NeuroImage 50 (2010) 1519–1535

Contents lists available at ScienceDirect

NeuroImage

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

High-dimensional pattern regression using machine learning: From medical images to continuous clinical variables

Ying Wang *, Yong Fan, Priyanka Bhatt, Christos Davatzikos

Section of Biomedical Image Analysis, Department of Radiology, University of Pennsylvania, Philadelphia, PA 19104, USA

ARTICLE INFO

Article history:
Received 5 September 2009
Revised 18 December 2009
Accepted 22 December 2009
Available online 4 January 2010

Keywords:
High-dimensionality pattern regression
Adaptive regional clustering
Relevance vector regression
Alzheimer's disease
MRI

ABSTRACT

This paper presents a general methodology for high-dimensional pattern regression on medical images via machine learning techniques. Compared with pattern classification studies, pattern regression considers the problem of estimating continuous rather than categorical variables, and can be more challenging. It is also clinically important, since it can be used to estimate disease stage and predict clinical progression from images. In this work, adaptive regional feature extraction approach is used along with other common feature extraction methods, and feature selection technique is adopted to produce a small number of discriminative features for optimal regression performance. Then the Relevance Vector Machine (RVM) is used to build regression models based on selected features. To get stable regression models from limited training samples, a bagging framework is adopted to build ensemble basis regressors derived from multiple bootstrap training samples, and thus to alleviate the effects of outliers as well as facilitate the optimal model parameter selection. Finally, this regression scheme is tested on simulated data and real data via cross-validation. Experimental results demonstrate that this regression scheme achieves higher estimation accuracy and better generalizing ability than Support Vector Regression (SVR).

Introduction

High-dimensional pattern classification has been increasingly used during the past decade as a means for quantifying and detecting spatially complex and often subtle imaging patterns of pathology from medical scans. It aims to provide sensitive and specific diagnostic and prognostic indicators for individuals, as opposed to analyzing statistical group trends and differences (Lao et al., 2004; Thomaz et al., 2004; Cox and Savoya, 2003; Davatzikos et al., 2006, 2008, 2009; Fan et al., 2007; Kloppel et al., 2008; Liu et al., 2004; Vemuri et al., 2008, 2009a,b; Golland et al., 2002). However, classification is a dichotomous process, i.e. it assigns an individual to one category of two or more. Many pathologies and diseases present a continuous spectrum of structural and functional change. For example, AD pathology is known to progress gradually over many years, sometimes starting decades before a final clinical stage. It is therefore important to estimate continuous clinical variables that might relate to disease stage, rather than categorical classification. Furthermore, the ability to predict the change of clinical scores from baseline imaging is even more important, as it would estimate disease progression, and then improve patient management. Towards this goal, this paper investigates high-dimensional pattern regression methods.

Some regression methods have been established in recent imaging literature (Duchesne et al., 2005, 2009; Ashburner, 2007; Davis et al., 2007; Formisano et al., 2008). Duchesne et al. (2005, 2009) proposed a multiple regression approach to build linear models to estimate yearly Mini Mental State Examination (MMSE) changes from structural MR brain images. The features used in these models were extracted from intensities of structural MR brain images and deformation fields with Principal Component Analysis (PCA) techniques. Unfortunately, the linear model derived from multiple regression is not always reliable in capturing nonlinear relationships between brain images and clinical scores, especially with limited training samples of high-dimensionality. Davis et al. (2007) presented a manifold kernel regression method for estimating brain structural changes caused by aging. Actually this study addressed the converse problem, i.e. the problem of finding an average shape as a function of age, whereas we try to predict a continuous variable from a structural or functional imaging measurement, a problem that is hampered by the sheer dimensionality of the feature space.

Kernel methods have recently attracted intensive attention. In particular, Support Vector Machines (SVM), have become widely established in classification problems because of their robustness in real applications (Cox and Savoya, 2003; Fan et al., 2007, 2008). Due to SVM's success in classification, it has been used for regression (Harris Drucker et al., 1996; Mangasarian and Musicant, 2000, 2002; Smola and Schölkopf, 2004) as well. However, support vector regression has some disadvantages that become especially

* Corresponding author. Fax: +1 215 614 0266.
E-mail address: ying.wang@uphs.upenn.edu (Y. Wang).

1053-8119/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2009.12.092

pronounced in high-dimensional problems. In particular, typical application of SVR to medical images leads to a very large number of support vectors, thereby potentially limiting the generalization ability of the classifiers and regressors. This problem is particularly prominent in regression, which requires much larger training sets than classification, since it estimates a continuous variable. In order to overcome this limitation, a sparser regression method, called Relevance Vector Regression (RVR) has been proposed in Tipping (2001), which effectively uses L_1 sparsity instead of L_2 that is used in SVM, and hence leads to significantly sparser models that are likely to generalize well. Ashburner has used RVR to predict subject age (Ashburner, 2007). Formisano et al. (2008) applied SVM on classification and Relevance Vector Machine (RVM) on regression of brain responses by using PCA features extracted from functional MRI time series. Unfortunately, in Formisano et al. (2008), there was no systematic evaluation and comparison between SVM and RVM, especially for regression problem. Within our proposed framework, RVR and SVR are investigated by comparing their sparseness, generalization ability and robustness on both simulated and real data.

Due to the curse of dimensionality, dimensionality reduction and feature extraction are necessary, even though RVM is a very sparse model with good generalization ability. PCA has been commonly used to linearly reduce dimensionality (Davatzikos et al., 2008; Thomaz et al., 2004; Formisano et al., 2008). However, PCA is not always able to identify localized abnormal regions of pathology, which limits the power of classifiers or regressors. To address this problem, we use an adaptive regional clustering method previously discussed in Fan et al. (2007). Compared with global features extracted by PCA, spatial regional measures potentially offer a better way to capture spatial patterns of structure and function, especially in diseases that have regional specificity. Moreover, regional features are constructed to have strong discriminative power. Based on these regional features, we use RVM to construct regression models. To improve stability of the training process, a bagging framework is adopted to build ensemble regression models to facilitate optimal parameter selection. From the experimental results utilizing different feature extraction methods, regional feature-based RVR is found to be an effective regression method that yields robust estimation of continuous clinical variables with reasonable diagnostic accuracy and good generalization ability.

In the next section, we introduce the proposed methods in detail. In Data and applications, the simulated data and real data from different clinical applications are described. In Experimental result analysis, we verify the proposed method by extensive experiments and comparisons with SVR via other common feature extraction techniques. A set of conclusions drawn from the extensive discussions of experimental results, and possible work for further improvements are presented in Discussion and conclusion.

Methods

Our machine learning framework is presented in Fig. 1, which includes three key steps: (1) the first step depends on the specific application and involves calculation of certain features of interest from the original images. In many neuroimaging applications, we are interested in measuring spatial patterns of brain atrophy of growth. Therefore, we first design Tissue Density Maps (TDMs), as commonly done in the neuroimaging literature (Ashburner and Friston, 2000; Davatzikos et al., 2001; Fan et al., 2007). Based on TDMs, an adaptive regional clustering method is then applied to capture informative regional features, and then a subset of features with top-ranking correlation power to interested regression problem are selected from the original regional clusterings; (2) RVM is then used to construct efficient and sparse regression models; (3) A bagging framework is used to build ensemble regressors in order to improve model stability, as well as facilitate optimal parameter selection for regression models with the limited training data.

Feature extraction and selection

Medical imaging offers a wealthy information of spatial, temporal and multiparametric measurements of structure and function. However, the resultant feature space is too large for robust and fast pattern analysis, especially when only limited samples are available. With this in mind, the essential step in an imaging-based machine learning framework is to find a suitable representation of high-dimensional data by means of dimensionality reduction. The ideal representation has to be in a low-dimensional feature subspace while keeping as much information as necessary to estimate the variables of interest. More importantly, feature extraction needs to be robust to measurement noise and preprocessing errors, and should be easily extended to other applications. To address these challenges, we use an adaptive regional feature clustering method, and then investigate its discrimination ability and its robustness by comparing it with three benchmarks among dimensionality reduction methods: Linear Discriminant Analysis (LDA) and PCA, and also nonlinear manifold-learning approach (Hotelling, 1933; Fisher, 1936; Tenenbaum et al., 2000).

Adaptive regional feature clustering

This feature extraction algorithm is developed in a similar way to an adaptive regional clustering method, which has been successfully used in classification of brain images (Fan et al., 2007). The difference is that, during the brain region generating procedure, our method is to consider the similarity of correlation coefficients between voxel-wise measures and continuous clinical scores being regressed, instead of class labels.

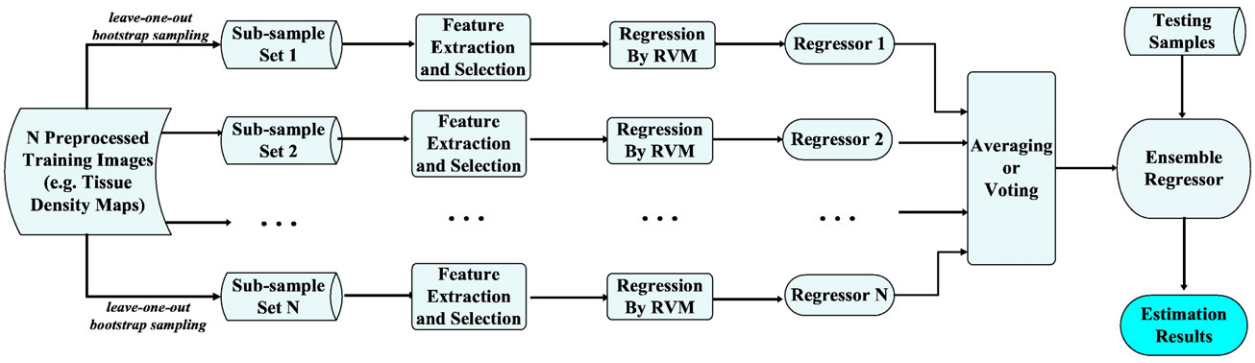

Fig. 1. A bagging framework of regression using RVM.

In particular, for a brain MR scan, we first calculate three TDMs corresponding to gray matter (GM), white matter(WM) and cerebrospinal fluid (CSF) respectively, as commonly done in neuroimaging (Davatzikos et al., 2001) in order to quantify regional tissue volumes. The correlation coefficients on voxel u between TDMs and continuous clinical variables can be represented as:

$$c^i(u) = \frac{\sum_j \left(f_j^i(u) - \overline{f^i(u)}\right)\left(r_j - \bar{r}\right)}{\sqrt{\sum_j \left(f_j^i(u) - \overline{f^i(u)}\right)^2 \sum_j \left(r_j - \bar{r}\right)^2}} \quad (1)$$

where j denotes the jth training sample; $i=1, 2, 3$ represents GM, WM and CSF respectively; $f_j^i(u)$ is a morphological feature of tissue i from the jth sample in the location u, and $\overline{f_i(u)}$ is the mean of $f_j^i(u)$ over all samples; $\bar{r}$ is the mean of all clinical variables r_j. A watershed segmentation algorithm is then applied to the correlation map computed above, adaptively generating spatial clusters adaptively (Vincent and Soille, 1991; Grau et al., 2004). Once these regional clusters are obtained, features can be computed from them. For example, the sum of tissue density maps represents the volume of the respective tissue within that cluster; the average image intensity within a cluster would be used for functional or molecular images.

Note that we seek the maximal regional cluster size that maintains a strong correlation with the variable being estimated. With this we aim at better robustness and generalization of the estimators to be built from these features. Even though an individual voxel could display a very strong correlation with the clinical variable being estimated, it is unlikely that this same voxel will be a good feature for other samples. Larger clusters are more likely to represent true biological processes and to lead to robust and generalizable models.

Feature selection

Although the number of generated clusters is much smaller than the number of original voxels, measures from some regions are still irrelevant or redundant for regression. This requires a feature selection method to select a small subset of regional clusters in order to improve generalization performance of regression.

