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

Addressing Confounding in Predictive Models with an Application to Neuroimaging

DOI 10.1515/ijb-2015-0030

Abstract: Understanding structural changes in the brain that are caused by a particular disease is a major goal of neuroimaging research. Multivariate pattern analysis (MVPA) comprises a collection of tools that can be used to understand complex disease effects across the brain. We discuss several important issues that must be considered when analyzing data from neuroimaging studies using MVPA. In particular, we focus on the consequences of confounding by non-imaging variables such as age and sex on the results of MVPA. After reviewing current practice to address confounding in neuroimaging studies, we propose an alternative approach based on inverse probability weighting. Although the proposed method is motivated by neuroimaging applications, it is broadly applicable to many problems in machine learning and predictive modeling. We demonstrate the advantages of our approach on simulated and real data examples.

Keywords: Multivariate pattern analysis (MVPA), structural magnetic resonance imaging (MRI), confounding, inverse probability weighting, support vector machine (SVM), machine learning, predictive modeling

1 Introduction

Quantifying population-level differences in the brain that are attributable to neurological or psychiatric disorders is a major focus of neuroimaging research. Structural magnetic resonance imaging (MRI) is widely used to investigate changes in brain structure that may aid the diagnosis and monitoring of disease. A structural MRI of the brain consists of many voxels, where a voxel is the three dimensional analogue of a pixel. Each voxel has a corresponding intensity, and jointly the voxels encode information about the size and structure of the brain. Functional MRI (fMRI) also plays an important role in the understanding of disease mechanisms by revealing relationships between disease and brain function. In this work we focus on structural MRI data, but many of the concepts apply to fMRI studies.

One way to assess group-level differences in the brain is to take a “mass-univariate” approach, where statistical tests are applied separately at each voxel. This is the basic idea behind statistical parametric mapping (SPM) [1–3] and voxel-based morphometry (VBM) [4, 5]. Voxel-based methods are limited in the sense that they do not make use of information contained jointly among multiple voxels. Figure 1 illustrates this concept using toy data with two variables, X_1 and X_2 . Marginally, X_1 and X_2 discriminate poorly between the groups, but perfect linear separability exists when X_1 and X_2 are considered jointly. Thus, there has been a shift away from voxel-wise methods to multivariate pattern analysis (MVPA) in the neuroimaging community. In general, MVPA refers to any approach that is able to identify disease effects that are manifested as spatially distributed patterns across multiple brain regions [6–27].

The goal of MVPA is often two-fold: (i) to understand underlying patterns in the brain that characterize a disease, and (ii) to develop sensitive and specific image-based biomarkers for disease diagnosis, the prediction of disease progression, or prediction of treatment response. Although the MVPA literature often

*Corresponding author: Russell T. Shinohara, Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, E-mail: rshi@upenn.edu

Kristin A. Linn, Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania
Bilwaj Gaonkar, Jimit Doshi, Christos Davatzikos, Department of Radiology, Perelman School of Medicine, University of Pennsylvania

Figure 1: Marginally, X_1 and X_2 discriminate poorly between the groups, but perfect separability is attained when X_1 and X_2 are considered jointly.

uses terminology that suggests a causal interpretation of disease patterns in the brain, little has been done to formalize a causal framework for neuroimaging, with the notable exception of recent work by Weichwald et al. [62]. In this paper, we elucidate subtle differences between the two goals of MVPA and provide guidance for future implementation of MVPA in neuroimaging studies. We focus attention on the consequences of confounding on goal (i) and give a few remarks regarding goal (ii).

Confounding of the disease-image relationship by non-imaging variables such as age and gender can have undesirable effects on the output of MVPA. In particular, confounding may lead to identification of false disease patterns, undermining the usefulness and reproducibility of MVPA results. We discuss the implications of “regressing out” confounding effects using voxel-wise parametric models, a widely used approach for addressing confounding, and propose an alternative based on inverse probability weighting.

The structure of this paper is the following. Section 2 provides a brief overview of the use of MVPA in neuroimaging with focus on the use of the support vector machine (SVM) as a tool for MVPA. In Section 3, we address the issue of confounding by reviewing current practice in neuroimaging and proposing an alternative approach. In Section 4, we illustrate our method using simulated data, and Section 5 presents an application to data from an Alzheimer's disease neuroimaging study. We conclude with a discussion in Section 6.

