and (13) always yield non-negative values.

For eq. (12), we have:

$$\begin{aligned}\alpha_i^{\mathrm{new}} &= \frac{1 - \alpha_i\boldsymbol{\Sigma}_{ii} - \lambda\left(\mathbf{P}^{-1}\mathbf{L}\right)_{ii}}{\mu_i^2} \\ &= \frac{\alpha_i(\mathbf{P}^{-1} - \boldsymbol{\Sigma})_{ii}}{\mu_i^2} \quad (43) \\ &= \frac{\alpha_i(\mathbf{P}^{-1} - \mathbf{P}^{-1} + \mathbf{Z}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{Z})_{ii}}{\mu_i^2} \quad (44) \\ &= \frac{\alpha_i}{\mu_i^2}\mathbf{z}_i^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{z}_i \geq 0, \end{aligned}$$

where we have used eq. (41) to obtain eq. (43), and expanded $\boldsymbol{\Sigma}$ using eq. (27) to obtain eq. (44). $\mathbf{C}$ is the positive semi-definite matrix defined in eq. (6) and $\mathbf{z}_i$ is the i'th column of $\mathbf{Z} = \mathbf{X}\mathbf{P}^{-1}$.

For eq. (13) we have:

$$\begin{aligned}\beta^{\mathrm{new}} &= \frac{N - \mathrm{trace}(\beta\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^{\mathrm{T}})}{\|\mathbf{t} - \mathbf{X}\boldsymbol{\mu}\|^2} \\ &= \frac{N - \mathrm{trace}(\beta\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\Sigma})}{\|\mathbf{t} - \mathbf{X}\boldsymbol{\mu}\|^2} \\ &= \frac{N - \mathrm{trace}\left((\boldsymbol{\Sigma}^{-1} - \mathbf{P})\boldsymbol{\Sigma}\right)}{\|\mathbf{t} - \mathbf{X}\boldsymbol{\mu}\|^2} \\ &= \frac{\mathrm{trace}(\mathbf{P}\boldsymbol{\Sigma})}{\|\mathbf{t} - \mathbf{X}\boldsymbol{\mu}\|^2} \\ &= \frac{\mathrm{trace}(\mathbf{S}^{\mathrm{T}}\mathbf{P}\mathbf{S})}{\|\mathbf{t} - \mathbf{X}\boldsymbol{\mu}\|^2} \\ &= \frac{\sum_i \mathbf{s}_i^{\mathrm{T}}\mathbf{P}\mathbf{s}_i}{\|\mathbf{t} - \mathbf{X}\boldsymbol{\mu}\|^2} \geq 0, \end{aligned} \quad (45)$$

where we have used $\boldsymbol{\Sigma} = (\mathbf{P} + \beta\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$, $\boldsymbol{\Sigma} = \mathbf{S}\mathbf{S}^{\mathrm{T}}$ is the Cholesky decomposition of $\boldsymbol{\Sigma}$, $\mathbf{s}_i$ is the i'th column of $\mathbf{S}$ and the inequality is due to $\mathbf{P}$ being positive semi-definite.

APPENDIX C
DERIVATIONS FOR THE CLASSIFICATION MODEL

We here explain how we compute the most probable voxel weights in the classification model (eq. (23)) and locally approximate the classification training problem by a regression one (eq. (22)).

For a given set of hyper-parameters $\{\boldsymbol{\alpha}, \lambda\}$, we compute the voxel weights $\mathbf{w}_{\mathrm{MP}}$ maximizing the posterior distribution $p(\mathbf{w}|\mathbf{X}, \mathbf{b}, \boldsymbol{\alpha}, \lambda) \propto p(\mathbf{b}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha}, \lambda)$ by using Newton's method, i.e., by repeatedly performing

$$\mathbf{w}^{\mathrm{new}} = \mathbf{w} - \Big(\nabla\nabla \ln p(\mathbf{w}|\mathbf{X}, \mathbf{b}, \boldsymbol{\alpha}, \lambda)\Big)^{-1}\nabla \ln p(\mathbf{w}|\mathbf{X}, \mathbf{b}, \boldsymbol{\alpha}, \lambda)$$

until convergence, with gradient

$$\nabla \ln p(\mathbf{w}|\mathbf{X}, \mathbf{b}, \boldsymbol{\alpha}, \lambda) = \mathbf{X}^{\mathrm{T}}(\mathbf{b} - \boldsymbol{\sigma}) - \mathbf{P}\mathbf{w}$$

