University of Pennsylvania
UPenn Biostatistics Working Papers

Year 2015 *Paper* 42

Control-Group Feature Normalization for Multivariate Pattern Analysis Using the Support Vector Machine

Kristin A. Linn* Bilwaj Gaonkar† Jimit Doshi‡
Christos Davatzikos** Russell T. Shinohara††

*Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, klinn@upenn.edu

†Department of Neurosurgery, UCLA

‡Department of Radiology, University of Pennsylvania

**Department of Radiology, Perelman School of Medicine, University of Pennsylvania

††Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, rshi@upenn.edu

http://biostats.bepress.com/upennbiostat/art42

Control-Group Feature Normalization for Multivariate Pattern Analysis Using the Support Vector Machine

Kristin A. Linn, Bilwaj Gaonkar, Jimit Doshi, Christos Davatzikos, and Russell T. Shinohara

Abstract

Normalization of feature vector values is a common practice in machine learning. Generally, each feature value is standardized to the unit hypercube or by normalizing to zero mean and unit variance. Classification decisions based on support vector machines (SVMs) or by other methods are sensitive to the specific normalization used on the features. In the context of multivariate pattern analysis using neuroimaging data, standardization effectively up- and down-weights features based on their individual variability. Since the standard approach uses the entire data set to guide the normalization it utilizes the total variability of these features. This total variation is inevitably dependent on the amount of marginal separation between groups. Thus, such a normalization may attenuate the separability of the data in high dimensional space. In this work we propose an alternate approach that uses an estimate of the control-group standard deviation to normalize features before training. We also show that control-based normalization provides better interpretation with respect to the estimated multivariate disease pattern and improves the classifier performance in many cases.

Control-Group Feature Normalization for Multivariate Pattern Analysis Using the Support Vector Machine

Kristin A. Linn, Bilwaj Gaonkar, Jimit Doshi, Christos Davatzikos,
Russell T. Shinohara
For the Alzheimer's Disease Neuroimaging Initiative

Abstract

Normalization of feature vector values is a common practice in machine learning. Generally, each feature value is standardized to the unit hypercube or by normalizing to zero mean and unit variance. Classification decisions based on support vector machines (SVMs) or by other methods are sensitive to the specific normalization used on the features. In the context of multivariate pattern analysis using neuroimaging data, standardization effectively up- and down-weights features based on their individual variability. Since the standard approach uses the entire data set to guide the normalization it utilizes the total variability of these features. This total variation is inevitably dependent on the amount of marginal separation between groups. Thus, such a normalization may attenuate the separability of the data in high dimensional space. In this work we propose an alternate approach that uses an estimate of the control-group standard deviation to normalize features before training. We also show that control-based normalization provides better interpretation with respect to the estimated multivariate disease pattern and improves the classifier performance in many cases.

1 Introduction

Machine learning classification algorithms such as the support vector machine (SVM) [4, 43] are often used to map high-dimensional neuroimaging data to a clinical diagnosis or decision. Structural and functional magnetic resonance imaging (MRI) are promising tools for building biomarkers to diagnose, monitor, and treat neurological and psychological illnesses. Mass-univariate methods such as statistical parametric mapping [18–20] and voxel-based morphometry [1, 10] test for marginal disease effects at each voxel, ignoring complex spatial correlations and multivariate relationships among voxels. As a result, methods have emerged for performing multivariate pattern analysis (MVPA) that leverage the information contained in the covariance structure of the images to discriminate between the groups being studied [6, 7, 9, 11–14, 17, 21, 25–27, 29, 31, 36, 38–40, 44, 45, 47, 49]. Identifying multivariate structural and functional signatures in the brain that discriminate between groups may lead to a better understanding of disease processes and is therefore of great interest in the field of neuroimaging research.

The SVM is a common choice for estimating multivariate patterns in the brain because it is amenable to high-dimensional, low sample size data. Our focus in this work is on patterns in the brain that reflect structural changes due to disease. However, the methods apply more generally to applications of MVPA to BOLD measurements from fMRI or measures

of connectivity across the brain. The SVM takes as input image-label pairs and returns a decision function that is a weighted sum of the imaging features. The estimated weights reflect the joint contribution of the imaging features to the predicted class label.

Machine learning methods in general, and SVMs in particular, are sensitive to differences in feature scales. For example, a SVM will place more importance on a feature that takes values in the range of $[1000, 2000]$ than a feature that takes values in the interval $[1, 2]$. This is because the former tends to have a stronger influence on distances in high-dimensional space. To give all voxels or regions of interest equal importance during classifier training, it is common practice to implement feature-wise standardization in some way, either by normalizing each to have mean zero and unit variance or by scaling to a common domain. For example, Peng et al. [34] scale each feature to be in the interval $[0, 1]$, and Etzel et al. [16], Hanke et al. [23], Wang et al. [46], Zacharaki et al. [50] and Sato et al. [41] normalize to mean zero and unit variance. Such a preprocessing step, while common in practice, tends to be applied without weighing the consequent ramifications in a careful manner. Careful consideration must be given to the choice of feature normalization, as it is directly tied to the relative magnitude of the estimated SVM weights and thus the performance and interpretation of the classifier. While the original idea of feature scaling dates back to the universal approximation theorem from the neural network literature, it has not been explored in detail in the context of neuroimaging and MVPA. This is the object of this manuscript.

The rest of this paper is organized as follows: in Section 2, we provide a brief introduction to MVPA using the SVM. In Section 3, we review two popular feature normalization methods and propose an alternative based on the control-group variability. Using simulations, we compare the performance of different feature normalization techniques in Section 4. In Section 5, we investigate the effects of feature normalization by analyzing data from healthy controls and patients with Alzheimer's disease. We conclude with a discussion in Section 6.

2 Multivariate Pattern Analysis using the SVM

Let $(Y_i, \boldsymbol{X}_i^\mathsf{T})^\mathsf{T}$, $i = 1, \ldots, n$, denote n independent and identically distributed observations of the random vector $(Y, \boldsymbol{X}^\mathsf{T})^\mathsf{T}$, where $Y \in \{-1, 1\}$ denotes the group label, and $\boldsymbol{X} \in \mathbb{R}^p$ denotes a vectorized image with p voxels. A popular MVPA tool used in the neuroimaging community is the SVM [4, 43]. SVMs are known to work well for high dimension, low sample size data [42]. Such data are common in the neuroimaging-based diagnostic setting. Henceforth, we focus on MVPA using the SVM.