2 Multivariate pattern analysis in neuroimaging

Let $(Y_i, \mathbf{X}_i^T, \mathbf{A}_i^T)^T$, $i=1, \dots, n$, denote n independent and identically distributed observations of the random vector $(Y, \mathbf{X}^T, \mathbf{A}^T)^T$, where $Y \in \{-1, 1\}$ denotes the group label, e.g., control versus disease, $\mathbf{X} \in \mathbb{R}^p$ denotes a vectorized image with p voxels, and $\mathbf{A} \in \mathbb{R}^r$ denotes a vector of non-image variables such as age and gender. Suppose Y and $\mathbf{A}$ both affect $\mathbf{X}$. For example, Alzheimer's disease is associated with patterns of atrophy in the brain that are manifested in structural MRIs. It is well known that age also affects brain structure [28]. Our primary aim is to develop a framework for studying multivariate differences in the brain between disease groups that are attributable solely to the disease and not to differences in non-imaging

variables between the groups. Thus, we advocate for creating balance between the groups with respect to non-imaging variables before performing MVPA. More formal details are given in the next section.

A popular MVPA tool used by the neuroimaging community is the support vector machine (SVM) [29, 30]. This choice is partly motivated by the fact that SVMs are known to work well for high dimension, low sample size data [31]. Often, the number of voxels in a single MRI can exceed one million depending on the resolution of the scanner and the protocol used to obtain the image. The SVM is trained to predict the group label from the vectorized set of voxels that comprise an image. Alternatives include penalized logistic regression [32] as well as functional principal components and functional partial least squares [33, 34]. Henceforth, we focus on MVPA using the SVM.

The hard-margin linear SVM solves the constrained optimization problem

$$\min_{\mathbf{v}, b} \frac{1}{2} \|\mathbf{v}\|^2$$

$$\text{such that } Y_i(\mathbf{v}^\top \mathbf{X}_i + b) \geq 1 \quad \forall i = 1, \dots, n, \quad (1)$$

Where $b \in \mathbb{R}$, and $\mathbf{v} \in \mathbb{R}^p$ are feature weights that describe the relative contribution of each voxel to the classification function. 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

$$\min_{\mathbf{v}, b, \xi} \frac{1}{2} \|\mathbf{v}\|^2 + C \sum_{i=1}^n \xi_i$$

such that :

$$Y_i(\mathbf{v}^\top \mathbf{X}_i + b) \geq 1 - \xi_i \quad \forall i = 1, \dots, n,$$

$$\xi_i \geq 0 \quad \forall i = 1, \dots, n, \quad (2)$$

where $C \in \mathbb{R}$ is a tuning parameter that penalizes misclassification, and $\xi = (\xi_1, \xi_2, \dots, \xi_n)^\top$. For details about solving optimization problems (1) and (2) we refer the reader to Hastie et al. [35].

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 [36]. Thus, MVPA is often applied using the hard-margin linear SVM in (1). For example, this is the approach implemented by: Bendfeldt et al. [37] to classify subgroups of multiple sclerosis patients; Cuingnet et al. [7] and Davatzikos et al. [8] in Alzheimer's disease applications; and Liu et al. [38], Gong et al. [39], and Costafreda et al. [40] for various classification tasks involving patients with depression. This is only a small subset of the relevant literature, which illustrates the widespread popularity of the approach.

3 Multivariate pattern analysis and confounding

3.1 Causal framework for descriptive aims

When the goal of MVPA is to understand patterns of change in the brain that are attributable to a disease, the ideal dataset would contain two images for each subject: one where the subject has the disease and another at the same point in time where the subject is healthy. Of course, this is the fundamental problem of causal inference, as it is impossible to observe both of these potential outcomes [41, 42]. In addition, confounding of the disease–image relationship presents challenges. Figure 2 depicts confounding of the Y – X relationship by a single confounder, A . Training a classifier in the presence of confounding may lead to biased estimation of the underlying disease pattern. This occurs when classifiers rely heavily on regions that are strongly correlated with confounders instead of regions that encode subtle disease changes [43].

Figure 2: The relationship between Y (disease) and X (image) is confounded by A (e.g., age), which affects both Y and X .

Failing to address confounding in MVPA can lead to a false understanding of image signatures that characterize the disease and a lack of generalizability of the estimated classifier.