In this paper, we employ a common feature-ranking method based on correlation coefficients as calculated by Formulation (1). Generally, the feature ranking method computes a ranking score for each feature according to its discriminative power, and then simply selects top ranked features as final features. Considering the correlation coefficient has been successfully employed as a feature ranking criterion in a number of applications (Guyon and Elisseeff, 2003; Fan et al., 2007), the absolute value of leave-one-out correlation coefficient is used here to rank features. We set the number of features as one less than the total number of training samples. Reverse feature elimination methods, like the one employed in Fan et al. (2007), are likely to further improve feature selection, albeit at the expense of significantly higher computational load.

Relevance vector regression

In high-dimensional pattern regression, the most challenging task is to find a suitable model with good generalization and accurate estimation performance from small-sample data sets (relative to the dimensionality of input images). SVM is a popular method with a simple implementation and high estimating accuracy in many real-world applications (Vapnik, 1998, 2000; Cox and Savoya, 2003; Fan et al., 2007, 2008). It uses a kernel function to find a hyperplane that minimizes the deviation from the actually obtained targets for all training data. Despite its success, SVM has some limitations: firstly, it is not a very sparse model, because the number of support vectors typically grows linearly with the size of training samples. Secondly, the basis functions must be continuous symmetric kernels with a positive real coefficients under Mercer's conditions. Thirdly, SVM does not output the probability distribution of the estimated variable, and therefore does not provide an estimate of the error variance. These limitations have been overcome by RVM (Tipping, 2001), which is formulated in a Bayesian estimation theory. RVM provides probabilistic prediction with a conditional distribution that allows the expression of uncertainty in estimation and prediction (Tipping, 2001). Furthermore, RVM can employ arbitrary basis functions without having to satisfy Mercer's kernel conditions.

SVM obtains sparsity by using the L_2-norm as a regularizer defined in a reproducing kernel Hilbert space of the kernel functions. In contrast, the sparsity of RVM is induced by the hyperpriors on model parameters in a Bayesian framework with the Maximum A Posteriori (MAP) principle. L_1-norm like regularization used in RVM provides significantly fewer basis functions, which is often important for good generalization. The reason is that L_1-norm encourages the sum of absolute values, instead of squares of the L_2-norm, to be small, which often drives many parameters to zero. L_1-norm also has a greater resistance against outliers than L_2-norm, since the effect of the outliers with a large norm is exaggerated by using L_2-norm. Table 1 summaries these differences between RVM and SVM.

RVM, originally introduced in the machine learning literature (Tipping, 2001), aims to find out the relationship between input feature vectors $\boldsymbol{x} = \{\mathbf{x}_n\}_1^N$ and the corresponding target values $t = \{t_n\}_1^N$:

$$t_n = y(\mathbf{x}_n) + \varepsilon_n \quad (2)$$

where ε_n is the measurement noise, assumed as independent, and to follow zero-mean Gaussian distribution, i.e., $\varepsilon_n \sim \mathrm{N}(0,\sigma^2)$; $y(\mathbf{x})$; is assumed as a linear combination of basis functions $\phi(\mathbf{x}, \mathbf{x_n})$ with the form:

$$y(\mathbf{x}) = \sum_{n=1}^{N} \omega_n \phi(\mathbf{x}, \mathbf{x_n}) + \omega_0 \quad (3)$$

Where $W = (\omega_0,\omega_1,\omega_2,\ldots,\omega_N)^T$ is a weight vector. Assume $\mathbf{t}_n$ is drawn from a Gaussian distribution with mean $y(\mathbf{x}_n)$ and variance σ_2, then

$$\mathbf{t} = y(\mathbf{x}) + \varepsilon = \Phi W \quad (4)$$

where ε is a noise vector with elements ε_n; $\Phi_{N\times(N+1)}$ is the design matrix, wherein $\Phi_{ij} = \phi(\mathbf{x}_i, \mathbf{x}_j)$, $i=1,\ldots, N$; $j=0,\ldots,N$, and $\phi(\mathbf{x}_i, \mathbf{x}_0) = 1$.

Assuming statistical independence of the samples, t_n, the maximum likelihood estimate for W is given by:

$$p\left(\mathbf{t}|W,\sigma^2\right) = \left(2\pi\sigma^2\right)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2}|\ \mathbf{t} - \Phi W|^{\ 2}\right\} \quad (5)$$

which is known as the common least square estimate, and suffers from over-fitting. To over-come this problem, a desired constraint on

Table 1
Comparison of SVM and RVM.

	SVM	RVM
Introduced	1996, Vapnik	2001, Tipping
Theory	Structural risk minimization	Bayesian formulation
Basis function	Mercer's condition limited	Arbitrary kernel
Regularization	L_2-norm	L_1-norm
Sparse representation	Larger set of support vectors	Fewer relevance vectors needed
Regression	Point estimation	Probabilistic prediction

the model parameters W is necessary. Therefore, sparsity-inducing hyperprior within a principled Bayesian MAP methodology is introduced to provide such constraint, which not only leads to smooth function estimate, but also drives many weight parameters into zeros. For more details, see Appendix A, "Further details of relevance vector learning".

Bagging framework for building ensemble regression models

As described earlier, a fundamental challenge for pattern regression by machine learning algorithms lies in the small size and the high dimensionality of the input data. Samples of small size, relative to the input dimensionality, are typically insufficient to extract discriminative features with good generalization ability when the underlying patterns are complex, i.e. require many variables to be described effectively. Hence, features extracted from limited training data sets often lead to "unstable" regression models, where small changes in the samples result in large changes in regression performance, even though RVR is known to be very sparse. In order to improve the stability of regression, we use bagging (bootstrap aggregating)-based RVR, which has been successfully applied in brain image classification (Fan et al., 2008).

Given a set of bootstrap samples $\{L_1,\ i=1,\ldots l\}$ generated from the training set, and a number of base models $\{f_1,\ i=1,\ldots l\}$ from $\{L_1,\ i=1,\ldots l\}$, a bagging regression model can be constructed:

$$f(v) = \sum_{i=1}^{l} a_i f_i(v) \tag{6}$$

where $\sum_{i=1}^{l} a_i = 1$, and $a_i \geq 0$, $i = 1, \ldots l$, are weighting factors for base models. The weights can be simply set as $1/l$, or values reflecting the performance of base models. It has been shown both empirically and theoretically that bagging improves the Mean Squared Error (MSE) for "unstable" regression models whereas it remains roughly the same for "stable" ones (Breiman, 1996a). Furthermore, if the bootstrap samples are generated by leaving out part of the training data, the left-out data can be used to estimate the performance of base models, which can be subsequently used to select the optimal parameters of learning algorithms (Breiman, 1996b). This is a valuable byproduct of bagging, besides its primary purpose to increase estimation accuracy, especially in the applications with limited samples.

To generate multiple versions of the training set, a leave-k-out bootstrap sampling scheme is used in this paper due to the limited availability of training data. For a training set with n samples, A succession of different bootstrap training sets are generated and each of them has $(n–k)$ samples. Based on each bootstrap training set, a RVR model, denoted as the base regressor, is trained based on the extracted features. Then RVR parameters are determined by optimizing the performance of base models on the left-out training data. The optimal regressors are then applied to the test samples. In current work, Gaussian Radial Basis Function (GRBF) is used to map the original features to the infinite-dimensional kernel space. MSE is used to measure the performance of a regression model:

$$MSE = \frac{1}{n^*}\sum_{i=1}^{n^*}\left(f\left(x_i^*\right) - t_i^*\right)^2 \tag{7}$$

where $(x_i^*,\ t_i^*,\ i=1,2,\ldots n^*)$ is the testing set.

Data and applications

To evaluate the regression performance of bagging based RVR, we use images from two sources: simulated data and real data.

Simulated data

Simulations allow us to generate images with local thinning in a precisely known relationship. To simulate atrophy of the brain cortex, we simulated twenty shapes $\{S_n,\ n=1,\ \ldots\ 20\}$, each of which represents an individual subject with increasing level of thinning in some regions. In practice, the regions and atrophy rates might be different from one subject to another. Towards this goal, for each shape S_n, we create ten "follow-up" images that have region shrinkage within three different areas according to pre-specified rates. Therefore, we totally have two hundred images with their corresponding target values: $\{I_{nj},\ T_{nj},\ n=1,\ \ldots\ 20,\ j=1,\ \ldots\ 10\}$. This simulation process is detailed next:

Step 1: Based on the initial shape, create twenty different shapes by applying small random geometrical transforms: $\{S_n,\ n=1,\ldots 20\}$;

Step 2: For each shape S_n, randomly select the locations of the center of three areas, and assign the respective size of each of them; thereby generate three regions $\{\mathbf{R}_i,\ i=1,2,3\}$ to represent the areas of pathology (thinning in this case);

Step 3: For each region R_i, define a 10-dimensional vector $\{\mathbf{V}_{ij}=(v_{i,1},\ldots,v_{i,10})\}$ associated with increasing thinning rates, and $0 \leq v_{i,1} < \ldots < v_{i,10} < 1$; then implement morphological operations on each region until the given rates are met;

Step 4: Define the exact (actual) value of the target variable for each image as follows:

$$T_{nj} = (\mathrm{Area}\mathbf{R}_1)^2 + m^*(\mathrm{Area}\mathbf{R}_2) + \mathrm{Area}\mathbf{R}_3 \tag{8}$$

where m is constant coefficient, Area$\mathbf{R}_1$, Area$\mathbf{R}_2$ and Area$\mathbf{R}_3$ represent these three regional sizes, respectively. It will be tested whether our algorithm can accurately discover the relationship between the images and target variables (regressors), without knowing the regions $\mathbf{R}_1$, $\mathbf{R}_2$ and $\mathbf{R}_3$ or the actual relationship in Eq. (8). Fig. 2 illustrates an example of the thinning simulation in detail.

Real data

Real data are brain MR images from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (ADNI, 2004; Jack et al., 2008). They are standard T1-weighted MR images acquired using volumetric 3D MPRAGE with 1.25×1.25 mm in-plane spatial resolution and 1.2 mm thick sagittal slices from 1.5T scanners. To ensure consistency among scans, all downloaded images were preprocessed by GradWarp (Jovicich et al., 2006), B1 non-uniform correction (Jack et al., 2008), and N3 bias field correction (Sled et al., 1998). General information is detailed under the ADNI website.

The goal of regression is to discover the potential relationship between brain atrophy patterns and clinical stage of the disease, the latter being linked to some clinical variables. The proposed method is measured by two continuous clinical scores: MMSE and Boston Naming Testing (BNT), which are reliable and frequently used in the

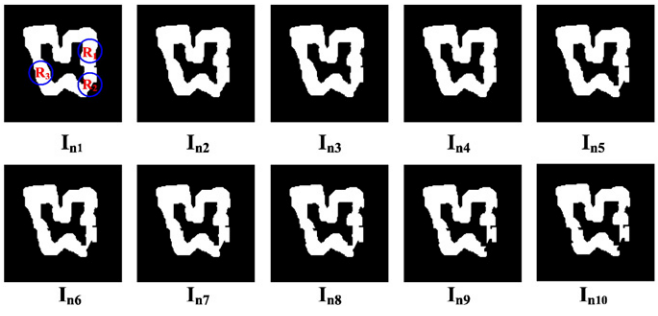

Fig. 2. A sample with ten "follow-up" images in the simulated data.

Table 2
Participant information.

Clinical score	Information	AD	MCI	CN
MMSE	No. of subjects	23	74	22
	Age (year), mean ± std	78.45 ± 6.02	75.87 ± 7.28	73.25 ± 5.45
	MMSE, mean ± std	15.06 ± 2.43	25.75 ± 2.44	28.54 ± 0.98
	Interval (month), mean ± std	16.17 ± 5.84	19.14 ± 4.93	15.82 ± 5.72
BNT	No. of subjects	22	31	16
	Age (year), mean ± std	77.60 ± 5.95	74.75 ± 8.63	71.36 ± 7.04
	BNT, mean ± std	16.55 ± 1.17	23 ± 2.60	28.88 ± 0.78
	Interval (month), mean ± std	15.27 ± 5.47	18.39 ± 4.88	16.88 ± 5.89

evaluation of Alzheimer's disease (Modrego, 2006; Petersen et al., 2001; Tombaugh and McIntyre, 1992). The total scores for the MMSE and BNT range from 0 to 30, and the lower scores indicate more severe cognitive impairment.