The hard-margin linear SVM solves the constrained optimization problem

$$\begin{aligned} &\arg\min_{\boldsymbol{v}, b} \frac{1}{2}||\boldsymbol{v}||^2 \\ &\text{such that } Y_i(\boldsymbol{v}^\mathsf{T}\boldsymbol{X}_i + b) \geq 1 \ \ \forall i = 1, \ldots, n, \end{aligned} \tag{1}$$

where $b \in \mathbb{R}$ and $\boldsymbol{v} \in \mathbb{R}^p$ are parameters that describe the classification function. For a given set of training data, let the solution to (1) be denoted by $(\tilde{\boldsymbol{v}}, \tilde{b})$. Then, for a new observation

$\boldsymbol{X}^{\text{new}}$ with unknown label Y^{new}, the classification function $c(\boldsymbol{X}^{\text{new}}) = \text{sign}(\tilde{\boldsymbol{v}}^{\mathsf{T}}\boldsymbol{X}^{\text{new}} + \tilde{b})$ returns a predicted group label.

When the data from the two groups are not linearly separable, the soft-margin linear SVM allows some observations to be misclassified during training through the use of slack variables ξ_i with associated penalty parameter C. In this case, the optimization problem becomes

$$\begin{aligned}
&\arg\min_{\boldsymbol{v},b,\boldsymbol{\xi}} \frac{1}{2}||\boldsymbol{v}||^2 + C\sum_{i=1}^{n}\xi_i \\
&\text{such that:} \\
&Y_i(\boldsymbol{v}^{\mathsf{T}}\boldsymbol{X}_i + b) \geq 1 - \xi_i \ \ \forall i = 1,\ldots,n, \\
&\xi_i \geq 0 \ \ \forall i = 1,\ldots,n,
\end{aligned} \tag{2}$$

where $C \in \mathbb{R}$ is a tuning parameter that penalizes misclassification, and $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_n)^{\mathsf{T}}$ is the vector of slack variables. For details about solving optimization problems (1) and (2) we refer the reader to Hastie et al. [24].

In high-dimensional problems where the number of features is greater than the number of observations, the data are almost always separable by a linear hyperplane [32]. Thus, MVPA is often applied using the hard-margin linear SVM in (1). Select examples include classification of multiple sclerosis patients into disease subgroups [2], the study of Alzheimer's disease [7, 9], and various classification tasks involving patients with depression [5, 22, 28]. This is only a small subset of the relavant literature, which demonstrates the widespread popularity of the approach.

3 SVM Feature Normalization for MVPA

The choice of feature normalization affects the estimated weight pattern of a SVM and can lead to vastly different conclusions about the underlying disease process. Two widely implemented approaches are to (i) normalize each feature to have mean zero and unit variance, and (ii) scale each feature to have a common domain such as $[0, 1]$. Henceforth, we will refer to (i) as *standard normalization* and (ii) as *domain standardization* [33].

Let μ_j and σ_j denote the mean and standard deviation of the j^{th} feature, $j = 1,\ldots,p$. Denote the corresponding empirical estimates by $\overline{X}_j = n^{-1}\sum_{i=1}^{n} X_{i,j}$ and $\widehat{\sigma}_j = \{(n-1)^{-1}\sum_{i=1}^{n}(X_{i,j} - \overline{X}_j)^2\}^{1/2}$. Then, subject i's standard-normalized j^{th} feature is calculated as

$$X^{Z}_{i,j} = \frac{X_{i,j} - \overline{X}_j}{\widehat{\sigma}_j}.$$

Alternatively, subject i's domain-scaled j^{th} feature is calculated as

$$X^{U}_{i,j} = \frac{X_{i,j} - \min_i X_{i,j}}{\max_i X_{i,j} - \min_i X_{i,j}}.$$

One potential drawback of using domain scaling is the instability of the minimum and maximum order statistics, especially in small sample sizes. This may introduce bias in the estimated weight pattern by up- and down-weighting features in an unstable way. In comparison, the standard normalization may seem relatively stable. However, it implicitly depends on the relative sample size of each group and the separability between groups. To see this, let f_{X_j} denote the marginal distribution of X_j, with mean μ_j and variance σ_j^2. Let $f_{X_j \mid Y=y}$ denote the conditional distribution of X_j given $Y = y$ with mean $\mu_{j,y}$ and variance $\sigma_{j,y}^2$. In addition, let $\gamma = \text{pr}(Y = 1)$. Then, $\mu_j = \gamma\mu_{j,1} + (1-\gamma)\mu_{j,-1}$ and

$$\begin{aligned}
\sigma_j^2 &= E(X_j - \mu_j)^2 \\
&= EX_j^2 - \mu_j^2 \\
&= \int x_j^2\{\gamma f_{X_j \mid Y=1}(x) + (1-\gamma) f_{X_j \mid Y=-1}(x)\}dx - \mu_j^2 \\
&= \gamma(\sigma_{j,1}^2 + \mu_{j,1}^2) + (1-\gamma)(\sigma_{j,-1}^2 + \mu_{j,-1}^2) - \mu_j^2.
\end{aligned}$$

After simplification, the previous expression can be written as

$$\sigma_j^2 = \gamma\sigma_{j,1}^2 + (1-\gamma)\sigma_{j,-1}^2 + \gamma(1-\gamma)(\mu_{j,1} - \mu_{j,-1})^2 \tag{3}$$

The right-hand side of expression (3) shows that the variance of feature j depends on a mixture of the conditional variances of both classes and a term that depends on the squared distance between their marginal means. Larger marginal separability of feature j will lead to a larger estimate of the pooled standard deviation used for normalization. Thus, normalizing by the pooled standard deviation can in some cases harshly penalize, or down-weight, features that have good separability, leading to a loss in predictive performance. We demonstrate this using simulated data examples in Section 4.

The right-hand side of equation (3) also illuminates how normalization is dependent on the relative within-group sample sizes, which may have adverse effects on classifier performance. Suppose data for MVPA are available from a case-control study where the cases have been oversampled. That is, there is one healthy control for each subject with the disease. Suppose further that the true disease prevalence in the population is rare. Then, the estimate of σ_j^2 will be an equal mixture of the group variances $\sigma_{j,1}^2$ and $\sigma_{j,-1}^2$, whereas the true σ_j^2 in the population depends more heavily on the control-group variance, $\sigma_{j,-1}^2$. Methods for dataset or covariate shift address this issue by weighting individual data points to reflect the distribution of covariates in the population [30, 37]. However, these methods are usually implemented after feature normalization. As a result, the estimated decision rule may be undesirably influenced by the use of a biased estimate of the pooled variance.