Let $X_i(y)$ denote the image that would have been observed had subject i been observed with group status $Y_i = y$, possibly contrary to fact. Let $F_{X(-1)}$ and $F_{X(1)}$ denote the distributions of the counterfactual images $X(-1)$ and $X(1)$, respectively. Assume there exists a unique hyperplane in $\mathbb{R}^p$ that maximally separates the counterfactual distributions in the sense that the centers of the two distributions lie on opposite sides of the hyperplane and the total combined mass on the “wrong” side of the hyperplane is minimized. The following notation will be useful for defining our target parameter. Let S be a map from the space of two distributions with the same support to this unique separating hyperplane, $S: (F_D, F_{D'}) \mapsto \mathbb{R}^p$ for distributions D and D' . Define $\theta^* = S(F_{X(-1)}, F_{X(1)})$. Thus, θ^* is the hyperplane that maximally separates the counterfactual image distributions, assuming this unique hyperplane exists, and is our target parameter.

The target parameter θ^* is inherently of interest in MVPA, which aims to understand patterns of change associated with a disease in a population of interest. Due to their cost, imaging studies often focus on populations that are at-risk for a particular disease. One example is the Alzheimer's Disease Neuroimaging Initiative (ADNI, <http://www.adni.loni.usc.edu>), which studied patients with mild cognitive impairment (MCI) and who were therefore at-risk for Alzheimer's disease. The MCI group was comprised of male and female patients across a wide age range with various other heterogeneities. Although non-imaging covariates are often easy to collect, the marginal parameter θ^* is usually of interest in MVPA, as opposed to a parameter that is conditional on the non-imaging variables. Modeling conditional changes in the brain due to the disease would require more assumptions or stratification; the latter case reduces the sample size which may already be limited by budget constraints.

We do not directly observe samples from $F_{X(-1)}$ and $F_{X(1)}$, but under certain identifying assumptions, we can estimate the counterfactual distributions using the observed data. In particular, assume

$$(i) \quad X_i = \frac{X_i(1) + X_i(-1)}{2} + Y_i \frac{X_i(1) - X_i(-1)}{2},$$

$$(ii) \quad \{X_i(1), X_i(-1)\} \perp\!\!\!\perp Y_i \mid \mathbf{A}_i,$$

for all $i = 1, \dots, n$. Assumption (i) is the usual consistency assumption, and (ii) is the assumption of no unmeasured confounding, i.e., ignorability of exposure given measured confounders. Using (i) and (ii),

$$\begin{aligned} F_{X(y)} &= \text{pr}\{X(y) \leq x\} \\ &= E[\text{pr}\{X(y) \leq x \mid \mathbf{A}\}] \\ &= E[\text{pr}\{X(y) \leq x \mid Y=y, \mathbf{A}\}] \quad (ii) \\ &= E\{\text{pr}(X \leq x \mid Y=y, \mathbf{A})\}. \quad (i) \end{aligned}$$

Note that the expectation is over the marginal distribution of $\mathbf{A}$ rather than the conditional distribution of $\mathbf{A}$ given $Y=y$. Thus, we reweight the integrand as follows:

$$\begin{aligned} E\{\text{pr}(X \leq x \mid Y=y, \mathbf{A})\} &= E\left\{\text{pr}(X \leq x \mid Y=y, \mathbf{A}) \frac{\text{pr}(Y=y \mid \mathbf{A}) \text{pr}(Y=y)}{\text{pr}(Y=y \mid \mathbf{A}) \text{pr}(Y=y)}\right\} \\ &= \frac{1}{\text{pr}(Y=y)} \int \text{pr}(X \leq x, Y=y \mid \mathbf{A}) \frac{\text{pr}(Y=y)}{\text{pr}(Y=y \mid \mathbf{A})} dP_{\mathbf{A}} \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{\text{pr}(Y=y)} \int \text{pr}(\mathbf{X} \leq x, Y=y | \mathbf{A}) dP_{\mathbf{A}}^* \\
&= F_{\mathbf{X}|Y=y}^*,
\end{aligned}$$