and Hessian matrix

$$\nabla\nabla \ln p(\mathbf{w}|\mathbf{X}, \mathbf{b}, \boldsymbol{\alpha}, \lambda) = -\left(\mathbf{X}^{\mathrm{T}}\mathbf{B}\mathbf{X} + \mathbf{P}\right),$$

where we have defined $\boldsymbol{\sigma} = (\sigma_1 \ldots, \sigma_N)^{\mathrm{T}}$, $\sigma_n = \sigma(\mathbf{x}_n^{\mathrm{T}}\mathbf{w})$, $\mathbf{B} = \mathrm{diag}(\beta_1, \ldots, \beta_N)$, and $\beta_n = \sigma_n(1-\sigma_n)$. Since the Hessian is always positive definite, $\ln p(\mathbf{w}|\mathbf{X}, \mathbf{b}, \boldsymbol{\alpha}, \lambda)$ is concave and therefore has a unique maximum [41].

Once the optimum weights $\mathbf{w}_{\mathrm{MP}}$ are obtained, we approximate the integral in eq. (21) by replacing the integrand with an unnormalized Gaussian centered around $\mathbf{w}_{\mathrm{MP}}$ (Laplace approximation), yielding:

$$\begin{aligned} &\ln p(\mathbf{b}|\mathbf{X}, \boldsymbol{\alpha}, \lambda) \\ &\quad\simeq \ln\left[p(\mathbf{b}|\mathbf{X}, \mathbf{w}_{\mathrm{MP}}) p(\mathbf{w}_{\mathrm{MP}}|\boldsymbol{\alpha}, \lambda)\sqrt{\frac{(2\pi)^M}{|\mathbf{H}|}}\right] \end{aligned} \quad (46)$$

for the log marginal likelihood, where we have defined

$$\mathbf{H} = -\nabla\nabla \ln p(\mathbf{w}|\mathbf{X}, \mathbf{b}, \boldsymbol{\alpha}, \lambda)\Big|_{\mathbf{w}=\mathbf{w}_{\mathrm{MP}}}.$$

Around the most probable voxel weights $\tilde{\mathbf{w}}_{\mathrm{MP}}$ corresponding to some hyper-parameters $\{\tilde{\boldsymbol{\alpha}}, \tilde{\lambda}\}$ (eq. (23)), we can linearize $\sigma(\mathbf{x}_n^{\mathrm{T}}\mathbf{w})$ as follows:

$$\sigma(\mathbf{x}_n^{\mathrm{T}}\mathbf{w}) \simeq \sigma(\mathbf{x}_n^{\mathrm{T}}\tilde{\mathbf{w}}_{\mathrm{MP}}) + \tilde{\beta}_n \mathbf{x}_n^{\mathrm{T}}\left(\mathbf{w} - \tilde{\mathbf{w}}_{\mathrm{MP}}\right),$$

and therefore

$$\boldsymbol{\sigma} \simeq \tilde{\boldsymbol{\sigma}} + \tilde{\mathbf{B}}\mathbf{X}(\mathbf{w} - \tilde{\mathbf{w}}_{\mathrm{MP}}).$$

As a result, we have that

$$\begin{aligned} \nabla \ln p(\mathbf{b}|\mathbf{X}, \mathbf{w}) &= \mathbf{X}^{\mathrm{T}}(\mathbf{b} - \boldsymbol{\sigma}) \\ &\simeq \mathbf{X}^{\mathrm{T}}(\mathbf{b} - \tilde{\boldsymbol{\sigma}} + \tilde{\mathbf{B}}\mathbf{X}\tilde{\mathbf{w}}_{\mathrm{MP}} - \tilde{\mathbf{B}}\mathbf{X}\mathbf{w}) \\ &= \mathbf{X}^{\mathrm{T}}\tilde{\mathbf{B}}(\tilde{\mathbf{t}} - \mathbf{X}\mathbf{w}) \end{aligned}$$

and therefore that

$$\ln p(\mathbf{b}|\mathbf{X}, \mathbf{w}) \simeq \ln \mathcal{N}(\tilde{\mathbf{t}}|\mathbf{X}\mathbf{w}, \tilde{\mathbf{B}}^{-1}) + \mathrm{const}, \quad (47)$$

where the constant depends only on $\tilde{\mathbf{w}}_{\mathrm{MP}}$. Using this result, we obtain

$$\mathbf{H} \simeq \mathbf{X}^{\mathrm{T}}\tilde{\mathbf{B}}\mathbf{X} + \mathbf{P} = \boldsymbol{\Sigma}^{-1}, \quad (48)$$