As an alternative, we propose normalizing the j^{th} feature as follows:

$$X_{i,j}^C = \frac{X_{i,j} - \overline{X}_j}{\widehat{\sigma}_j^C},$$

for all subjects $i = 1, \ldots, n$, where $\overline{X}_j$ is the pooled sample mean of feature j, and $\widehat{\sigma}_j^C$ is the sample standard deviation of the j^{th} feature calculated using the control-group data only.

Note that $\overline{X}_j$ and $\widehat{\sigma}_j^C$ are computed using only the control-group, but the normalization is applied to subjects from both groups. We refer to this as *control normalization*. Note that for features that contribute greatly to the separability of the groups, the control-group standard deviation will be smaller than the pooled-group standard deviation. Scaling by this smaller value will implicitly up-weight the most discriminative features in comparison to the standard-normalization. In some studies or applications, there may not be a control group. In this case, a reference group may be chosen based on expert knowledge or the scientific goals of the study. In Section 4 we demonstrate how the choice of feature normalization technique may lead to a tradeoff of classifier properties such as sensitivity and specificity.

Figure 1 displays an example of the influence of feature normalization on the estimated SVM weight pattern. We generated $n_0 = 50$ independent control-group observations and

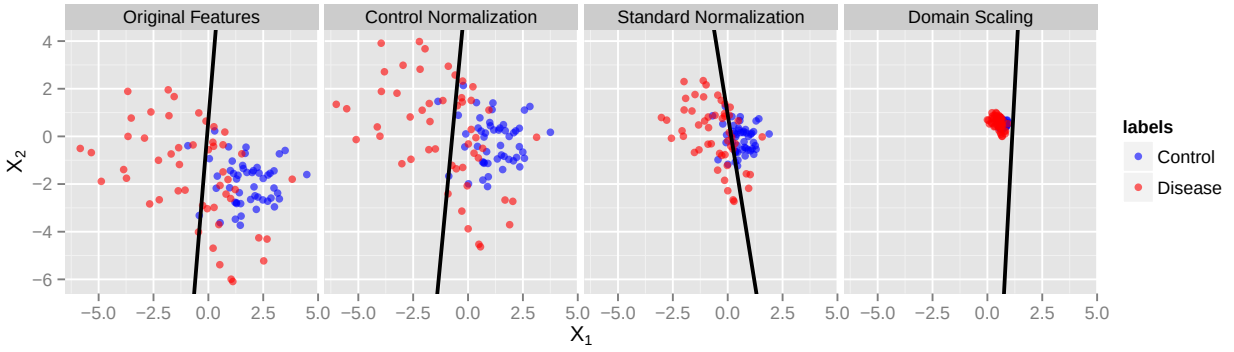

Figure 1: *Influence of feature normalization on the SVM decision boundary. From left to right: original feature scales, control-normalized features, standard-normalized features, domain-scaled features.*

$n_1 = 50$ disease-group observation from models (6) and (7), respectively, defined in Section 4. All features are independent noise features except the first two, X_1 and X_2, which are plotted pre-normalization in the first panel of Figure 1. The correlation between X_1 and X_2 in the control group is $\rho_0 = -0.2$, and it is $\rho_1 = -0.6$ in the disease-group. The control-normalized, standard-normalized, and domain-scaled versions are plotted in the second, third, and fourth panels. The estimated SVM decision boundary is projected onto the space of these two features and is given by the black line in each panel. We carefully chose these parameters because they represent a scenario where the choice of feature normalization changed the sign of the estimated optimal SVM line. While the difference in results may not always be so drastic, this example motivates the need for researchers to adopt a single, interpretable technique for feature normalization when performing multivariate pattern analysis using SVMs. We study the effects of feature normalization for a range of parameter settings in Section 4.

Another advantage of the control normalization is the resulting interpretability of the feature values. Fixing all other SVM features at a constant value, the estimated weight

corresponding to feature j conveys the magnitude and direction of change in the decision function score for a one unit increase in feature j, where the units are in terms of the control-group standard deviation of that feature. In many studies, it is likely that more knowledge exists about the distribution of values in the normal population, as the disease being studied may be highly heterogeneous, rare, or not yet well-understood. Being able to interpret the estimated disease pattern relative to the healthy control distribution may improve the reproducibility and clinical value of the MVPA results.

4 Simulations

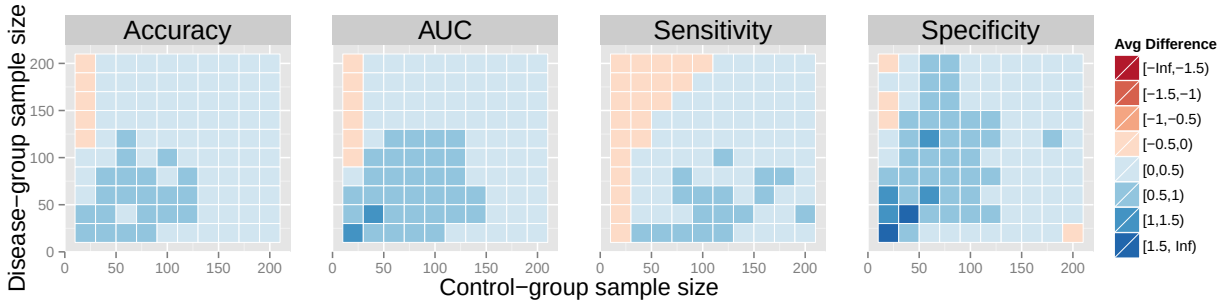

Figure 2: *Average difference in performance measures between the control normalization and standard normalization for a range of sample sizes. Data generated from models 4 and 5.*

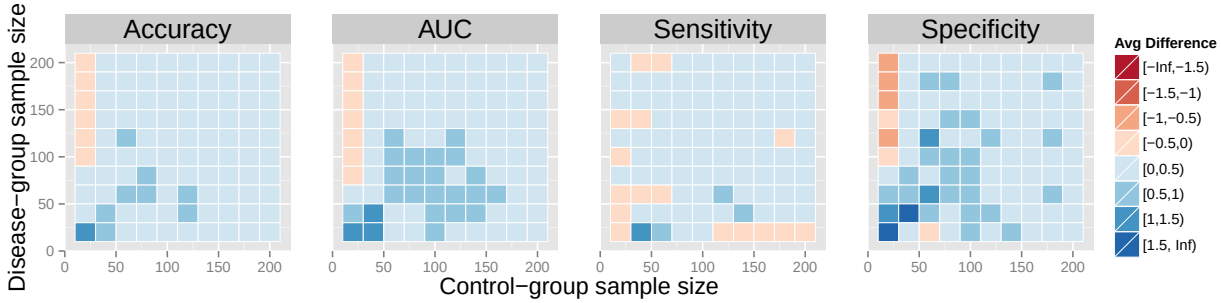

Figure 3: *Average difference in performance measures between the control normalization and domain scaling for a range of sample sizes. Data generated from models 4 and 5.*