where $F_{\mathbf{X}|Y=y}^*$ is the conditional distribution of $\mathbf{X}$ given $Y=y$ that results from averaging over a weighted version of the distribution of $\mathbf{A}$. The weights are the inverse of the probability of being in observed group $Y=y$ given confounders $\mathbf{A}$, multiplied by the normalizing constant, $\text{pr}(Y=y)$. We assume positivity, meaning $\text{pr}(Y=y|\mathbf{A})$ is bounded away from zero for all possible values of $\mathbf{A}$. We have shown under assumptions (i) and (ii) that $F_{\mathbf{X}(-1)}$ and $F_{\mathbf{X}(1)}$ are identifiable from the observed data. Thus, our target parameter corresponds to $\theta^* = S(F_{\mathbf{X}|Y=-1}^*, F_{\mathbf{X}|Y=1}^*)$, which can be estimated by $\hat{\theta}^* = S(\hat{F}_{\mathbf{X}|Y=-1}^*, \hat{F}_{\mathbf{X}|Y=1}^*)$.

To illustrate the effects of confounding on MVPA, consider a toy example with a single confounder A . Let $\mathbf{X}$ consist of two features, $\mathbf{X} = (X_1, X_2)^T$, and define the corresponding potential outcomes, $\mathbf{X}(Y) = \{X_1(Y), X_2(Y)\}^T$. In the study of Alzheimer's disease, A might be age, Y an indicator of disease group, and X_1 and X_2 gray matter volumes of two brain regions. We generate $N = 1,000$ independent observations from the generative model

$$\begin{aligned}
X_1 &= 4 - Y + \epsilon_1, & X_2 &= -.25 - 7A - .25Y - 2AY + \epsilon_2, \\
A &\sim \text{Unif}[0, 1], & Y &\sim \text{Unif}\{-1, 1\} \\
\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} &\sim \text{Normal}\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}\right\}.
\end{aligned} \tag{3}$$

Note that model (3) has the property that Y and A are independent, so that A is not a confounder of the $Y - \mathbf{X}$ relationship. Next, we generate an additional $N = 1,000$ independent observations from model (3) except with $Y = 2Y^* - 1$, where $Y^* \sim \text{Bernoulli}(A)$, so that A is a confounder of the $Y - \mathbf{X}$ relationship in this second sample. The first sample is plotted in the top three panels of Figure 3 and the linear SVM decision boundary estimated from the unconfounded data is drawn in gray in the top right panel. The $Y - \mathbf{X}$ relationship is confounded by A in second sample which is displayed in the bottom three panels of Figure 3. Here, A mimics the confounding effect of age in Alzheimer's disease in two ways: (i) we give

Figure 3: Top row: unconfounded data generated from model (3). **Bottom row:** data with the $Y - \mathbf{X}$ relationship confounded by A . The target parameter is the linear SVM decision rule learned from the data in the top right plot, shown in gray. The black line is the linear SVM decision rule learned from the confounded sample in the bottom right plot.

larger values of A a higher probability of being observed with $Y = 1$, and (ii) A has a decreasing linear effect on X_2 . The decision boundary estimated from the confounded sample is shown in black in the bottom right panel. Confounding by A shifts the estimated decision boundary and obscures the true relationship between the features X_1 , X_2 and outcome Y .

There is some variation in the definition of confounding in the imaging literature, making it unclear in some instances if, when, and why an adjustment is made. For example, some researchers recommend correcting images for age effects even after age-matching patients and controls [44]. In an age-matched study, age is not a confounder, and adjusting for its relationship with $\mathbf{X}$ is unnecessary. To address confounding, one approach proposed in the neuroimaging literature is to “regress-out” the effects of confounders from the image $\mathbf{X}$. This is commonly done by fitting a (usually linear) regression of voxel intensity on confounders separately at each voxel and subtracting the fitted value at each location [44, 3]. The resulting “residual image” is then used in MVPA. Formally, the following model is fit using least squares, separately for each $j = 1, \dots, p$:

$$X_j = \beta_{0,j} + \beta_{1,j}^T \mathbf{A} + \epsilon_j, \quad (4)$$

where the ϵ_j are assumed to be independent for all j . The least squares estimates $\hat{\beta}_{0,j}$ and $\hat{\beta}_{1,j}$ define the j^{th} residual voxel,

$$\tilde{X}_j = X_j - (\hat{\beta}_{0,j} + \hat{\beta}_{1,j}^T \mathbf{A}).$$

Combining all residuals gives the vector $\tilde{\mathbf{X}} = (\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_p)$ which is used as the feature vector to train the MVPA classifier. We henceforth refer to this method as the *adjusted MVPA*.