and because

$$\nabla p(\mathbf{w}|\mathbf{X}, \mathbf{b}, \boldsymbol{\alpha}, \lambda)\Big|_{\mathbf{w}=\mathbf{w}_{\mathrm{MP}}} = 0$$

also that

$$\mathbf{X}^{\mathrm{T}}\tilde{\mathbf{B}}(\tilde{\mathbf{t}} - \mathbf{X}\mathbf{w}_{\mathrm{MP}}) - \mathbf{P}\mathbf{w}_{\mathrm{MP}} \simeq 0$$

and therefore

$$\mathbf{w}_{\mathrm{MP}} \simeq \boldsymbol{\Sigma}\mathbf{X}^{\mathrm{T}}\tilde{\mathbf{B}}\tilde{\mathbf{t}}. \quad (49)$$

Plugging eq. (47) and (48) into (46), we have that

$$\begin{aligned} \ln p(\mathbf{b}|\mathbf{X}, \boldsymbol{\alpha}, \lambda) &\simeq -\frac{N}{2}\ln(2\pi) + \frac{1}{2}\ln|\tilde{\mathbf{B}}| - \\ &\quad \frac{1}{2}(\tilde{\mathbf{t}} - \mathbf{X}\mathbf{w})^{\mathrm{T}}\tilde{\mathbf{B}}(\tilde{\mathbf{t}} - \mathbf{X}\mathbf{w}) + \\ &\quad \frac{1}{2}\ln|\mathbf{P}| - \frac{1}{2}\mathbf{w}_{\mathrm{MP}}^{\mathrm{T}}\mathbf{P}\mathbf{w}_{\mathrm{MP}} - \\ &\quad \frac{1}{2}\ln|\boldsymbol{\Sigma}^{-1}| + \mathrm{const} \\ &\simeq -\frac{N}{2}\ln(2\pi) + \frac{1}{2}\ln|\tilde{\mathbf{B}}| - \\ &\quad \frac{1}{2}\ln|\boldsymbol{\Sigma}^{-1}| + \frac{1}{2}\ln|\mathbf{P}| - \\ &\quad \frac{1}{2}\mathbf{t}^{\mathrm{T}}(\tilde{\mathbf{B}} - \tilde{\mathbf{B}}\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^{\mathrm{T}}\tilde{\mathbf{B}})\mathbf{t} + \mathrm{const}, \end{aligned}$$

where we have used eq. (49) in the last step. Comparing this result to eq. (31) finally yields eq. (22).

REFERENCES

[1] J. Ashburner and K.J. Friston. Voxel-based morphometry–the methods. *Neuroimage*, 11(6):805–821, 2000.

[2] C. Davatzikos, A. Genc, D. Xu, and S.M. Resnick. Voxel-based morphometry using the RAVENS maps: methods and validation using simulated longitudinal atrophy. *NeuroImage*, 14(6):1361–1369, 2001.

[3] MK Chung, KJ Worsley, T. Paus, C. Cherif, DL Collins, JN Giedd, JL Rapoport, and AC Evans. A unified statistical approach to deformation-based morphometry. *NeuroImage*, 14(3):595–606, 2001.

[4] B. Fischl and A.M. Dale. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *PNAS*, 97(20):11050, 2000.

[5] KJ Worsley and KJ Friston. Analysis of fMRI time-series revisited - again. *Neuroimage*, 2(3):173–181, 1995.

[6] J.D. Haynes and G. Rees. Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7):523–534, 2006.

[7] T.M. Mitchell, R. Hutchinson, R.S. Niculescu, F. Pereira, X. Wang, M. Just, and S. Newman. Learning to decode cognitive states from brain images. *Machine Learning*, 57(1):145–175, 2004.

[8] J.V. Haxby, M.I. Gobbini, M.L. Furey, A. Ishai, J.L. Schouten, and P. Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425, 2001.

[9] J. Mourão-Miranda, A.L.W. Bokde, C. Born, H. Hampel, and M. Stetter. Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data. *NeuroImage*, 28(4):980–995, 2005.

[10] K.A. Norman, S.M. Polyn, G.J. Detre, and J.V. Haxby. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in cognitive sciences*, 10(9):424–430, 2006.

[11] N. Batmanghelich, B. Taskar, and C. Davatzikos. A general and unifying framework for feature construction, in image-based pattern classification. In *IPMI*, pages 423–434. Springer, 2009.