In this section, we study a range of data-generating models to compare the performance of the control normalization, standard normalization, and domain scaling when using the linear SVM for MVPA. For all simulations, we generate p features, $(X_1, X_2, \ldots, X_p)^\mathsf{T}$, the first two of which have varying levels of joint discriminative power. The remaining $p-2$

are independent noise features. The first two features are generated as mixtures of multivariate normal distributions. The following steps describe the procedure used to obtain the results in Figures 2–5. For each of M=1,000 iterations, n_0 control subjects are generated as independent draws from the model

$$\begin{pmatrix} X_1^0 \\ X_2^0 \end{pmatrix} \sim \text{Normal}\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_0 \\ \rho_0 & 1 \end{pmatrix} \right\}, \qquad X_j^0 \sim \text{Normal}(0,1), \quad j = 3, \ldots, p. \quad (4)$$

Non-control group subjects are generated as n_1 independent draws from the model

$$\begin{pmatrix} X_1^1 \\ X_2^1 \end{pmatrix} \sim \text{Normal}\left\{ \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{pmatrix} \right\}, \qquad X_j^1 \sim \text{Normal}(0,1), \quad j = 3, \ldots, p, \quad (5)$$

where X_j^0, X_j^1, $j = 3, \ldots, p$, are all mutually independent. Additionally, we generate $t_0 = 500$ independent control-group samples from model (4) and $t_1 = 500$ independent samples from model (5) for testing. We then train an SVM using the $n_0 + n_1$ training samples using the `scikit learn` library in Python, which internally calls `libSVM` [3]. When $n_0 \neq n_1$, we train a class-weighted SVM that weights the cost parameter by $(n_1 + n_0)/n_0$ for the control group and by $(n_1 + n_0)/n_1$ for the disease group. In Figures 2 and 3 the correlations are fixed at $\rho_0 = 0$, $\rho_1 = 0$, and we vary $n_0, n_1 \in \{20, 40, \ldots, 200\}$. In Figures 4 and 5 the sample sizes are fixed at $n_0 = 50$, $n_1 = 50$, and we vary $\rho_0, \rho_1 \in \{-0.9, -0.8, \ldots, 0.9\}$.

We compare the average difference in accuracy, area under the ROC curve (AUC), sensitivity, and specificity of the SVM on the test set. Given the true test labels, accuracy is defined as the percentage of correct classifications using the SVM decision rule learned from the training data. Sensitivity is the percentage of correct positive predictions, and specificity is the percentage of correct negative predictions. The ROC curve is the proportion of true positives as a function of the false positive rate which ranges in $[0, 1]$ as the SVM intercept b is varied across the real line. Larger values of the criteria are desirable and indicate better classifier performance.

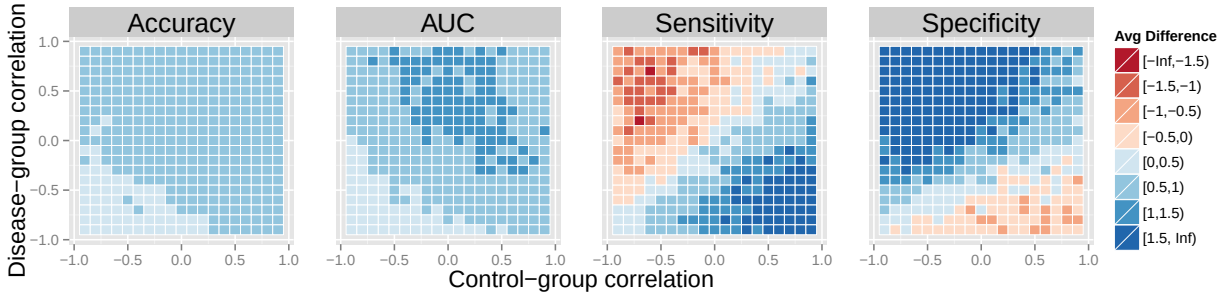

Figure 4: *Average difference in performance measures between the control normalization and standard normalization for a range of feature correlations. Data generated from models 4 and 5.*

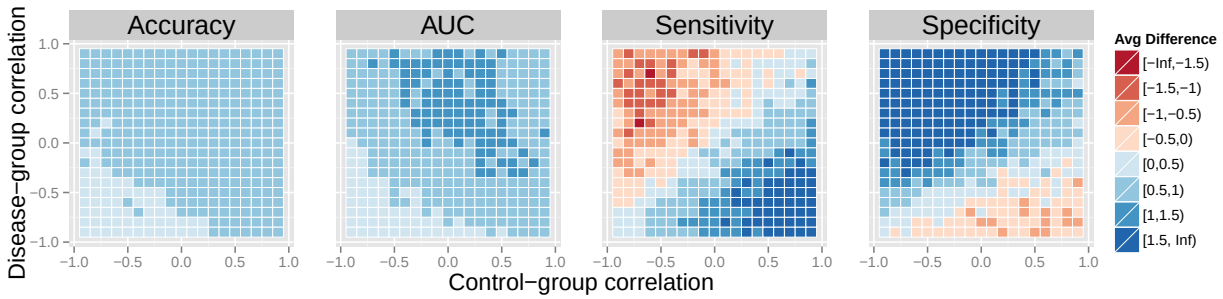

Figure 5: *Average difference in performance measures between the control normalization and domain scaling for a range of feature correlations. Data generated from models 4 and 5.*

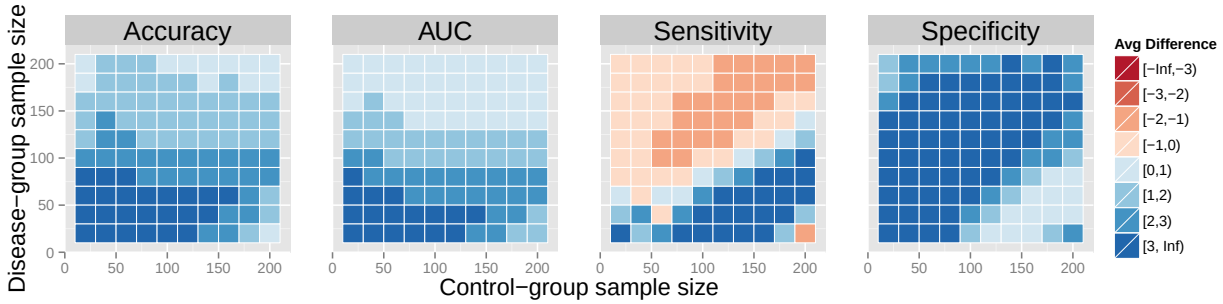

Figure 6: *Average difference in performance measures between the control normalization and standard normalization for a range of sample sizes. Data generated from models 6 and 7.*