A similar procedure is to fit model (4) using the control group only [44]. We refer to this approach as the *control-adjusted MVPA*. In applications where there is not a clear control group, i.e., comparing two disease subclasses, a single reference group is chosen. Let $\hat{\beta}_{0,j}^c$ and $\hat{\beta}_{1,j}^c$ denote the least squares estimates of $\beta_{0,j}$ and $\beta_{1,j}$ when model (4) is fit using only control-group data. The control-group adjusted features used in the MVPA classifier are then $\tilde{\mathbf{X}}^c = (\tilde{X}_1^c, \tilde{X}_2^c, \dots, \tilde{X}_p^c)$, where $\tilde{X}_j^c = X_j - (\hat{\beta}_{0,j}^c + \hat{\beta}_{1,j}^{cT} \mathbf{A})$.

A comparison of the adjusted and control-adjusted MVPA features is displayed in Figure 4. The first two plots of Figure 4 show the original feature X_2 and the adjusted MVPA feature, $\tilde{X}_2$. Although the residuals $\tilde{X}_2$ are orthogonal to A by definition of least squares residuals, separability of the classes by $\tilde{X}_2$ alone is much less than marginal separability of the classes on the original feature X_2 . This implies that using adjusted features for marginal MVPA may have undesirable consequences on discrimination accuracy and the estimated disease pattern. The right two plots in Figure 4 show that the control-adjusted MVPA fails to remove the association between X_2 and A . Higher $\tilde{X}_2^c$ values correspond to lower values of A and lower values of $\tilde{X}_2^c$ correspond to higher values of A . Thus, Figure 4 suggests that regression-based methods for addressing confounding are ineffective, motivating our proposed method described next.

Figure 4: Comparison of adjusted and control-adjusted MVPA features. Left to right: original X_2 with estimated age effect; residuals, $\tilde{X}_2$; original X_2 with control-group estimated age effect; residuals, $\tilde{X}_2^c$. Dashed lines are the least squares fit of X_2 on A using the full and control-group data, respectively.

3.2 Inverse probability weighted classifiers

Having formally defined the problem of confounding in MVPA, we now propose a general solution based on inverse probability weighting (IPW) [45–48]. We have already shown that weighting observations by the inverse probability of Y given $\mathbf{A}$ relates the observed data to the counterfactual distributions $F_{X(-1)}$ and $F_{X(1)}$. The idea of weighting observations for classifier training is not new and in practice, applying IPW in this way is similar to weighting approaches that address dataset shift, a well-established concept in the machine learning literature [see, for example: 49–51].

The inverse probability weights are often unknown and must be estimated from the data. One way to estimate the weights is by positing a model and obtaining fitted values for the probability that $Y=1$ given confounders $\mathbf{A}$, also known as the propensity score [52, 53]. Logistic regression is commonly used to model the propensity score, however, more flexible approaches using machine learning have also received attention [54]. Using logistic regression, the model would be specified as

$$\text{logit}[\text{pr}(Y=1|\mathbf{A})] = \gamma_0 + \mathbf{A}^\top \gamma_1.$$

Then, the estimated inverse probability weights would follow as

$$\hat{w}_i^{-1} = [\mathbb{1}_{Y_i=1} \expit(\hat{\gamma}_0 + \mathbf{A}_i^\top \hat{\gamma}_1) + \mathbb{1}_{Y_i=0} \{1 - \expit(\hat{\gamma}_0 + \mathbf{A}_i^\top \hat{\gamma}_1)\}]^{-1},$$

where $\expit(x)$ is the inverse of the logit function, $\expit(x) = e^x / (1 + e^x)$, and $\mathbb{1}_z$ is the indicator function that takes value 1 if condition z is true and 0 otherwise.

IPW can be naturally incorporated into some classification models such as logistic regression. Subject-level weighting can be accomplished in the soft-margin linear SVM framework defined in expression (2) by weighting the slack variables. Suppose the true weights w_i are known. To demonstrate how IPW can be incorporated in the soft-margin linear SVM, we first consider approximate weights, T_i , defined as subject i 's inverse probability weight rounded to the nearest integer. For example, suppose subject i 's inverse weight is $1/w_i = 3.2$; then, $T_i = 3$. Next, consider creating an approximately balanced pseudo-population which consists of T_i copies of each original subject's data, $i=1, \dots, n$. This pseudo-population has $n^* = \sum_{i=1}^n T_i$ observations. The soft-margin SVM in the pseudo-population is then