Since the follow-up interval of ADNI is 6 months, to get robust and stable regression results, at least two follow-up scans besides baseline scanning are required during sample selection. However, the ADNI dataset has more Mild Cognitive Impairment (MCI) patients than AD and Cognitively Normal (CN) subjects. To avoid underweight the relatively fewer samples with very high or very low clinical score, we uniformly sample the cases from the entire range of clinical variables. Note that, only baseline scans are used in experiments, but in order to remove the noise that is commonly encountered in measuring the MMSE scores, the mean of all the measured scores from baseline and each follow-up scanning is taken as the corresponding regressor for each subject. We have two reasons for doing this. Firstly, the individual cognitive evaluations are known to be extremely unstable and depend on a number of factors unrelated to the underlying brain pathology; secondly, the proposed methodology emphasizes the prediction ability of the regressor at baseline rather than from longitudinal scans. Despite that the baseline scan with the averaged clinical variables is not the optimal way, it was one of the best solutions we found to remove random variations in the cognitive measures, especially in the study with short clinical follow-up period.

In this paper, 23 AD patients, 74 patients with MCI and 22 CN subjects are included for MMSE, while 22 ADs, 31 MCIs and 16 CNs for BNT. More participant information is shown in Table 2. We also provide the lists of these samples with Tables 5 and 6 in the Appendix C.

Experimental result analysis

Simulated clinical stage estimation

Simulated shapes were first employed to evaluate the performance of RVR by using four feature extraction methods via leave-10-out cross-validation. As indicated in Simulated data, there were 200 images (20 subjects, each of them has ten follow-up images) in simulated data, 190 images (19 subjects) of which were used to construct four feature subspaces based on regional clusters. The training data and testing data (10 images) were then projected to these subspaces for their corresponding feature extraction. This process was repeated 20 times, each time leaving out 10 different samples. For each time, regression models were trained using Gaussian kernel-based RVR, and model parameters were optimized by bagging framework with the second leave-10-out cross validation (each time 10 training samples from the same subject were left out to create a training subgroup L_i).

To get suitable RVR models, different kernel sizes were tested to determine the appropriate range, in which the optimal regression

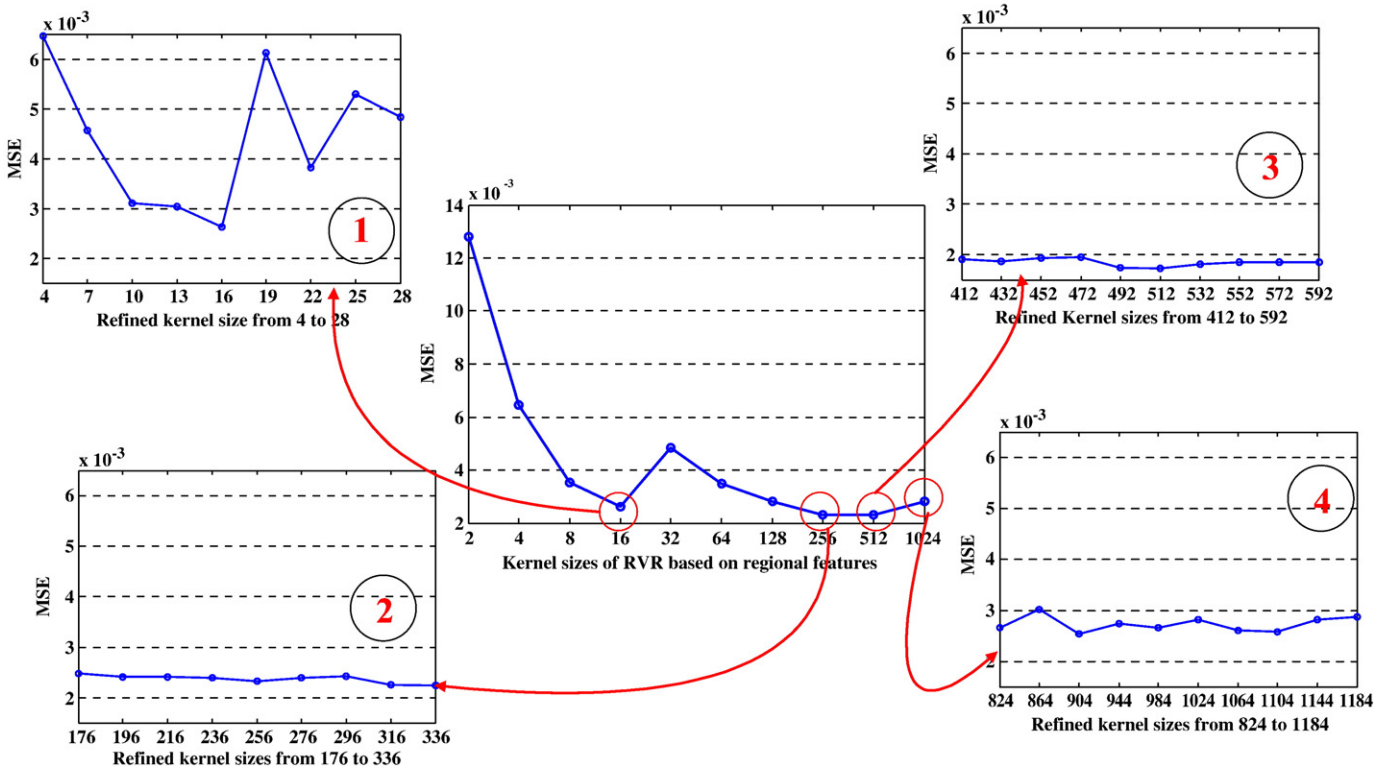

Fig. 3. MSEs for different kernel sizes of RVR based on regional features. The small graphs are correspond to different refined ranges of kernel sizes.

performance can be achieved robustly. Firstly, a wide range from 2 to 1024 was scanned in order to find out the suitable kernel sizes for our regression problem, and then the ranges around these sizes were refined to select the relatively stable sizes. From the middle graph of Fig. 3, MSEs ranging from 0.002 to 0.014 were taken as kernel sizes varied from 2 to 1024. The lowest MSE was obtained when kernel size was 512. To avoid local minima, four kernel sizes (16,128,512 and 1024) were selected for further refinement, and the corresponding MSEs are illustrated respectively in the small graphs of Fig. 3. Obviously, although regression accuracy fluctuates greatly at the smaller values of kernel sizes (graph 1), in the other three ranges, the regression accuracy curves are very smooth (graphs 2, 3 and 4). Especially during the range from 412 to 592, MSEs are lower than 0.002. It can thus be concluded that, RVM is a stable regression model within a reasonable range of kernel sizes. In general, avoiding the relatively smaller kernel sizes is advisable since they lead to overtraining.

For comparison purposes, GRBF-kernel based SVR with ε-insensitive loss function was also tested by the same leave-10-out cross-validation procedure, as we described in the RVR case. Besides kernel function parameters, two more SVR parameters, namely C and ε, were optimized within the bagging framework based on the training data. C is the tradeoff parameter between the smoothness of the SVR function and the training error, and ε is the error insensitivity parameter, i.e., the training error below ε will be treated as zero (Vapnik, 2000; Smola and Schölkopf, 2004). In the training process, the performance of base regression models was evaluated within the bagging framework, with respect to different parameters (Chang and Lin, 2001). From the regression results summarized in Table 3, regional feature-based regression got the highest correlation and the smallest estimation error. Meanwhile, regression by means of LDA and PCA features overall achieved reasonable accuracy, while ISOMAP features performed the worst.

The detailed regression results are illustrated by Fig. 13 in Appendix B, in which the first row shows the estimated scores by SVR and RVR respectively with regional features. The other three rows summarize the regression results based on LDA, PCA and ISOMAP features, where RVR is shown on the left side and SVR on the right. Note that, the regression lines are shown with black solid lines.

In order to verify the validity of our learning algorithm, a discriminative direction method is applied to display the changing regional features found by the regressor (Golland et al., 2002). This method provides a group difference map, indicating the degree that each regional feature participates in regression estimation. A leave-one-out validation is performed in our experiments for testing the generalizability of RVR, and the group difference will be evaluated by averaging all group difference maps obtained from all leave-one-out cases. First, for each relevance vector in each leave-one-out case, we investigate its corresponding projection vector in the high-dimensional space. The rule is to follow the steepest gradient of regressor from small to large values. The difference between this relative vector and its corresponding projection vector reflects the changes on the selected regional features. These features become typical when a normal brain gradually changes to the respective configuration with disease progression. Second, an overall group difference vector can be obtained by summing up all regional differences calculated from all relative vectors. Finally, the group difference vector is mapped to its corresponding brain regions in the template space and subsequently added to other leave-one-out repetitions. Referring to Fig. 4, our regression algorithm captures the significant features of brain image with the increasing regressor value, similar with our pathology areas simulated in Stimulated data.

Table 3
Numerical results for simulated data regression with four feature extraction methods by RVR and SVR, respectively.

Methods	Features			
	Regional clusters	LDA	PCA	ISOMAP
Mean square errors				
RVR	0.001686	0.0018073	0.0019334	0.01017
SVR	0.00287	0.0052233	0.0063355	0.016377
Correlation coefficients				
RVR	0.98438	0.9842	0.9824	0.90386
SVR	0.97763	0.97661	0.97056	0.52889

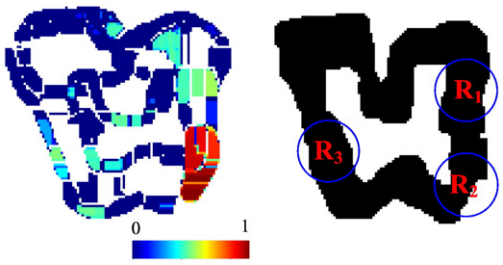

Fig. 4. Most representative sections with regions of the group difference of the simulated data, as shown in the left figure. Parts R_1, R_2, and R_3 represent the areas of pathology, more details can be found in Simulated data.

Clinical stage estimation

GM based regression with ADNI data

Since RVR showed high estimation accuracy and robustness with simulated data, similar experiments were carried out using clinical images from ADNI. Images were first preprocessed into tissue density maps as described in Davatzikos et al. (2001). Aiming to improve the robustness of local TDMs, a 8-mm full-width at half-maximum Gaussian smoothing was applied to all the images in the experiments. For simplicity, only gray matter after tissue segmentation was used to investigate regression performance of the methods mentioned above. Different from simulated data, leave-one-out cross-validation scheme was performed to test RVR performance on real data. For 119 samples selected from ADNI data, 118 samples were taken as training set for feature extraction and bagging-based model building, and the one left was chosen as testing sample to evaluate regression performance. This process was repeated 119 times, and each time ensemble regressors were trained by bagging framework via the second leave-10-out cross-validation within the training set.

We compared our method with SVR based on different feature extraction approaches in the same experiment setting. As illustrated in Fig. 5, there are four rows with double columns to show estimated MMSEs by RVR and SVR combining regional features, LDA, PCA and ISOMAP features, respectively. Although estimation performance was not as good as that of simulated data, we still obtained the promising results for real data. The best result was adaptive regional feature-based RVR with correlation 0.73868.

For BNT score, although the best estimation accuracy was also obtained by RVR and regional features as shown in Fig. 6, the overall regression performances for BNT was lower than those for MMSE. Even in the best case, the correlation coefficient between estimated BNT and measured BNT was only 0.58908. This inferior performance might be due to the comparatively small size of samples used for BNT, which are too limited to ensure efficient feature extraction and stable model parameter selection (there are 119 samples in MMSE set, but only 69 in BNT set).