[12] C. Davatzikos, Y. Fan, X. Wu, D. Shen, and S.M. Resnick. Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging. *Neurobiology of aging*, 29(4):514–523, 2008.

[13] Y. Fan, D. Shen, and C. Davatzikos. Classification of structural images via high-dimensional image warping, robust feature extraction, and svm. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2005*, pages 1–8, 2005.

[14] S. Klöppel, C.M. Stonnington, C. Chu, B. Draganski, R.I. Scahill, J.D. Rohrer, N.C. Fox, C.R. Jack, J. Ashburner, and R.S.J. Frackowiak. Automatic classification of MR scans in Alzheimer's disease. *Brain*, 131(3):681, 2008.

[15] Y. Liu, L. Teverovskiy, O. Carmichael, R. Kikinis, M. Shenton, C.S. Carter, V.A. Stenger, S. Davis, H. Aizenstein, J.T. Becker, et al. Discriminative MR image feature analysis for automatic schizophrenia and Alzheimer's disease classification. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2004*, pages 393–401, 2004.

[16] K. Pohl and M. Sabuncu. A unified framework for MR based disease classification. In *Information Processing in Medical Imaging*, pages 300–313. Springer, 2009.

[17] N.U.F. Dosenbach, B. Nardos, A.L. Cohen, D.A. Fair, J.D. Power, J.A. Church, S.M. Nelson, G.S. Wig, A.C. Vogel, C.N. Lessov-Schlaggar, et al. Prediction of individual brain maturity using fMRI. *Science*, 329(5997):1358, 2010.

[18] J. Ashburner, C. Hutton, R. Frackowiak, I. Johnsrude, C. Price, and K. Friston. Identifying global anatomical differences: deformation-based morphometry. *Human Brain Mapping*, 6(5-6):348–357, 1998.

[19] P. Golland. Discriminative direction for kernel classifiers. *Advances in Neural Information Processing Systems*, 1:745–752, 2002.

[20] Z. Lao, D. Shen, Z. Xue, B. Karacali, S.M. Resnick, and C. Davatzikos. Morphological classification of brains via high-dimensional shape transformations and machine learning methods. *Neuroimage*, 21(1):46–57, 2004.

[21] P. Golland, W.E.L. Grimson, M.E. Shenton, and R. Kikinis. Detection and analysis of statistical differences in anatomical shape. *Medical Image Analysis*, 9(1):69–86, 2005.

[22] Y. Fan, D. Shen, R.C. Gur, R.E. Gur, and C. Davatzikos. COMPARE: classification of morphological patterns using adaptive regional elements. *IEEE Transactions on Medical Imaging*, 26(1):93–105, 2007.

[23] O. Yamashita, M. Sato, T. Yoshioka, F. Tong, and Y. Kamitani. Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. *NeuroImage*, 42:1414–1429, 2008.

[24] F. De Martino, G. Valente, N. Staeren, J. Ashburner, R. Goebel, and E. Formisano. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *NeuroImage*, 43:44–58, 2008.

[25] E. Janousova, M. Vounou, R. Wolz, K. Gray, D. Rueckert, and G. Montana1. Fast brain-wide search of highly discriminative regions in medical images: an application to Alzheimer's disease. In *Proceedings of Medical Image Understanding and Analysis 2011*, pages 17–21, 2011.

[26] B. Magnin, L. Mesrob, S. Kinkingnéhun, M. Pélégrini-Issac, O. Colliot, M. Sarazin, B. Dubois, S. Lehéricy, and H. Benali. Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. *Neuroradiology*, 51(2):73–83, 2009.

[27] R.S. Desikan, H.J. Cabral, C.P. Hess, W.P. Dillon, C.M. Glastonbury, M.W. Weiner, N.J. Schmansky, D.N. Greve, D.H. Salat, R.L. Buckner, and B. Fischl. Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease. *Brain*, 132:2048–2057, 2009.

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

[29] K. Franke, G. Ziegler, S. Kloppel, and C. Gaser. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters. *NeuroImage*, 50(3):883–892, 2010.

[30] L. Liang, V. Cherkassky, and D.A. Rottenberg. Spatial SVM for feature selection and fMRI activation detection. In *Neural Networks, 2006. IJCNN'06. International Joint Conference on*, pages 1463–1469. IEEE, 2006.

[31] G. Fung and J. Stoeckel. SVM feature selection for classification of SPECT images of Alzheimer's disease using spatial information. *Knowledge and Information Systems*, 11(2):243–258, 2007.