$$\min_{\mathbf{v}, b, \xi^*} \frac{1}{2} \|\mathbf{v}\|^2 + C \sum_{j=1}^{n^*} \xi_j^*$$

such that:

$$y_j^*(\mathbf{v}^\top \mathbf{x}_j^* + b) \geq 1 - \xi_j^* \quad \forall j=1, \dots, n^*,$$

$$\xi_j^* \geq 0 \quad \forall j=1, \dots, n^*. \quad (5)$$

However, in the approximately balanced pseudo-population, some of the $(y_j^*, \mathbf{x}_j^*)$ pairs are identical copies which implies some of the constraints are redundant. For example, if $(y_1^*, \mathbf{x}_1^*)$ and $(y_2^*, \mathbf{x}_2^*)$ are identical copies that correspond to $(y_1, \mathbf{x}_1)$ in the original sample, then it can be seen that $\xi_1^* = \xi_2^*$ must hold in (5). Let $\xi_1 = \xi_1^* = \xi_2^*$. Then, the constraints

$$y_1^*(\mathbf{v}^\top \mathbf{x}_1^* + b) \geq 1 - \xi_1^*,$$

$$y_2^*(\mathbf{v}^\top \mathbf{x}_2^* + b) \geq 1 - \xi_2^*,$$

$$\xi_1^* \geq 0,$$

$$\xi_2^* \geq 0,$$

in (5) are equivalent to

$$y_1(\mathbf{v}^T \mathbf{x}_1 + b) \geq 1 - \xi_1,$$

$$\xi_1 \geq 0.$$

In fact, assuming all observations in the original n samples are unique, there are n unique constraints of the form $y_i(\mathbf{v}^T \mathbf{x}_i + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$, corresponding to the original $i = 1, \dots, n$ samples. In addition, it is straightforward to show that $\sum_{j=1}^{n^*} \xi_j^* = \sum_{i=1}^n T_i \xi_i$. Thus, (5) is equivalent to the original data soft-margin linear SVM with weighted slack variables in the objective function:

$$\min_{\mathbf{v}, b, \xi} \frac{1}{2} \|\mathbf{v}\|^2 + C \sum_{j=1}^n T_i \xi_j$$

such that :

$$y_i(\mathbf{v}^T \mathbf{x}_i + b) \geq 1 - \xi_i \quad \forall i = 1, \dots, n,$$

$$\xi_i \geq 0 \quad \forall i = 1, \dots, n. \quad (6)$$

The previous argument suggests one could use the true weights w_i , rather than the truncated weights, T_i . To our knowledge, an implementation of the SVM in R [55] that enables weighting the slack variables at the subject level does not exist. Subject-level weighting is available in the popular library libSVM [56]. Practitioners familiar with C++, MATLAB, or Python can implement the weighted SVM directly or by calling one of these languages from R using tools such as the “Rcpp” or “rPython” packages (rcpp.org, rpython.r-forge.r-project.org). We are currently working on an R implementation of the inverse probability weighted SVM (IPW-SVM) that uses the true weights, w_i . Development code is available at www.github.com/kalinn/weightedSVM, and a full example is given at www.github.com/kalinn/IPW-SVM.

The IPW-SVM algorithm only works when the data are not linearly separable. Otherwise, there are no slack variables in the optimization problem to weight. To provide intuition, suppose we are trying to separate two points in two-dimensional space. The optimization problem is then the hard-margin linear SVM formulation:

$$\min_{\mathbf{v}, b} \frac{1}{2} \|\mathbf{v}\|^2$$

such that :

$$y_1(\mathbf{v}^T \mathbf{x}_1 + b) = 1,$$

$$y_2(\mathbf{v}^T \mathbf{x}_2 + b) = 1.$$

Adding copies of the data only adds redundant constraints that do not affect the optimization. This is a major issue in neuroimaging because the data often have more features than observations and are thus almost always linearly separable. When $p \geq n$, one idea would be to preprocess the data using a variable selection or other dimension reduction technique that accounts for possible confounding in the data. Then the IPW-SVM could be implemented on the reduced feature space. We are currently exploring alternatives to address confounding when $p \geq n$ that retain the original interpretability of the features.