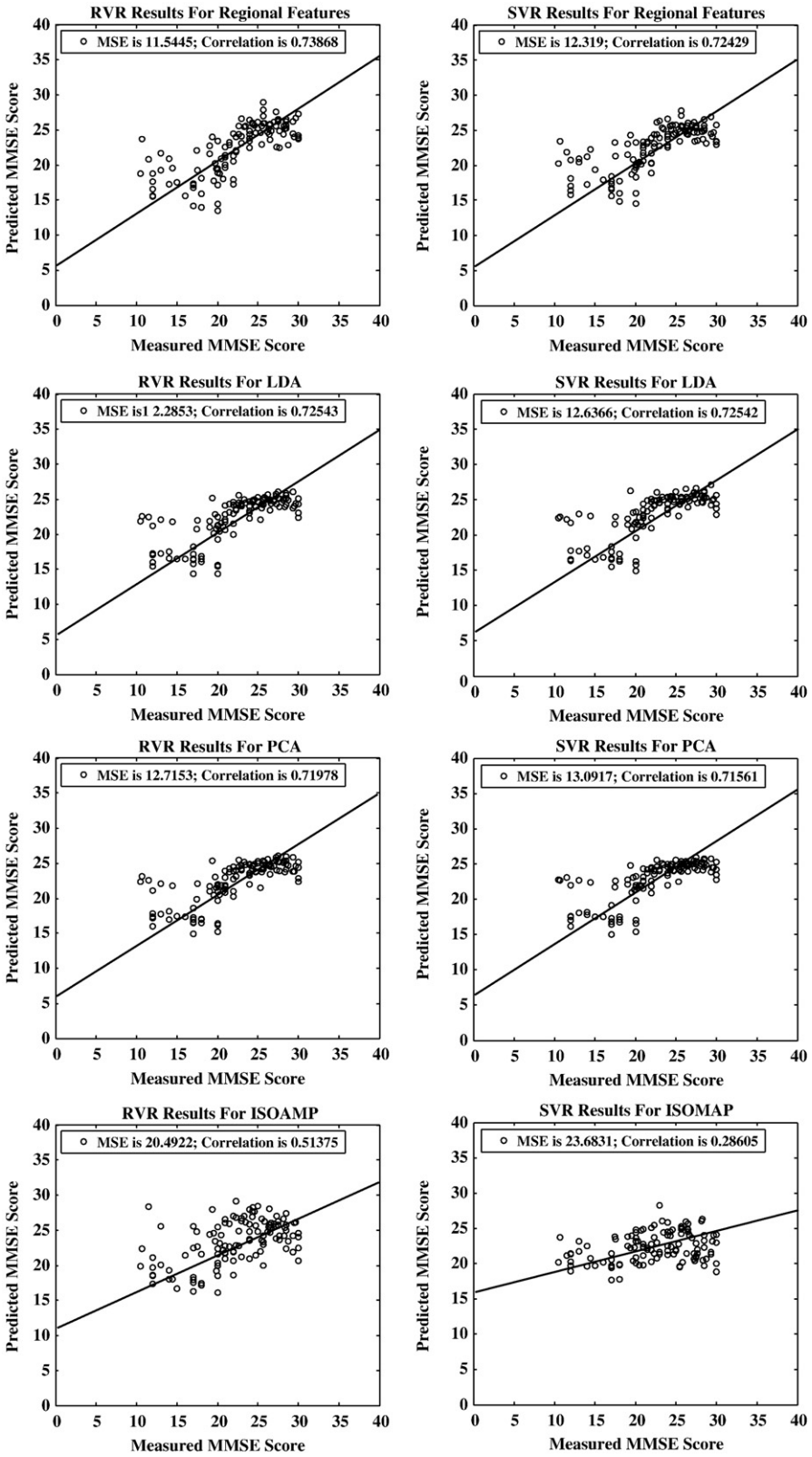

Fig. 5. Scatter plots of clinically measured MMSE scores vs. estimated scores by RVR/SVR with four feature extraction methods: region clustering, LDA, PCA and ISOMAP. The graphs in the right column are for RVR, and those of the left column are for SVR. The sloping black solid lines represent the regression lines.

No matter which feature extraction method was applied, RVR always obtained more accurate estimation with higher correlation than SVR. The reason is likely to be that fewer Relevance Vectors (RVs) are used in RVR, and thus RVR has better generalization ability, which leads to more accurate estimation, especially when only limited training samples are available. To further investigate the generalization ability of RVR and SVR, we designed the following experiment: we initially chose 10 training samples from MMSE-

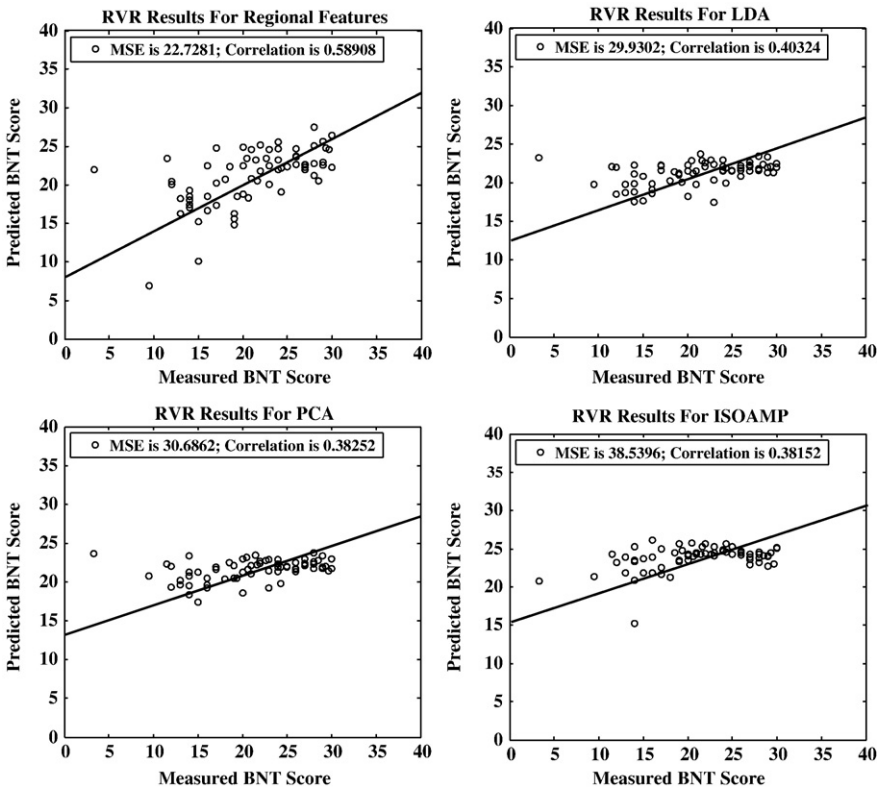

Fig. 6. Scatter plots of clinically measured BNT scores vs. estimated scores by RVR based on four feature extraction methods.

based ADNI data, then added 10 new samples each time and recorded the number of RV and Support Vectors (SVs) after the learning process. Regional clusters and LDA features were selected for these experiments. The generalization performance of RVR/SVR under regional and LDA features is summarized in Fig. 7, in which the blue curves correspond to RVR while the red curves are for SVR. It is evident from the figures that, for both regional and LDA features, the number of RVs/SVs increases with the changes of the number of training samples. However, the number of RVs is always smaller than that of SVs, especially for smaller sample sizes. This just confirmed the theoretic analysis of sparsity expected from these two regression methods. Experimental results also demonstrate that the number of RVs with LDA-based features is always smaller than that of regional features. Theoretically, LDA feature-based regression should result in better generalization ability since it needs fewer RVs. Nevertheless, LDA feature-based regression got inferior estimation performance while regional feature-based regression obtained lower MSE and higher correlation coefficient (Fig. 5). The reason might be that regional feature clusters could be more informative, since AD pathology tends to have a regional distribution.

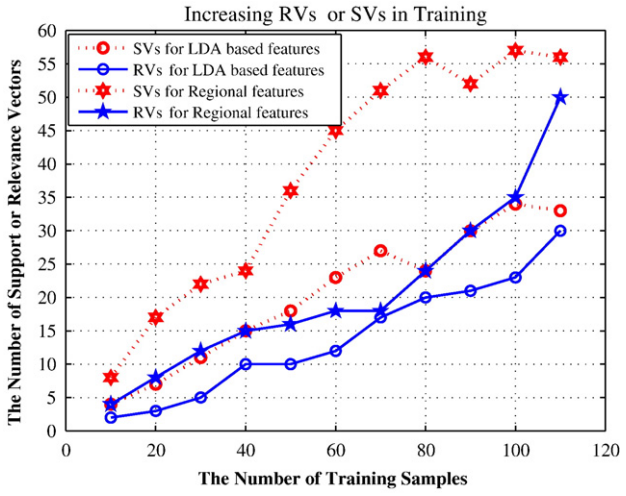

Fig. 7. Comparison of support/relevance vectors used in SVR and RVR. Curves show how the number of support/relevance vectors changes as the size of training samples increases.

GM, WM and CSF based regression

AD is associated with brain volume loss of gray matter, but white matter and CSF might also carry significant information. In this section, we tested RVR on the combined tissues of GM, WM and CSF by the same experiment setting as we did with GM alone. As can be seen from Fig. 8, RVR with regional features has the highest correlation and lowest MSE, followed by SVR with regional features. Regarding all experimental results in Simulated clinical stage estimation and GM based regression with ADN data, regional feature-based regression is superior to regression methods based on other feature extraction schemes. It proved that adaptive regional clustering is the best feature extraction method for both RVR and SVR in our experiments.

To compare the performance between GM-based regression and three tissue (GM, WM and CSF)-based regression methods in an organized way, Table 4 was constructed to record the numerical results appearing in Fig. 8 and Fig. 5, where the best results of each category are marked in blue. For both RVR and SVR, there were marginal improvements with regression performance based on three tissues compared to GM only, possibly because any additional information gained from WM and CSF is either marginal or counter-balanced by the curse of dimensionality issue. The best correlation coefficients and MSE between clinically measured and estimated MMSE scores are 0.75775 and 10.8338 from three tissue-based RVR

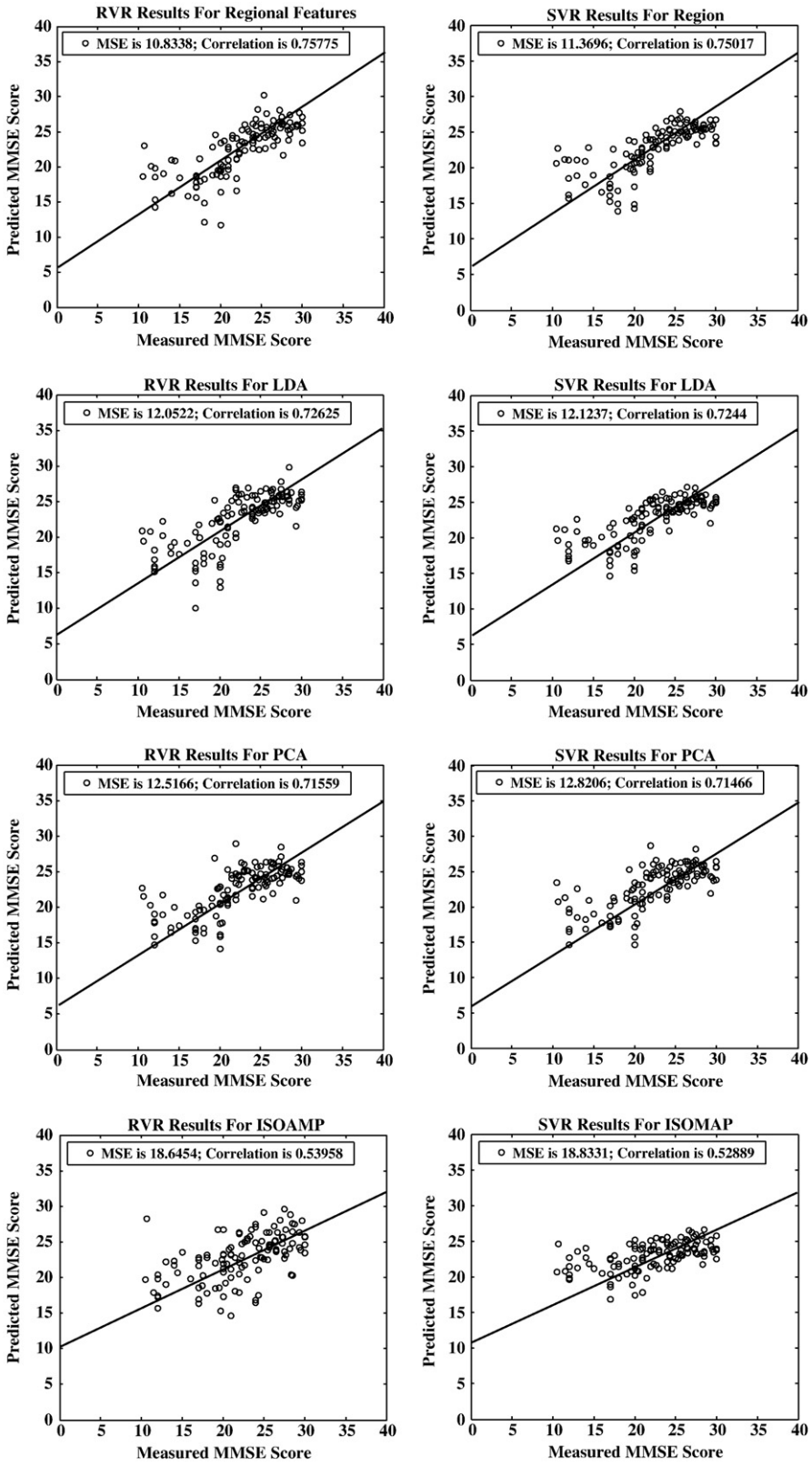

Fig. 8. GM, WM and CSF based regression results. Scatter plots of clinical measured MMSE vs. estimated MMSE by RVR/SVR with four feature extraction methods: regional clustering, LDA, PCA and ISOMAP. The two graphs in the first row show the best performance of RVR and SVR, respectively.

with regional features, while 0.73868 and 11.5445 for GM-based RVR.