Each colored square in the heatmaps represents a self-contained simulation with 1,000 iterations. The color indicates the average difference between a given performance measure between the control-normalized SVM and either the standard-normalized or domain-scaled SVM. Dark blue indicates superior performance of the control normalization. Across the simulations summarized in Figures 2 and 3, average accuracies ranged from approximately 60%–80%, average AUCs ranged from approximately 70%-80%, average sensitivities ranged from approximately 55%-80%, and average specificities ranged from approximately 55%-80%. In Figures 2 and 3, the control normalization performs better on average than the standard normalization and domain scaling for most combinations of within-group sample size. Notable exceptions are when the sample size of the control-group is much smaller than that of the disease-group. The standard normalization appears to improve sensitivity when the disease-group sample size is large but seemingly at the cost of reduced specificity. Overall, the results appear similar when comparing the control normalization to domain scaling.

Next, we present a case where the control normalization demonstrates significant improvement over the alternative feature standardizations. The following procedure was used to obtain the results in Figures 6–7. For each of M=1,000 iterations, n_0 control subjects are

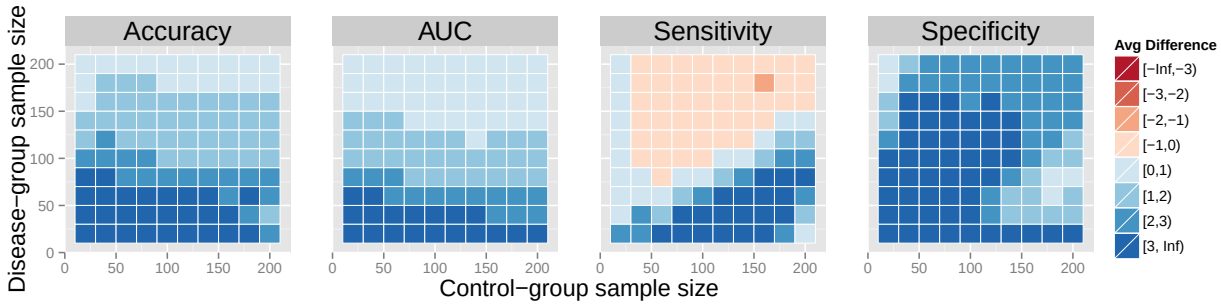

Figure 7: *Average difference in performance measures between the control normalization and domain scaling for a range of sample sizes. Data generated from models 6 and 7.*

generated as independent draws from the model

$$\begin{pmatrix} X_1^0 \\ X_2^0 \end{pmatrix} \sim \text{Normal}\left\{ \begin{pmatrix} 2 \\ -2 \end{pmatrix}, \begin{pmatrix} 1 & \rho_0 \\ \rho_0 & 1 \end{pmatrix} \right\}, \qquad X_j^0 \sim \text{Normal}(0,1), \quad j = 3, \ldots, p. \quad (6)$$

Non-control group subjects are generated as n_1 independent draws from the model

$$\begin{pmatrix} X_1^1 \\ X_2^1 \end{pmatrix} \sim \text{Normal}\left\{ \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} 5 & \rho_1 \\ \rho_1 & 5 \end{pmatrix} \right\}, \qquad X_j^1 \sim \text{Normal}(0,1), \quad j = 3, \ldots, p. \quad (7)$$

Additionally, we generate $t_0 = 500$ independent control-group samples from model (6) and $t_1 = 500$ independent samples from model (7) for testing. We vary the within-group correlation parameters, $\rho_0, \rho_1 \in \{-0.9, -0.8, \ldots, 0.9\}$. For this set of model parameters, the control-normalization demonstrates greater than three percent improvement in accuracy, AUC, and specificity for a variety of sample sizes when compared to the standard normalization and domain scaling. In some cases, the gains in specificity are due to lower sensitivity. However, the overall accuracy and AUC in these cases are still improved.

5 Case Study

The Alzheimer's Disease Neuroimaging Initiative (ADNI) (http://www.adni.loni.usc.edu) is a multi-million dollar study funded by a number of public and private resources from the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), Food and Drug Administration (FDA), the pharmaceutical industry, and non-profit organizations. Aims of the study include developing sensitive and specific image-based biomarkers for early diagnosis of Alzheimer's disease (AD), as well as monitoring the progression of mild cognitive impairment (MCI) and AD. Understanding and predicting disease trajectories is imperative for the discovery of effective treatments that intervene in the early stages of the disease to prevent irreversible damage to the brain.

The ADNI data are publicly available and as a result have been thoroughly analyzed in the neuroimaging literature [48]. A detailed comparison of SVM classification results using

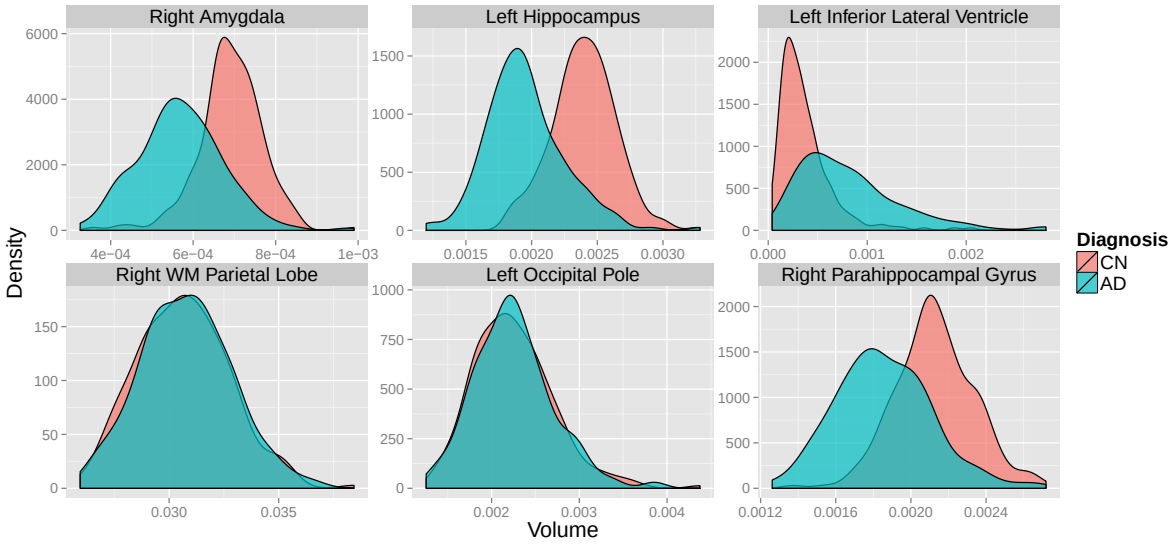

Figure 8: *Density plots of ROI volumes by group.*

different categories of imaging features is given in Cuingnet et al. [7]. In this section, we compare the performance of different SVM feature normalization techniques using volumes obtained from a multi-atlas segmentation pipeline applied to structural MRIs from the ADNI database [15].