[32] Z. Xiang, Y. Xi, U. Hasson, and P. Ramadge. Boosting with spatial regularization. *Advances in Neural Information Processing Systems*, 22:2107–2115, 2009.

[33] R. Cuingnet, M. Chupin, H. Benali, and O. Colliot. Spatial and anatomical regularization of SVM for brain image analysis. *Advances in Neural Information Processing Systems*, 23:460–468, 2010.

[34] R. Cuingnet, C. Rosso, M. Chupin, S. Lehricy, D. Dormont, H. Benali, Y. Samson, and O. Colliot. Spatial regularization of SVM for the detection of diffusion alterations associated with stroke outcome. *Medical Image Analysis*, 15:729–737, 2011.

[35] V. Michel, A. Gramfort, G. Varoquaux, E. Eger, and B. Thirion. Total variation regularization for fMRI-based prediction of behaviour. *Medical Imaging, IEEE Transactions on*, 30(7):1328–1340, 2011.

[36] N. Batmanghelich, B. Taskar, and C. Davatzikos. Generative-discriminative basis learning for medical imaging. *IEEE Transactions on Medical Imaging*, 31(1):51–69, 2012.

[37] M.A.J. van Gerven, B. Cseke, F.P. de Lange, and T. Heskes. Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior. *NeuroImage*, 50:150–161, 2010.

[38] D. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[39] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehéricy, M.O. Habert, M. Chupin, H. Benali, and O. Colliot. 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, 2011.

[40] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[41] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[42] M.R. Sabuncu and K. Van Leemput. The Relevance Voxel Machine (RVoxM): A Bayesian method for image-based prediction. *Lecture Notes in Computer Science*, 6893:99–106, 2011. Proceedings of MICCAI2011.

[43] R.M. Neal. *Bayesian Learning for Neural Networks*. Number 118 in Lecture Notes in Statistics. Springer-Verlag, 1996.

[44] M.E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

[45] D.J.C. MacKay. *Bayesian Methods for Adaptive Models*. PhD thesis, California Institute of Technology, 1992.

[46] J. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 2 edition, 1985.

[47] D.A. Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, pages 320–338, 1977.

[48] M.A. Woodbury. Inverting modified matrices. *Memorandum report*, 42:106, 1950.

[49] M. D'Esposito, E. Zarahn, G.K. Aguirre, and B. Rypma. The effect of normal aging on the coupling of neural activity to the bold hemodynamic response. *Neuroimage*, 10(1):6–14, 1999.

[50] D.H. Salat, R.L. Buckner, A.Z. Snyder, D.N. Greve, R.S.R. Desikan, E. Busa, J.C. Morris, A.M. Dale, and B. Fischl. Thinning of the cerebral cortex in aging. *Cerebral Cortex*, 14(7):721, 2004.

[51] J. Ashburner. A fast diffeomorphic image registration algorithm. *Neuroimage*, 38(1):95–113, 2007.

[52] D.S. Marcus, T.H. Wang, J. Parker, J.G. Csernansky, J.C. Morris, and R.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*, 19(9):1498–1507, 2007.

[53] J. Ashburner and K.J. Friston. Unified segmentation. *Neuroimage*, 26(3):839–851, 2005.

[54] R.A. Sperling, P.S. Aisen, L.A. Beckett, D.A. Bennett, S. Craft, A.M. Fagan, T. Iwatsubo, C.R. Jack, J. Kaye, T.J. Montine, et al. Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging and the Alzheimer's Association workgroup. *Alzheimer's and Dementia*, 2011.

[55] C.D. Good, I.S. Johnsrude, J. Ashburner, R.N.A. Henson, K.J. Friston, and R.S.J. Frackowiak. A voxel-based morphometric study of ageing in 465 normal adult human brains. *Neuroimage*, 14(1):21–36, 2001.

[56] KA Jobst, AD Smith, M. Szatmari, MM Esiri, A. Jaskowski, N. Hindley, B. McDonald, and AJ Molyneux. Rapidly progressing atrophy of medial temporal lobe in Alzheimer's disease. *The Lancet*, 343(8901):829–830, 1994.

[57] A.M. Dale, B. Fischl, and M.I. Sereno. Cortical surface-based analysis I: Segmentation and surface reconstruction. *Neuroimage*, 9(2):179–194, 1999.

[58] B. Fischl, M.I. Sereno, and A.M. Dale. Cortical surface-based analysis II: Inflation, flattening, and a surface-based coordinate system. *Neuroimage*, 9(2):195–207, 1999.