In view of the superior performance of RVR, more detailed results from it are illustrated by Fig. 9. For each different feature extraction method, histograms of MSE and correlation coefficients clearly indicate the better regression performance of RVR by using GM, WM and CSF than GM only, albeit the improvement is marginal.

Table 4
Numerical results for three tissue-based regression, and only GM-based, by RVR and SVR, with four feature extraction methods.

Data and methods		Features			
		Regional clusters	LDA	PCA	ISOMAP
Mean square errors					
GM, WM and CSF	RVR	10.8338	12.0522	12.5166	18.6454
	SVR	11.3696	12.1237	12.8206	18.8331
GM only	RVR	11.5445	12.2853	12.7153	20.4922
	SVR	12.319	12.6366	13.0917	23.6831
Correlation coefficients					
GM, WM and CSF	RVR	0.75775	0.72625	0.71559	0.53958
	SVR	0.75017	0.7244	0.71466	0.52889
GM only	RVR	0.73868	0.72543	0.71978	0.51375

Another interesting result is that, when MMSEs range from 25 to 30 (most of them are the normal control cases), our proposed method is not sensitive to them, as shown in Figs. 5 and 8. The reason is likely to be that in those mostly normal subjects, variability in clinical scores can be attributed to measurement error and "bad testing days," both of which are well known in this field. For MCI and AD patients, however, the decline in cognitive ability is due to a large extent to underlying pathology and brain atrophy, which is measured by our algorithm. However, regression using all three tissues improved the estimation accuracy of MMSE in this range (comparing the results as shown in Figs. 5 and 8). To account for this improvement in a detailed way, we plot the distribution histograms of estimated MMSEs by RVR from three tissues and GM, respectively. As illustrated by the purple rectangles in Fig. 10 (interval scale is 1), we can see that most estimations for MMSEs ranging from 25 to 30 are around 26, as shown in the middle graph of Fig. 10. In contrast, three tissue-based regression got a more even estimation distribution from 26 to 28, which means they are correctly estimated by comparing with the clinical measured MMSEs (the left graph of Fig. 10). These results indicate additional information can provide more structural patterns, and thus improve regression accuracy. We expect that this result will hold even more strongly for large sample sizes.

Predicting future decline from baseline scans

Patients with MCI are at high risk from conversion to AD; they generally tend to progress to AD at a rate of about 15% annually. Therefore, it is of great clinical interest to be able to predict which of the MCI individuals are likely to progress rapidly and decline cognitively, using their baseline exams. We address this problem in the following set of experiments, by attempting to predict future decline in MMSE, by applying our approach to baseline MRI scans.

A large part of the available ADNI images are from patients who did not display significant cognitive decline, and would overwhelm the regression algorithm if all used in the current experiment. We therefore randomly selected a representative subset of individuals spanning the entire range of MMSE decline in a relatively uniform way. In other words we selected a mix of individuals from ADs and MCIs, who are supposed to display variable degrees of MMSE change. In particular, we ended up with 26 patient samples with stably degenerative variance during longitudinal study. Our cohort is composed of 16 MCI-converters (diagnosed as MCI at the baseline scanning, and then progressed to AD at 6-month follow-up), 5 MCI-nonconverters and 5 AD. A list of sample IDs could be found in Table 7 of Appendix C.

Although feature ranking has been used to select the optimal regional features based on the correlation coefficients among brain regions, some redundant features can be inevitably selected, which ultimately decrease the regression results. Moreover, ranking-based feature selection method is subject to local optima, especially when the size of training samples is small. Therefore, we firstly used the proposed correlation-based feature ranking method to select the most relevant features, and then applied the RVR-based Recursive Feature Elimination (RVR-RFE) algorithm to further optimize the set of initial regional features for good regression performance. This algorithm was inspired by the Support Vector Machine-Recursive Feature Elimination (SVM-RFE) approach (Guyon et al., 2002; Rakotomamonjy, 2003; Fan et al., 2007), which includes backward

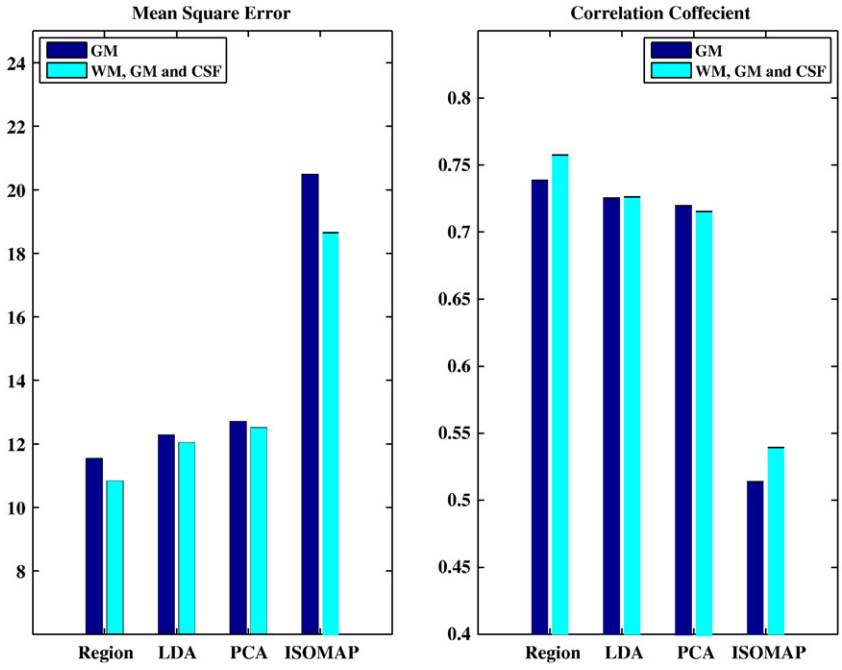

Fig. 9. Histograms of MSEs and correlation coefficients for RVR with four different feature extraction methods performed on two information source: three tissues and GM only.

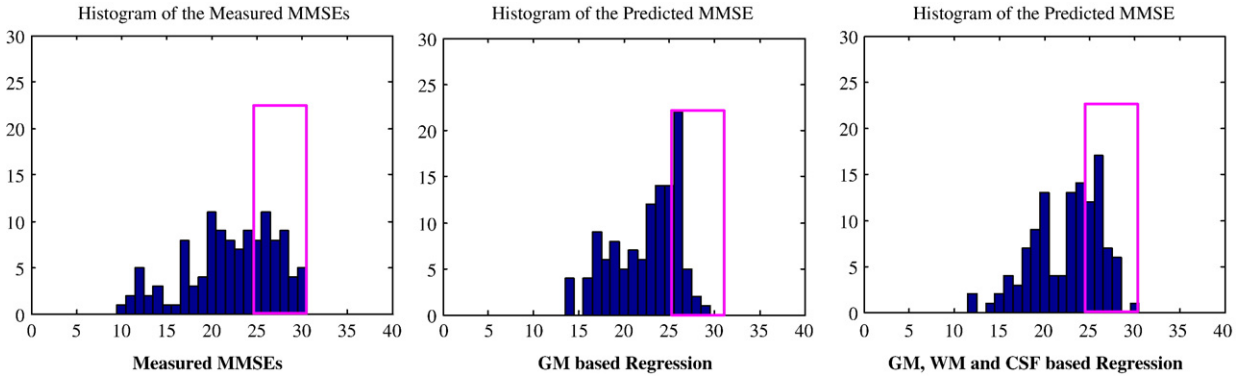

Fig. 10. Clinically measured and estimated MMSE histograms by RVR with regional features from three tissues and GM, respectively.

and forward steps. During the backward sequential selection, at each time, one feature is removed to make the variation of RVR-based leave-one-out error bound smallest, compared to removing other features. To avoid missing informative features whose ranks are low, but perform well jointly with the top ranked features for regression, a forward sequential feature selection method is used for compensation. This step adds one feature back at a time. Each time the added feature makes the variation of RVR-based leave-one-out error bound smallest, compared with adding other features. The search space of the forward selection is limited to a predefined feature subset in order to obtain a solution with cheaper computation cost.

In each leave-one-out validation experiment, one subject was first selected as a testing subject, and the other subjects were used for the entire adaptive regional feature extraction, feature selection, and training procedure, as described in framework Fig. 1. Fig. 11 illustrates the results of estimating the change and the rate of clinical variable change.

Note that the progression estimation results for MCI or AD patients are worse than the clinical stage prediction as discussed in Clinical stage estimation. The reason might be that, we have such assumption as follows on progression estimation: the rate of MMSE declines linearly along with brain volume atrophy for MCI and AD patients over longitudinal progression. Unfortunately, in our study, most samples are recorded within no more than three years, and also there are so many non-clinical factories that influence cognitive function testing, and all these factors will undermine the prediction results to some extent.

The group difference maps for all these clinical variable (MMSE, BNT and MMSE change) regression are shown in Fig. 12, which shows the brain regions that collectively contributed to the regression. The color scale indicates the frequency that brain regions were used in bagging regression experiments with different ADNI data. These figures reveal that brain atrophy patterns highly correlated with clinical scores are complex. However, we can see that, most of the group difference locations are consistent with previous clinical findings and computational analysis, such as the hippocampus, which plays a crucial role in memory processes, and usually is taken as a biomarker of early AD.

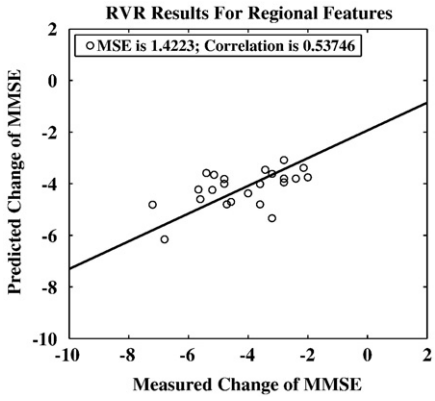

Fig. 11. MMSE change regression based on regional features.

Discussions and conclusion

This paper presented a general regression methodology by utilizing RVR with regional features to estimate continuous clinical variables from high-dimensional imaging patterns. The proposed method was validated on simulated datasets with known truth, where it was found to have excellent ability to estimate the relationship between the input patterns of atrophy and the output scalar score, when a very consistent relationship exists. It was also applied to data from the ADNI study.

In order to achieve a reasonably low dimensionality, a regional feature extraction methodology was used to sample spatial imaging patterns. Regions that correlated well with the estimated variables are sampled and integrated via a sparse regressor. Robustness and good generalization were achieved by a bagging methodology used on both the feature selection and the regression model construction. Alternative dimensionality reduction methods were also investigated, however the regional features provided the highest estimation accuracy.

We compared two pattern regression methods: RVR and SVR, which is a regression counterpart of the popular SVM classifier. RVR slightly outperforms SVR, and achieves robust estimation for different clinical variables. The best correlation between the estimated and the measured MMSE scores was around 0.73 obtained by using regional features. This correlation is as good as our expectation, in view of the significant noise present in cognitive measurements. Estimating BNT was quite promising, in the case of the limited data set. Our result also indicated that, pattern regression can predict future cognitive decline only by using baseline scans. Moreover, group difference maps could provide the most significant brain regions with respect to disease progression, which emphasizes the potential clinical importance of the proposed approach to find relatively early markers of structural brain changes.