The final dataset used for this analysis consists of labels indicating the presence or absence of AD and the volumes of 137 regions of interest (ROIs) in the brain for each subject. Each region is divided by the subject's total intracranial volume to adjust for differences in individual brain size. The data consist of 230 healthy controls (CN) and 200 patients diagnosed with AD with ages ranging between 55 and 90. AD is associated with atrophy in the brain, and thus the AD group has smaller volumes on average in particular ROIs compared to the CN group.

To give intuition about the differences between the control and standard normalization procedures in the ADNI data, we plot the densities of six features, stratified by group, in Figure 8. Whereas the pooled and control-group estimated variability is nearly identical for features such as the white matter parietal lobe and occipital pole, the control-group variability is less than the pooled variability for more marginally separable features such as the amygdala, hippocampus, inferior lateral ventricle, and parahippocampal gyrus. Thus, we expect a SVM trained after control normalization to place relatively heavier weights on these marginally discriminative features than a SVM trained after standard normalization. Figure 9 displays SVM weight patterns from the three methods. Based on Figure 9, it appears all methods obtain similar estimated disease patterns with a few subtle differences. Table 1 lists the top 10 features in order of the magnitude of their weights. As anticipated, the control normalization places more emphasis on the two amygdala regions because their marginal separability ensures a smaller denominator is used in the control-group normalization step,

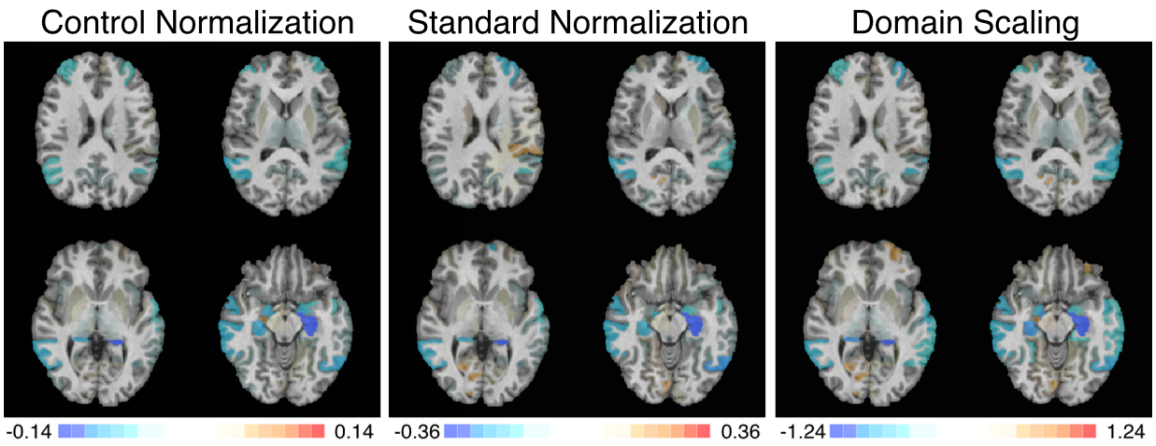

Figure 9: *SVM weight patterns for discriminating between AD and CN subjects by feature standardization method.*

up-weighting these features compared to the standard normalization.

Rank	Control Normalization	Standard Normalization	Domain Scaling
1	Left Hippocampus	Left Hippocampus	Left Hippocampus
2	Right Hippocampus	Left Inferior Temporal Gyrus	Right Hippocampus
3	Left Inferior Lateral Ventricle	Left Inferior Lateral Ventricle	Left Inferior Lateral Ventricle
4	Left Inferior Temporal Gyrus	Left Middle Frontal Gyrus	Left Inferior Temporal Gyrus
5	Left Amygdala	Right Hippocampus	Left Middle Frontal Gyrus
6	Right Amygdala	Left Superior Frontal Gyrus	Right Middle Temporal Gyrus
7	Right Middle Temporal Gyrus	Right Middle Temporal Gyrus	Left Superior Temporal Gyrus
8	Left Middle Frontal Gyrus	Left Amygdala	Right Amygdala
9	Right Angular Gyrus	Left Superior Temporal Gyrus	Left Amygdala
10	Right Inferior Lateral Ventricle	Right Calcarine Cortex	Left Middle Temporal Gyrus

Table 1: *Top 10 ranked features by the SVM weights in decreasing absolute value.*

We compare the control normalization proposed in Section 3 to the standard normalization and domain scaling. Table 2 displays 5-fold cross-validated estimates of classifier accuracy, area under the curve (AUC), sensitivity, and specificity. The control normalization and domain scaling outperform the standard normalization across all performance measures, increasing the cross-validated accuracy, sensitivity, and specificity by more than one percent. By a small margin, domain-scaling performs best in terms of prediction for this dataset but at the cost of fitted model interpretability. To quantify the uncertainty in the estimates in Table 2, we repeated the 5-fold cross-validation proceedure 1000 times using random subsamples of 140 patients and 140 controls. The point estimates are shown with a single standard error on each side in Figure 10. For this particular data set, the performance differences are not statistically significant across the three methods.