[59] B. Fischl, M.I. Sereno, R.B.H. Tootell, and A.M. Dale. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human Brain Mapping*, 8(4):272–284, 1999.

[60] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[61] R.S. Desikan, F. Ségonne, B. Fischl, B.T. Quinn, B.C. Dickerson, D. Blacker, R.L. Buckner, A.M. Dale, R.P. Maguire, B.T. Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.

[62] R.L. Buckner, A.Z. Snyder, B.J. Shannon, G. LaRossa, R. Sachs, A.F. Fotenos, Y.I. Sheline, W.E. Klunk, C.A. Mathis, J.C. Morris, et al. Molecular, structural, and functional characterization of Alzheimer's disease: evidence for a relationship between default activity, amyloid, and memory. *The Journal of Neuroscience*, 25(34):7709, 2005.

[63] D.R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220, 1972.

[64] D. Wipf and S. Nagarajan. A new view of automatic relevance determination. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1625–1632. MIT Press, Cambridge, MA, 2008.

[65] V. Michel, E. Eger, C. Keribin, and B. Thirion. Multiclass sparse Bayesian regression for fMRI-based prediction. *Journal of Biomedical Imaging*, pages 2:1–2:13, January 2011.

[66] D.J.C. MacKay. Comparison of approximate methods for handling hyperparameters. *Neural Computation*, 11(5):1035–1068, 1999.

[67] P.M. Williams. Bayesian regularization and pruning using a Laplace prior. *Neural Computation*, 7(1):117–143, 1995.

[68] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[69] Y. Qi, T.P. Minka, R.W. Picard, and Z. Ghahramani. Predictive automatic relevance determination by expectation propagation. In *Proceedings of the 21st International Conference on Machine Learning*, pages 671–678, 2004.

[70] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320, 2005.

[71] K. Friston, C. Chu, J. Mourao-Miranda, O. Hulme, G. Rees, W. Penny, and J. Ashburner. Bayesian decoding of brain images. *NeuroImage*, 39:181–205, 2008.

[72] F. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. *Advances in Neural Information Processing Systems*, pages 105–112, 2008.
[73] B.P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37:3468–3497, 2009.
[74] J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. *Journal of Machine Learning Research*, 12:3371–3412, 2011.
[75] R. Jenatton, J.-Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research*, 12:2777–2824, 2011.
[76] R. Jenatton, A. Gramfort, V. Michel, G. Obozinski, F. Bach, and B. Thirion. Multi-scale mining of fMRI data with hierarchical structured sparsity. In *IEEE International Workshop on Pattern Recognition in NeuroImaging*, 2011.
[77] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67, 2006.
[78] F.R. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.
[79] M. Hebiri and S. Van De Geer. The smooth-lasso and other l1 + l2-penalized methods. *ArXiv*, 1003, 2010.
[80] Y. Wang, Y. Fan, P. Bhatt, and C. Davatzikos. High-dimensional pattern regression using machine learning: From medical images to continuous clinical variables. *NeuroImage*, 50:1519–1535, 2010.
[81] F. Bunea, Y. She, H. Ombao, A. Gongvatana, K. Devlin, and R. Cohen. Penalized least squares regression methods and applications to neuroimaging. *NeuroImage*, 55:1519–1527, 2011.
[82] G. Varoquaux, A. Gramfort, and B. Thirion. Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.
[83] Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507, 2007.
[84] J. Hua, W.D. Tembe, and E.R. Dougherty. Performance of feature-selection methods in the classification of high-dimension data. *Pattern Recognition*, 42(3):409–424, 2009.
[85] J. Himberg and A. Hyvärinen. Icasso: software for investigating the reliability of ICA estimates by clustering and visualization. In *Neural Networks for Signal Processing, 2003. NNSP'03. 2003 IEEE 13th Workshop on*, pages 259–268. IEEE, 2003.
[86] Y. Saeys, T. Abeel, and Y. Van de Peer. Robust feature selection using ensemble feature selection techniques. *Machine Learning and Knowledge Discovery in Databases*, pages 313–325, 2008.
[87] C. Plant, S.J. Teipel, A. Oswald, C. Böhm, T. Meindl, J. Mourao-Miranda, A.W. Bokde, H. Hampel, and M. Ewers. Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease. *Neuroimage*, 50(1):162–174, 2010.
[88] M.E. Tipping and A.C. Faul. Fast marginal likelihood maximisation for sparse Bayesian models. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, 2003.