Some of the conclusions concerning these methods combining different feature extraction and model learning approaches are discussed next.

Discussion on feature extraction

- PCA features are good at capturing general variance of the data, however, they might miss directions with particularly good

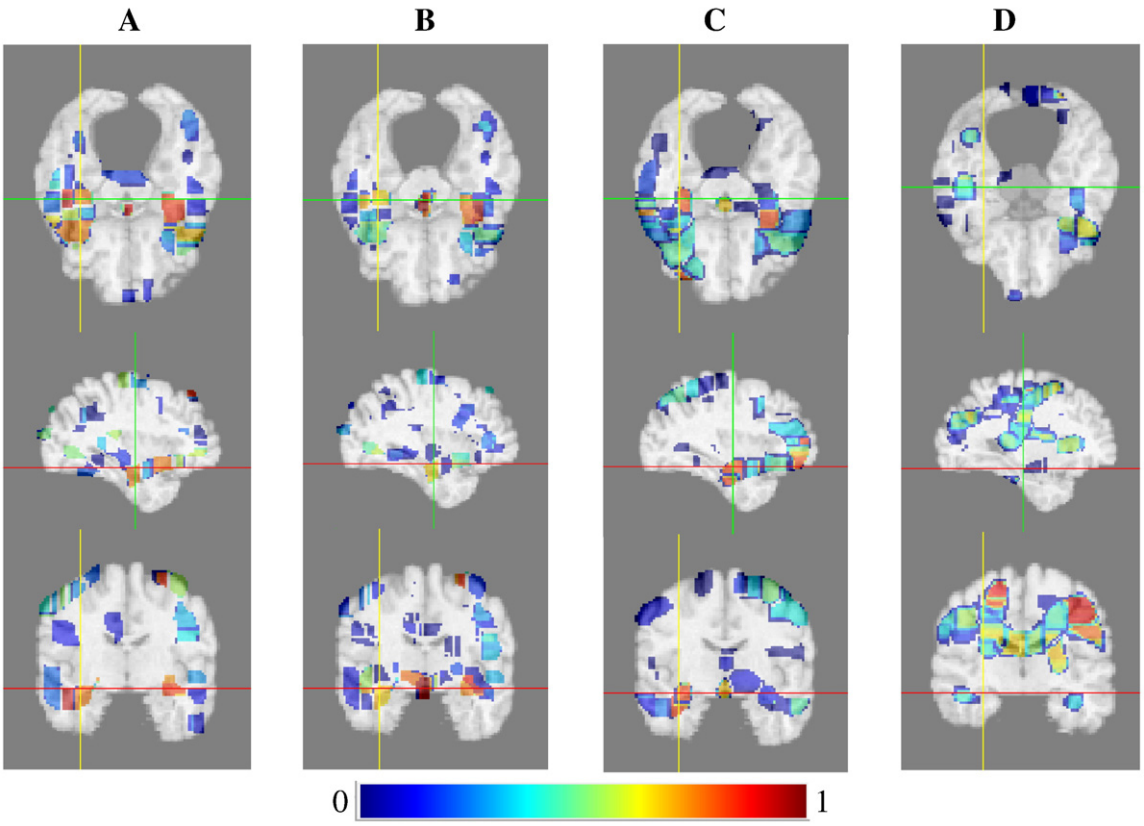

Fig. 12. Regions most representative of the group difference for different regression experiments based on ADNI data. (A: MMSE regression by using GM only; B: MMSE regression by using GM, WM and CSF; C: BNT regression by using GM; D: MMSE Change regression by using GM).

estimation value for the regressor, even though one would expect high variance along these directions. Moreover, PCA is unable to capture sufficient information from a small set of samples with extremely high-dimensional patterns, since it is limited by the number of samples. This is why PCA feature-based regression achieved acceptable performance when relatively larger training data sets were available (simulated data and MMSE estimation), but failed at BNT estimation.

- LDA is a better way to extract discriminative information because of its discriminatory nature. However, a number of pattern categories involved in the construction of the feature space is required before feature extraction (each corresponding to a different range of the regressor). Generally, this parameter was fixed according to prior knowledge of the disease under study. Hence, the lack of automatic parameter selection in LDA influences the further regression accuracy and limits its applicability.
- PCA and LDA have a common disadvantage that they assume data is linearly separable, however, this is not necessarily the case for most real-world problems. To address this problem, we tested a nonlinear manifold method, ISOMAP, but it did not work very well in most experiments. The reason is that, the success of ISOMAP depends on the correct computation of geodesic distances, which are approximated by the length of the shortest path between pairs of samples on a neighborhood graph. However, either *K–NN* or ε-neighbor approach is only valid when the training samples are dense enough. Undersampling of nonlinear manifolds results in erroneous connections when constructing the neighborhood graph, and finally lead to unstable manifold topology. In practice, for real image data, it is hard to preserve topological stability because of noise and outliers, especially for high-dimensional but limited sample. Consequently, ISOMAP failed in BNT estimation, and got the worst results in most experiments.
- Different from PCA, LDA and ISOMAP, the regional clustering method computes volumetric measures by adaptively grouping brain voxels into different regions based on the spatial consistency of brain structure, which is determined by the correlation between voxel-wise features and the continuous clinical scores being regressed. Thus, in theory, regional features are more discriminative and robust to the impact of noise and registration errors. In practice, taking all the experimental results into account, regional features outperform these three popular feature extraction methods, even in the very limited case, BNT estimation. However, these regional features were initially selected one by one in sequence. Therefore, if a collection of regions of weak individual estimation value, but of high collective value, exists in the data, it might be missed. Fully multivariate construction and selection of features is an extremely difficult problem for high-dimensional machine learning setting. Some recent formulation using non-negative matrix factorization might offer promising finding in the future (Batmanghelich et al., 2009).

Discussion on sparse model

A sparse model is also essential to get good generalization ability. To compare sparsity between SVR and RVR, many experiments were designed and analyzed in Experimental result analysis. From these experimental results, RVR was found to be a superior regression method with higher diagnostic power and better generalization ability. Although SVR achieved reasonably good performance both in simulated data and real data, it is less sparse than RVR. In practice, many more support vectors were used in SVR than relevance vectors were used in RVR, and the difference increased quickly as the size of

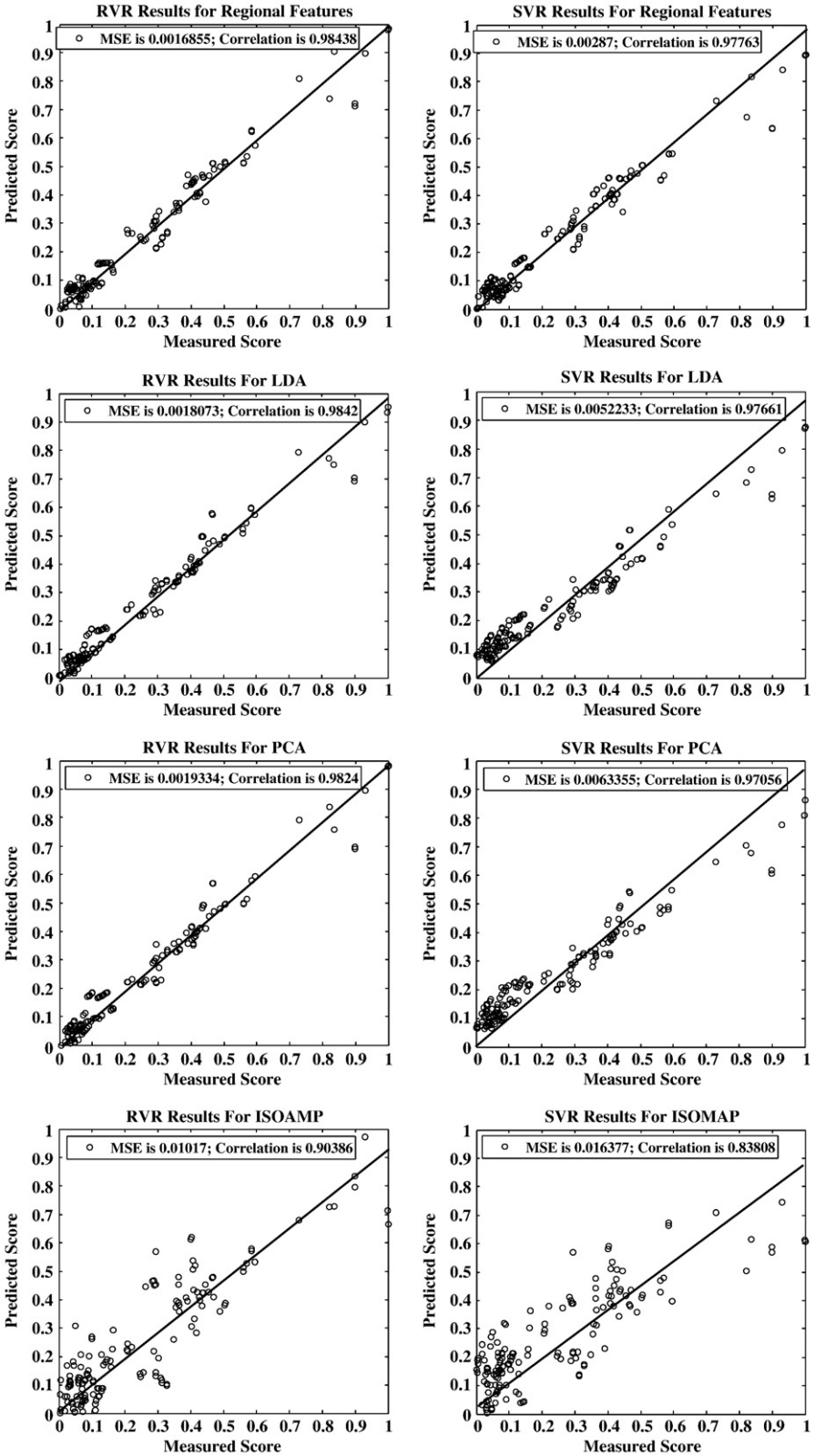

Fig. 13. Regression results based on simulated data. Scatter plots of true scores vs. estimated scores by RVR/SVR, with four kinds of feature extraction methods: adaptive regional clustering, LDA, PCA and ISOMAP. The sloping black solid lines represent the regression lines.

training samples grew. The reason is that the parameters of RVR are optimized by the principled Bayesian formulation with hierarchical hyperprior distributions, which leads to significantly fewer relevance vectors, thereby improving generalization ability. As discussed in Simulated clinical stage estimation with Fig. 3, our experiments also indicated that RVR is a stable regression method within a wide range of kernel sizes, in which estimation accuracy of continuous clinical scores varies smoothly.

Even if discriminative regional features and sparse regression model are used, it is still difficult to build a robust regressor, since there are relatively few sample with high dimensionality. Although our sample sizes might be adequate for classification, they are very limited for regression, especially in the case of progression estimation. To improve the robustness of regression models, a bagging framework with cross-validation is used to facilitate model parameter optimization, and to model expected variations in the training and testing images. By combining the outputs of multiple regressors learned from the subsets of training data, bagging based regression method can efficiently remove the influence of outliers.

In summary, bagging-based RVR with regional features was found to be a superior regression method in our experiments. Although this regression scheme is developed based on AD imaging data, it can be adapted and extended to other imaging problems. In the future, we plan to try Boosting-based algorithms in model learning, which are supposed to reduce both the bias and variance of unstable methods. Moreover, feature selection method with lower computation cost and better representation ability should be investigated, which is generally believed as a better way to improve model accuracy.

Acknowledgments

This study was financially supported by the NIH grant R01AG14971. The authors would like to thank Evi Parmpi for her help with data processing.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI; Principal Investigator: Michael Weiner; NIH grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering (NIBIB), and through generous contributions from the following: Pfizer Inc., Wyeth Research, Bristol-Myers Squibb, Eli Lilly and Company, GlaxoSmithKline, Merck Co. Inc., AstraZeneca AB, Novartis Pharmaceuticals Corporation, Alzheimer's Association, Eisai Global Clinical Development, Elan Corporation plc, Forest Laboratories, and the Institute for the Study of Aging, with participation from the U.S. Food and Drug Administration. Industry partnerships are coordinated through the Foundation for the National Institutes of Health. 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 of NeuroImaging at the University of California, Los Angeles.