Method	Accuracy	AUC	Sensitivity	Specificity
Control Normalization	88%	94%	84%	93%
Domain Scaling	89%	94%	85%	93%
Standard Normalization	87%	94%	82%	91%

Table 2: *5-fold cross-validation results.*

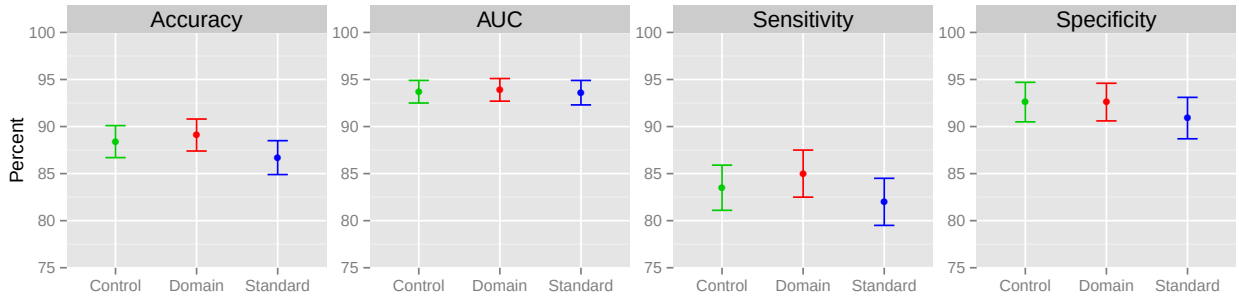

Figure 10: *5-fold cross-validation results with measures of uncertainty estimated by subsampling the original data.*

6 Discussion

The roots of feature scaling for preprocessing lie in the neural network literature of the 1990s. The universal approximation theorem, which broadly states that simple neural networks can approximate a rich set of functions, was initially proven for functions defined on a unit hypercube domain using a multilayer perceptron constructed with sigmoidal neurons [8]. It became natural to assume that centering and scaling of data would lead to a faster convergence even though neural networks are theoretically affine invariant [35]. For the most part, the optimization turned out to be more graceful with these scaled inputs, since it slowed down network saturation and avoided the vanishing gradients problem to a certain extent.

However, applying scaling to kernel methods such as SVMs or distance-based methods such as k-means tends to yield completely different results depending on the scaling method used. This is because these methods are not transformation-invariant. In such a case, scaling essentially imposes a form of soft feature selection since it implicitly changes the metric used for computing the kernel matrix. This fact is important in the context of image-based diagnosis using SVMs with region of interest (ROI) data. Scaling implicitly enforces the fact that variation in the amygdala, which is a relatively small structure in terms of volume, is as important as that in the prefrontal lobe, which is much larger in volume. Thus, appropriate scaling of features is an important but under-emphasized issue that we have attempted to call attention to in this manuscript.

It is critical for researchers wishing to interpret the results of MVPA from SVMs to

understand how the choice of feature normalization influences the results, as well as how to determine the best method for their scientific question. We have proposed a control-based normalization and demonstrated several advantages of the approach. Most notably, we have highlighted the possibility of improved classifier performance according to criteria such as accuracy and AUC for a comprehensive set of data generating distributions. The control normalization improves classifier performance by giving higher weight, relative to other standardization techniques, to features with greater marginal separability between groups. Depending on the underlying data generating distribution and relative sample size between groups, different classifiers will experience tradeoffs between sensitivity and specificity. The optimal choice of feature normalization may depend on the unknown data generating distribution as well as certain clinical considerations. As a result, the interpretability of the control normalization is an attractive property that makes it amenable to a vast majority of clinical applications. Given the overall increase in AUC demonstrated by the control normalization in the simulations, it is possible that simply shifting the estimated optimal hyperplane would lead to a classifier that has the same sensitivity as the standard-normalized hyperplane but with increased specificity.

Interpretability of estimated disease patterns is a desirable quality for most applications of MVPA to neuroimaging data. Standardizing features by the control-group variability improves interpretability over other feature normalization methods. We showed in Section 3 that the standard normalization method depends on the relative sample size of the two groups as well as the marginal separability; in contrast, the control normalization is unaffected by these qualities of the data and hence provides better generalizability across samples. We believe that including a control normalization step in the MVPA preprocessing pipeline is a simple alternative to current practice that promises increased interpretability, generalizability, and performance of the results.

7 Acknowledgements

The authors would like to acknowledge funding by NIH grant R01 NS085211 and a seed grant from the Center for Biomedical Image Computing and Analytics at the University of Pennsylvania. This work represents the opinions of the researchers and not necessarily that of the granting institutions.

Data used in the preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimers disease (AD). For up-to-date information, see www.adni-info.org.

References

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

[2] Bendfeldt, K., Klöppel, S., Nichols, T. E., Smieskova, R., Kuster, P., Traud, S., Mueller-Lenke, N., Naegelin, Y., Kappos, L., Radue, E.-W., et al. (2012). Multivariate pattern classification of gray matter pathology in multiple sclerosis. *Neuroimage*, 60(1):400–408.

[3] Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

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

[5] Costafreda, S. G., Chu, C., Ashburner, J., and Fu, C. H. (2009). Prognostic and diagnostic potential of the structural neuroanatomy of depression. *PLoS One*, 4(7):e6353.

[6] Craddock, R. C., Holtzheimer, P. E., Hu, X. P., and Mayberg, H. S. (2009). Disease state prediction from resting state functional connectivity. *Magnetic Resonance in Medicine*, 62(6):1619–1628.

[7] Cuingnet, R., Rosso, C., Chupin, M., Lehricy, S., Dormont, D., Benali, H., Samson, Y., and Colliot, O. (2011). Spatial regularization of {SVM} for the detection of diffusion alterations associated with stroke outcome. *Medical Image Analysis*, 15(5):729 – 737. Special Issue on the 2010 Conference on Medical Image Computing and Computer-Assisted Intervention.

[8] Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314.

[9] Davatzikos, C., Bhatt, P., Shaw, L. M., Batmanghelich, K. N., and Trojanowski, J. Q. (2011). Prediction of {MCI} to {AD} conversion, via mri, {CSF} biomarkers, and pattern classification. *Neurobiology of Aging*, 32(12):2322.e19 – 2322.e27.

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

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

[12] Davatzikos, C., Ruparel, K., Fan, Y., Shen, D., Acharyya, M., Loughead, J., Gur, R., and Langleben, D. D. (2005). Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. *Neuroimage*, 28(3):663–668.

[13] Davatzikos, C., Xu, F., An, Y., Fan, Y., and 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.

[14] De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., and Formisano, E. (2008). Combining multivariate voxel selection and support vector machines for mapping and classification of fmri spatial patterns. *Neuroimage*, 43(1):44–58.

[15] Doshi, J., Erus, G., Ou, Y., and Davatzikos, C. (2013). Ensemble-based medical image labeling via sampling morphological appearance manifolds. In *MICCAI Challenge Workshop on Segmentation: Algorithms, Theory and Applications. Nagoya, Japan*.