Appendix A. Further details of relevance vector learning

In this appendix, we discuss how to optimize the weights of RVM, and show their L_1 sparsity as well.

To avoid over-fitting in the maximum likelihood estimation, a zero-mean Gaussian prior distribution is imposed on the weights W:

$$p(W|\alpha) = \prod_{n=0}^{N} N(w_n|0,\alpha_n) \tag{9}$$

where α is a vector of $N+1$ hyperparameters, which are associated independently with each weight, modifying the strength of the prior over its associated weight. In order to further simplify computation, Gamma prior distributions are assumed for α, and a noise variance of σ^{-2} (Tipping, 2001):

$$p(\alpha) = \prod_{n=0}^{N} \text{Gamma}(\alpha_n|a,b), p\left(\sigma^{-2}\right) = \text{Gamma}\left(\sigma^{-2}|c,d\right) \tag{10}$$

where a, b, c and d are constants and are usually set to zero. Thus the 'true' weight prior is:

$$p(w_i) = \int p(w_i|\alpha_i)p(\alpha_i)d\alpha_i = \frac{b^{\alpha}\Gamma(a+\frac{1}{2})}{(2\Pi)^{\frac{1}{2}}\Gamma(a)}\left(b+w_i^2/2\right)^{-(a+\frac{1}{2})} \tag{11}$$

Eq. (11) showed, although a non-sparse Gaussian prior is assigned over the weights, the "true" weight prior achieves a sparse Student t-test prior by integrating these hierarchical hyperpriors.

To solve the problem described by Eq. (5), the parameter posterior distribution $p(W, \alpha, \sigma^2 \mid \mathbf{t})$ needs to be computed based on the Bayesian framework. Unfortunately, it cannot be computed analytically, and thus an effective approximation must be made. According to the procedure described in Tipping (2001), the parameter posterior is decomposed as follows:

$$p\left(W,\alpha,\sigma^2|\mathbf{t}\right) = p\left(W|\mathbf{t},\alpha,\sigma^2\right)p\left(\alpha,\sigma^2|\mathbf{t}\right) \tag{12}$$

in which the first part can be expressed as:

$$p\left(W|\mathbf{t},\alpha,\sigma^2\right) = \frac{p\left(\mathbf{t}|W,\sigma^2\right)p(W|\alpha)}{p(\mathbf{t}|\alpha,\sigma^2)} = (2\pi)^{-(N+1)/2}|\Sigma|^{-1/2} \times \exp\left\{-\frac{1}{2}(W-\mu)^T\Sigma^{-1}(W-\mu)\right\} \tag{13}$$

where the posterior covariance and mean are respectively:

$$\Sigma = \left(\sigma^{-2}\Phi^{\mathrm{T}}\Phi + \mathbf{A}\right)^{-1} \tag{14}$$

$$\mu = \sigma^{-2}\Sigma\Phi^T\mathbf{t} \tag{15}$$

with $\mathbf{A} = \text{diag}(\alpha_0, \alpha_1,\ldots,\alpha_N)$. In order to evaluate Σ and μ, we need to find the hyperparameters which maximize the second part of Eq. (12), decomposed as follows:

$$p\left(\alpha,\sigma^2|\mathbf{t}\right) \propto p\left(\mathbf{t}|\alpha,\sigma^2\right)p(\alpha)p\left(\sigma^2\right) \tag{16}$$

Considering the hyperpriors over α and σ^2, $p(\alpha)$ and $p(\sigma^2)$ can be ignored, the parameter learning of RVR only needs to maximize the term $p(\mathbf{t}|\alpha, \sigma^2)$:

$$p\left(\mathbf{t}|\alpha,\sigma^2\right) = \int_{-\infty}^{+\infty} p\left(\mathbf{t}|W,\sigma^2\right)p(W|\alpha)dW = (2\pi)^{-N/2}|\sigma^2 I + \Phi\mathbf{A}^{-1}\Phi^T|^{-1/2} \exp\left\{-\frac{1}{2}\mathbf{t}^T\left(\sigma^2 I + \Phi\mathbf{A}^{-1}\Phi^T\right)^{-1}\mathbf{t}\right\} \tag{17}$$

This can be formulated as a typeII maximum likelihood procedure, that is, a most probable point estimate α_{MP} and $\sigma_{\mathbf{MP}}$ may be found throughout the maximization of the marginal likelihood with respect to $\log(\alpha)$ and $\log(\sigma^2)$. More optimization details can be found in Tipping (2001). The crucial information is that the optimal values of many hyperparameters are infinite. Accordingly, the parameter posterior may result in infinite peak at zero for many weights ω_i, which are ultimately responsible for the sparseness property of RVM.

Appendix B. Regression results based on simulated data (Fig. 13)

Appendix C. ADNI subject IDs used in experiments

Table 5
ADNI subject IDs used in MMSE experiment.

No.	Subject ID	Date of baseline scan	Clinical stage	No.	Subject ID	Date of baseline scan	Clinical stage
1	016_S_1028	2006-11-02	MCI	2	023_S_0887	2006-09-20	MCI
3	027_S_0485	2006-05-08	MCI	4	029_S_1384	2007-10-04	MCI
5	094_S_1398	2007-05-03	MCI	6	098_S_0269	2006-03-04	MCI
7	007_S_0293	2006-03-14	MCI	8	022_S_0750	2006-08-07	MCI
9	022_S_0924	2006-09-27	MCI	10	023_S_0855	2006-09-05	MCI
11	035_S_0997	2006-11-29	MCI	12	036_S_0656	2006-07-07	MCI
13	067_S_0077	2006-08-21	MCI	14	128_S_1148	2008-01-17	MCI
15	005_S_0222	2006-02-21	MCI	16	005_S_0324	2006-11-13	MCI
17	007_S_0041	2005-10-21	MCI	18	007_S_0344	2006-03-31	MCI
19	023_S_0331	2006-03-23	MCI	20	023_S_0388	2006-04-10	MCI
21	023_S_1126	2006-12-05	MCI	22	027_S_0256	2006-03-21	MCI
23	127_S_0394	2006-05-17	MCI	24	033_S_0567	2006-12-05	MCI
25	002_S_0729	2006-07-17	MCI	26	002_S_0954	2007-05-03	MCI
27	003_S_1057	2006-12-04	MCI	28	005_S_0572	2006-06-20	MCI
29	007_S_0414	2006-05-22	MCI	30	011_S_0861	2006-09-27	MCI
31	013_S_0325	2006-04-19	MCI	32	013_S_1120	2006-11-22	MCI
33	006_S_0675	2006-08-31	MCI	34	011_S_0326	2006-03-20	MCI
35	016_S_0702	2006-07-24	MCI	36	022_S_0544	2006-05-17	MCI
37	023_S_0126	2006-08-08	MCI	38	035_S_0033	2005-11-22	MCI
39	052_S_0671	2006-07-05	MCI	40	006_S_1130	2008-01-31	MCI
41	002_S_1155	2006-12-14	MCI	42	005_S_0448	2006-05-04	MCI
43	016_S_1121	2007-01-11	MCI	44	022_S_1394	2008-05-27	MCI
45	023_S_0625	2006-07-12	MCI	46	027_S_0179	2006-09-15	MCI
47	032_S_0187	2006-10-16	MCI	48	099_S_0111	2006-01-18	MCI
49	130_S_0102	2005-12-28	MCI	50	133_S_0638	2006-06-26	MCI
51	141_S_0697	2008-04-29	MCI	52	141_S_1244	2007-02-18	MCI
53	002_S_1070	2006-11-28	MCI	54	007_S_0128	2006-01-16	MCI
55	011_S_0856	2006-09-15	MCI	56	011_S_1080	2006-11-22	MCI
57	011_S_1282	2007-02-09	MCI	58	021_S_0332	2006-04-19	MCI
59	023_S_0078	2006-01-12	MCI	60	027_S_0461	2006-06-02	MCI
61	027_S_1387	2007-02-26	MCI	62	031_S_0568	2006-12-06	MCI
63	031_S_1066	2006-12-04	MCI	64	052_S_0952	2007-04-23	MCI
65	094_S_0434	2006-04-20	MCI	66	098_S_0667	2006-06-24	MCI
67	099_S_0054	2005-11-16	MCI	68	116_S_0752	2006-09-08	MCI
69	127_S_0393	2006-11-21	MCI	70	128_S_0947	2006-10-06	MCI
71	136_S_0195	2006-03-13	MCI	72	136_S_0873	2007-10-26	MCI
73	141_S_0982	2008-05-14	MCI	74	141_S_1004	2007-05-18	MCI
75	011_S_0053	2005-11-14	AD	76	011_S_0183	2006-03-03	AD
77	013_S_1161	2006-12-20	AD	78	016_S_0991	2006-11-01	AD
79	021_S_0753	2006-09-11	AD	80	022_S_0007	2005-09-13	AD
81	022_S_0543	2006-05-22	AD	82	023_S_0093	2006-01-03	AD
83	023_S_0139	2006-01-24	AD	84	023_S_0916	2006-09-22	AD
85	027_S_0404	2006-05-16	AD	86	027_S_1082	2006-12-13	AD
87	036_S_0577	2006-05-26	AD	88	036_S_0759	2006-08-22	AD
89	036_S_0760	2006-08-25	AD	90	062_S_0730	2006-07-19	AD
91	067_S_0029	2005-10-14	AD	92	067_S_0812	2006-10-02	AD
93	067_S_0828	2006-09-05	AD	94	109_S_0777	2006-09-07	AD
95	109_S_1157	2007-01-03	AD	96	116_S_1083	2006-12-21	AD
97	141_S_1137	2006-12-19	AD	98	011_S_0002	2005-08-26	CN
99	011_S_0016	2005-09-27	CN	100	011_S_0023	2005-10-31	CN
101	035_S_0048	2006-06-16	CN	102	035_S_0555	2006-06-06	CN
103	036_S_0576	2006-06-01	CN	104	036_S_0813	2006-08-25	CN
105	041_S_0125	2006-01-13	CN	106	067_S_0019	2005-10-12	CN
107	067_S_0056	2005-11-09	CN	108	082_S_0363	2006-03-28	CN
109	128_S_0863	2006-09-25	CN	110	011_S_0021	2005-10-10	CN
111	013_S_0502	2006-07-25	CN	112	016_S_0538	2006-08-02	CN
113	022_S_0066	2005-11-23	CN	114	022_S_0130	2006-01-23	CN
115	023_S_0031	2005-10-12	CN	116	022_S_0096	2006-01-18	CN
117	073_S_0312	2006-05-31	CN	118	109_S_0840	2006-09-07	CN
119	114_S_0601	2006-06-05	CN				

Table 6
ADNI subject IDs used in BNT experiment.

No.	Subject ID	Date of baseline scan	Clinical stage	No.	Subject ID	Date of baseline scan	Clinical stage
1	007_S_0414	2006-05-22	MCI	2	029_S_0878	2006-09-15	MCI
3	067_S_0077	2006-08-21	MCI	4	005_S_0448	2006-05-04	MCI
5	099_S_0054	2005-11-16	MCI	6	013_S_1120	2006-11-22	MCI
7	016_S_1117	2006-12-01	MCI	8	023_S_0030	2005-10-10	MCI
9	005_S_0324	2006-11-13	MCI	10	007_S_0698	2007-03-23	MCI
11	016_S_1121	2007-01-11	MCI	12	023_S_0388	2006-04-10	MCI
13	006_S_0675	2006-08-31	MCI	14	022_S_0044	2005-11-03	MCI
15	027_S_1045	2006-11-03	MCI	16	035_S_0033	2005-11-22	MCI
17	002_S_0782	2006-08-14	MCI	18	002_S_0954	2007-05-03	MCI
19	003_S_1122	2006-12-06	MCI	20	005_S_0222	2006-02-21	MCI
21	011_S_0326	2006-03-20	MCI	22	011_S_1282	2007-02-09	MCI
23	013_S_0860	2006-09-21	MCI	24	016_S_1028	2006-11-02	MCI
25	023_S_0887	2006-09-20	MCI	26	027_S_0179	2006-09-15	MCI
27	099_S_0111	2006-01-18	MCI	28	109_S_1114	2006-12-27	MCI
29	127_S_0394	2006-05-17	MCI	30	130_S_0102	2005-12-28	MCI
31	041_S_0697	2008-04-29	MCI	32	013_S_0996	2006-11-06	AD
33	016_S_0991	2006-11-01	AD	34	022_S_0543	2006-05-22	AD
35	023_S_0093	2006-01-03	AD	36	027_S_0404	2006-05-16	AD
37	027_S_1082	2006-12-13	AD	38	033_S_0724	2006-08-11	AD
39	036_S_0577	2006-05-26	AD	40	062_S_0730	2006-07-19	AD
41	067_S_0828	2006-09-05	AD	42	099_S_0372	2006-04-21	AD
43	099_S_0470	2006-05-10	AD	44	116_S_0392	2006-06-26	AD
45	141_S_1137	2006-12-19	AD	46	062_S_0690	2006-07-18	AD
47	067_S_0110	2006-01-25	AD	48	011_S_0053	2005-11-14	AD
49	141_S_0790	2006-09-10	AD	50	062_S_0690	2006-07-18	AD
51	109_S_1157	2007-01-03	AD	52	099_S_1144	2006-12-19	AD
53	114_S_0979	2006-11-01	AD	54	003_S_0907	2006-09-11	CN
55	011_S_0008	2005-09-13	CN	56	011_S_0022	2005-10-10	CN
57	013_S_0502	2006-07-25	CN	58	022_S_0066	2005-11-23	CN
59	023_S_0081	2006-01-10	CN	60	032_S_0677	2006-10-09	CN
61	041_S_0125	2006-01-13	CN	62	062_S_0768	2006-08-02	CN
63	067_S_0177	2006-03-10	CN	64	073_S_0312	2006-05-31	CN
65	082_S_0363	2006-03-28	CN	66	099_S_0352	2006-04-07	CN
67	109_S_0876	2006-09-19	CN	68	114_S_0173	2006-02-10	CN
69	128_S_0863	2006-09-25	CN				

Table 7
ADNI subject IDs used in future decline prediction.

No.	Subject ID	Date of baseline scan	Clinical stage
1	141_S_1244	2007-02-18	MCI Converter
2	141_S_0982	2006-11-15	MCI Converter
3	027_S_0256	2006-03-21	MCI Converter
4	127_S_0394	2006-05-17	MCI Converter
5	133_S_0913	2007-07-18	MCI Converter
6	023_S_0030	2005-10-10	MCI Converter
7	027_S_0461	2006-06-02	MCI Converter
8	099_S_0054	2005-11-16	MCI Converter
9	067_S_0077	2006-08-21	MCI Converter
10	128_S_0947	2006-10-06	MCI Converter
11	052_S_1054	2007-06-06	MCI Converter
12	011_S_1080	2006-11-22	MCI Converter
13	098_S_0269	2006-03-04	MCI Converter
14	023_S_0625	2006-07-12	MCI Converter
15	022_S_0750	2006-08-07	MCI Converter
16	127_S_0393	2006-11-21	MCI Converter
17	016_S_1028	2006-11-02	MCI Non-Converter
18	941_S_1295	2007-02-09	MCI Non-Converter
19	141_S_0851	2006-09-26	MCI Non-Converter
20	127_S_1427	2007-08-20	MCI Non-Converter
21	027_S_0485	2006-05-08	MCI Non-Converter
22	062_S_0730	2006-07-19	AD
23	067_S_0812	2006-10-02	AD
24	016_S_0991	2006-11-01	AD
25	027_S_1082	2006-12-13	AD
26	141_S_1137	2006-12-19	AD

References

ADNI, 2004. http://www.loni.ucla.edu/ADNI/.

Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. NeuroImage 38 (1), 95–113.

Ashburner, J., Friston, K., 2000. Voxel-based morphometry: the methods. NeuroImage 11, 805–821.

Batmanghelich, N., Taskar, B., Davatzikos, C., 2009. A general and unifying framework for feature construction, in image-based pattern classification. In: Prince, J.L., Pham, D.L., Myers, K.J. (Eds.), Information Processing in Medical Imaging. Vol. 5636 of Lecture Notes in Computer Science. Springer, pp. 423–434.

Breiman, L., 1996a. Bagging predictors. Mach. Learn. 26 (2), 123–140.

Breiman, L., 1996b. Out-of-bag Estimation. Technical Report. Statistics Department, University of California.

Chang, C.-C., Lin, C.-J., 2001. LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm.

Cox, D., Savoya, R., 2003. Functional magnetic resonance imaging (fMRI) brain reading? detecting and classifying distributed patterns of fMRI activity in human visual cortex. NeuroImage 19, 261–270.

Davatzikos, C., Genc, A., Xu, D., Resnick, S., 2001. Voxel-based morphometry using the RAVENS maps: methods and validation using simulated longitudinal atrophy. NeuroImage 14, 1361–1369.

Davatzikos, C., Fan, Y., Wu, X., Shen, D., Resnick, S., 2006. Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging. Neurobiol. Aging 29, 514–523.

Davatzikos, C., Resnick, S., Wu, X., Parmpi, P., Clark, C., 2008. Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. NeuroImage 41 (4), 1220–1227.

Davatzikos, C., Xu, F., An, Y., Fan, Y., Resnick, S.M., 2009. Longitudinal progression of Alzheimer's-like patterns of atrophy in normal older adults: the SPARE-AD index. Brain 132 (8), 2026–2035.

Davis, B., Fletcher, P., Bullitt, E., Joshi, S., 2007. Population shape regression from random design data. IEEE 11th International Conference on Computer Vision.

Duchesne, S., Caroli, A., Geroldi, C., Frisoni, G., Collins, D., 2005. Predicting clinical variable from MRI features: application to MMSE in MCI. Medical Image Computing and Computer-Assisted Intervention, pp. 392–399.

Duchesne, S., Caroli, A., Geroldi, C., Collins, D.L., Frisoni, G.B., 2009. Relating one-year cognitive change in mild cognitive impairment to baseline MRI features. NeuroImage 47 (4), 1363–1370.

Fan, Y., Shen, D., Gur, R.C., Gur, R.E., Davatzikos, C., 2007. Compare: Classification of morphological patterns using adaptive regional elements. IEEE Trans. Med. Imag. 26, 93–105.

Fan, Y., Resnick, S., Davatzikos, C., 2008. Feature selection and classification of multiparametric medical images using bagging and SVM. In: Reinhardt, J.M., Pluim, J.P.W. (Eds.), SPIE medical imaging, 6914. 69140Q.1-69140Q.10.

Fisher, R., 1936. The use of multiple measurements in taxonomic problems. Ann. Eugen. 7, 179–188.

Formisano, E., De Martino, F., Valente, G., 2008. Multivariate analysis of fMRI time series: classification and regression of brain responses using machine learning. Magn. Reson. Imaging 26 (7), 921–934.

Golland, P., Fischl, B., Spiridon, M., Kanwisher, N., Buckner, R.L., Shenton, M.E., Kikinis, R., Dale, A., Grimson, W.E.L., 2002. Discriminative analysis for image-based studies. International Conference on Medical Image Computing and Computer Assisted Intervention. Vol. 2488 of Lecture Notes in Computer Science, pp. 508–515.

Grau, V., Mewes, A., Alcaniz, M., Kikinis, R., Warfield, S., 2004. Improved watershed transform for medical image segmentation using prior information. IEEE Trans. Med. Imag. 23 (4), 447–458.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. Mach. Learn. 3, 1157–1182.

Guyon, I., Weston, J., Barnhill1, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machine. Mach. Learn. 46, 389–422.

Harris Drucker, C.J.B., Kaufman, L., Smola, A., Vapnik, V., 1996. Support vector regression machines. Adv. Neural Inf. Process. Syst. 9, 155–161.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. 24, 417–441.

Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., Whitwell, J.L., Ward, C., Dale, A.M., Felmlee, J.P., Gunter, J.L., Hill, D.L. G., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Krueger, G., Ward, H.A., Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J.P., Fleisher, A.S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., Weiner, M.W., 2008. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. J. Magn. Reson. Imag. 27 (4), 685–691.

Jovicich, J., Czanner, S., Greve, D., Haley, E., van der Kouwe, A., Gollub, R., Kennedy, D., Schmitt, F., Brown, G., MacFall, J., Fischl, B., Dale, A., 2006. Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. NeuroImage 30 (2), 436–443.

Kloppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Jack Jr., C.R., Frackowiak, R.S.J., 2008. Automatic classification of MRI scans in Alzheimer's disease. Brain 131 (3), 681–689.

Lao, Z., Shen, D., Xue, Z., Karacali, B., Resnick, S.M., Davatzikos, C., 2004. Morphological classification of brains via high-dimensional shape transformations and machine learning methods. NeuroImage 21, 46–57.

Liu, Y., Teverovskiy, L., Carmichael, O., Kikinis, R., Shenton, M., Carter, C.S., Stenger, V.A., Davis, S., Aizenstein, H., Becker, J.T., Lopez, O.L., Meltzer, C.C., 2004. Discriminative MR image feature analysis for automatic schizophrenia and Alzheimer's disease classification. Medical Image Computing and Computer-Assisted Intervention. Springer-Verlag GmbH, Saint-Malo, France.

Mangasarian, O.L., Musicant, D.R., 2000. Robust linear and support vector regression. IEEE Trans. Pattern Anal. Mach. Intell. 22, 950–955.

Mangasarian, O.L., Musicant, D.R., 2002. Large scale kernel regression via linear programming 46, 255–269.

Modrego, P.J., 2006. Predictors of conversion to dementia of probable Alzheimer type in patients with mild cognitive impairment. Curr. Alzheimer Res. 3, 161–170.

Petersen, R.C., Stevens, J.C., Ganguli, M., Tangalos, E.G., Cummings, J.L., DeKosky, S.T., 2001. Practice parameter: early detection of dementia: mild cognitive impairment (an evidence-based review). Neurology 56, 1133–1142.

Rakotomamonjy, A., 2003. Variable selection using SVM-based criteria. Mach. Learn. 3, 1357–1370.

Sled, J., Zijdenbos, A., Evans, A., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imag. 17 (1), 87–97.

Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Statistics and Computing 14, 199–222.

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

Thomaz, C.E., Boardman, J.P., Hill, D.L., Hajnal, J.V., Edwards, D.D., Rutherford, M.A., Gillies, D.F., Rueckert, D., 2004. Using a maximum uncertainty LDA-based approach to classify and analyse MR brain images. Medical Image Computing and Computer-Assisted Intervention. Vol. 3216 of Lecture Notes in Computer Science. Springer, pp. 291–300.

Tipping, M.E., 2001. Sparse Bayesian learning and the relevance vector machine. J. Mach. Learn. Res. 1, 211–244.

Tombaugh, T., McIntyre, N., 1992. The mini-mental state examination: a comprehensive review. J. Am. Geriatr. Soc. 40 (5), 922–935.

Vapnik, V., 1998. Statistical Learning Theory. Wiley, New York.

Vapnik, V., 2000. The Nature of Statistical Learning Theory. 2nd ed. Springer, New York.

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

Vemuri, P., Gunter, J.L., Senjem, M.L., Whitwell, J.L., Kantarci, K., Knopman, D.S., Boeve, B.F., Petersen, R.C., Jack Jr., C.R., 2009a. MRI and CSF biomarkers in normal, MCI, and AD subjects: diagnostic discrimination and cognitive correlations. Neurology 73, 287–293.

Vemuri, P., Gunter, J.L., Senjem, M.L., Whitwell, J.L., Kantarci, K., Knopman, D.S., Boeve, B.F., Petersen, R.C., Jack Jr., C.R., 2009b. MRI and CSF biomarkers in normal, MCI, and AD subjects: predicting future clinical change. Neurology 73, 294–301.

Vincent, L., Soille, P., 1991. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. IEEE Trans. Pattern Anal. Mach. Intell. 13 (6), 583–589.