[16] Etzel, J. A., Valchev, N., and Keysers, C. (2011). The impact of certain methodological choices on multivariate analysis of fmri data with support vector machines. *Neuroimage*, 54(2):1159–1167.

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

[18] Frackowiak, R., Friston, K., Frith, C., Dolan, R., and Mazziotta, J., editors (1997). *Human Brain Function*. Academic Press USA.

[19] Friston, K. J., Frith, C., Liddle, P., and Frackowiak, R. (1991). Comparing functional (pet) images: the assessment of significant change. *Journal of Cerebral Blood Flow & Metabolism*, 11(4):690–699.

[20] Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., and Frackowiak, R. S. (1994). Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping*, 2(4):189–210.

[21] Gaonkar, B. and Davatzikos, C. (2013). Analytic estimation of statistical significance maps for support vector machine based multi-variate image analysis and classification. *NeuroImage*, 78(0):270 – 283.

[22] Gong, Q., Wu, Q., Scarpazza, C., Lui, S., Jia, Z., Marquand, A., Huang, X., McGuire, P., and Mechelli, A. (2011). Prognostic prediction of therapeutic response in depression using high-field mr imaging. *Neuroimage*, 55(4):1497–1503.

[23] Hanke, M., Halchenko, Y. O., Sederberg, P. B., Olivetti, E., Fründ, I., Rieger, J. W., Herrmann, C. S., Haxby, J. V., Hanson, S. J., and Pollmann, S. (2009). Pymvpa: a unifying approach to the analysis of neuroscientific data. *Frontiers in neuroinformatics*, 3.

[24] Hastie, T., Tibshirani, R., and Friedman, J. (2001). Springer New York Inc.

[25] Klöppel, S., Stonnington, C. M., Chu, C., Draganski, B., Scahill, R. I., Rohrer, J. D., Fox, N. C., Jack, C. R., Ashburner, J., and Frackowiak, R. S. J. (2008). Automatic classification of MR scans in Alzheimer's disease. *Brain*, 131(3):681–689.

[26] Koutsouleris, N., Meisenzahl, E. M., Davatzikos, C., Bottlender, R., Frodl, T., Scheuerecker, J., Schmitt, G., Zetzsche, T., Decker, P., Reiser, M., et al. (2009). Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. *Archives of general psychiatry*, 66(7):700–712.

[27] Langs, G., Menze, B. H., Lashkari, D., and Golland, P. (2011). Detecting stable distributed patterns of brain activation using gini contrast. *NeuroImage*, 56(2):497–507.

[28] Liu, F., Guo, W., Yu, D., Gao, Q., Gao, K., Xue, Z., Du, H., Zhang, J., Tan, C., Liu, Z., et al. (2012). Classification of different therapeutic responses of major depressive disorder with multivariate pattern analysis method based on structural mr scans. *PLoS One*, 7(7):e40968.

[29] Mingoia, G., Wagner, G., Langbein, K., Maitra, R., Smesny, S., Dietzek, M., Burmeister, H. P., Reichenbach, J. R., Schlösser, R. G., Gaser, C., et al. (2012). Default mode network activity in schizophrenia studied at resting state using probabilistic ica. *Schizophrenia research*, 138(2):143–149.

[30] Moreno-Torres, J. G., Raeder, T., Alaiz-RodríGuez, R., Chawla, N. V., and Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530.

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

[32] Orrù, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G., and Mechelli, A. (2012). Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neuroscience & Biobehavioral Reviews*, 36(4):1140–1152.

[33] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[34] Peng, X., Lin, P., Zhang, T., and Wang, J. (2013). Extreme learning machine-based classification of adhd using brain structural mri data.

[35] Perantonis, S. J. and Lisboa, P. J. (1992). Translation, rotation, and scale invariant pattern recognition by high-order neural networks and moment classifiers. *Neural Networks, IEEE Transactions on*, 3(2):241–251.

[36] Pereira, F. (2007). *Beyond brain blobs: machine learning classifiers as instruments for analyzing functional magnetic resonance imaging data*. ProQuest.

[37] Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. (2009). *Dataset shift in machine learning*. The MIT Press.

[38] Reiss, P. T. and Ogden, R. T. (2010). Functional generalized linear models with images as predictors. *Biometrics*, 66(1):61–69.

[39] Richiardi, J., Eryilmaz, H., Schwartz, S., Vuilleumier, P., and Van De Ville, D. (2011). Decoding brain states from fmri connectivity graphs. *Neuroimage*, 56(2):616–626.

[40] Sabuncu, M. R. and Van Leemput, K. (2011). The relevance voxel machine (rvoxm): a bayesian method for image-based prediction. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011*, pages 99–106. Springer.

[41] Sato, J. R., Kozasa, E. H., Russell, T. A., Radvany, J., Mello, L. E., Lacerda, S. S., and Amaro Jr, E. (2012). Brain imaging analysis can identify participants under regular mental training. *PloS one*, 7(7):e39832–e39832.

[42] Schölkopf, B., Tsuda, K., and Vert, J.-P. (2004). *Kernel methods in computational biology*. MIT press.

[43] Vapnik, V. (2000). *The nature of statistical learning theory*. springer.

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

[45] Venkataraman, A., Rathi, Y., Kubicki, M., Westin, C.-F., and Golland, P. (2012). Joint modeling of anatomical and functional connectivity for population studies. *Medical Imaging, IEEE Transactions on*, 31(2):164–182.

[46] Wang, L., Shen, H., Tang, F., Zang, Y., and Hu, D. (2012). Combined structural and resting-state functional mri analysis of sexual dimorphism in the young adult human brain: an mvpa approach. *Neuroimage*, 61(4):931–940.

[47] Wang, Z., Childress, A. R., Wang, J., and Detre, J. A. (2007). Support vector machine learning-based fmri data group analysis. *NeuroImage*, 36(4):1139–1151.

[48] Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., Liu, E., et al. (2013). The alzheimer's disease neuroimaging initiative: a review of papers published since its inception. *Alzheimer's & Dementia*, 9(5):e111–e194.

[49] Xu, L., Groth, K. M., Pearlson, G., Schretlen, D. J., and Calhoun, V. D. (2009). Source-based morphometry: The use of independent component analysis to identify gray matter differences with application to schizophrenia. *Human brain mapping*, 30(3):711–724.

[50] Zacharaki, E. I., Wang, S., Chawla, S., Soo Yoo, D., Wolf, R., Melhem, E. R., and Davatzikos, C. (2009). Classification of brain tumor type and grade using mri texture and shape in a machine learning scheme. *Magnetic Resonance in Medicine*, 62(6):1609–1618.