4 Simulation study

In this section we evaluate the finite sample performance of the IPW-SVM relative to the regression methods discussed in Section 3.1. We simulate training data from the following generative model with $p = 100$:

$$A \sim \text{Unif}(0, 1), \quad X_j(Y) = \begin{cases} 1 - Y + A\epsilon_{1,j} + \epsilon_{2,j} & j = 1, 2 \\ 5 - 11.25A - .75YA + A\epsilon_{1,j} + \epsilon_{2,j} & j = 3, \dots, p \end{cases}$$

$$\epsilon_1 \sim \text{Normal}(0_{p \times 1}, \Sigma_1), \quad \epsilon_2 \sim \text{Normal}(0_{p \times 1}, \Sigma_2), \quad (7)$$

where Σ_1 is a $p \times p$ identity matrix, and Σ_2 is a $p \times p$ matrix with 1s on the diagonal and 0.2s on all off-diagonal elements.

For each of $M=1,000$ iterations, we generate a sample of size $N=300$ of the trajectory $(A, \mathbf{X}(-1)^\top, \mathbf{X}(1)^\top)^\top$ from model (7). We train a SVM using the features $\mathbf{X}_i(-1)$ and $\mathbf{X}_i(1)$, $i=1, \dots, N$, and take the resulting SVM weights to be the “true” weight vector. Next, we simulate confounding by setting $\mathbf{X} = \mathbf{X}(Y^{obs})$, where $Y^{obs} = 2Y^* - 1$, $Y^* = \text{Bernoulli}(\tilde{A}^2)$ and $\tilde{A} = 0.5\mathbb{1}_{A < 0.5} + A\mathbb{1}_{A \geq 0.5}$. Thus, subjects with larger values of A are more likely to be observed with $Y=1$. Finally, we create a test set with no confounding by A by generating a separate sample of $N=300$ trajectories from model (7) and setting $\mathbf{X} = \mathbf{X}(Y^{test})$, where $Y^{test} = 2Y^* - 1$, $Y^* = \text{Bernoulli}(.5)$.

We compare the performance of the IPW-SVM (IPW) to an unadjusted SVM (Unadjusted), a SVM after “regressing out” A from each feature separately using a linear model (Adjusted), and a SVM after “regressing out” A from each feature separately using a linear model fit in the group observed with $Y^{obs} = -1$ (CN-Adjusted). Section 3.1 gives details about the regression-based adjustment methods. The full estimated inverse probability weights (i.e., non-truncated weights) are used for training the IPW-SVM. We use L_2 distance between the true and estimated weight vectors as one criterion for comparison. Figure 5 displays boxplots of the test accuracy and L_2 distance from the true weights for $M = 1,000$ iterations. Results are presented relative to the unadjusted SVM for each iteration. That is, the results in the left plot of Figure 5 are obtained by dividing the L_2 distance of the IPW-SVM weights from the true weights by the L_2 distance of the unadjusted SVM weights from the true weights at each of the M iterations, and similarly for the other SVM methods. Similarly, the right plot in Figure 5 is obtained by dividing the test accuracy of the IPW-SVM by the test accuracy of the unadjusted SVM at each iteration, and similarly for the other SVM methods. Thus, improved performance over the unadjusted SVM is indicated by values below one in the left plot and values above one in the right plot of Figure 5. Overall, the IPW-SVM performs the best with respect to the relative distribution of L_2 distance and attains the highest median test accuracy. The left plot of Figure 5 has been zoomed-in to better compare the interquartile ranges. The IPW-SVM resulted in more outliers than the regression-based adjustment methods. The IPW-SVM seems sensitive to very large weights which occurred by chance in several iterations. Using stabilized weights provided modest improvement (results not presented here) in this simulation study.

Figure 5: Left: L_2 distance between the true and estimated weight vectors from each SVM implementation that addresses confounding, scaled by the L_2 distance between the true and estimated weight vectors from the unadjusted SVM at each iteration of the simulation. Right: Distribution of test accuracy between the estimated SVM decision rules on an unconfounded test set, relative to the test accuracy of the unadjusted SVM on the same unconfounded test set.

5 Application

The Alzheimer’s Disease Neuroimaging Initiative (ADNI) (<http://www.adni.loni.usc.edu>) is a \$60 million study funded by public and private resources including the National Institute on Aging, the National

Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies, and non-profit organizations. The goals of the ADNI are to better understand progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD) as well as to determine effective biomarkers for disease diagnosis, monitoring, and treatment development. MCI is characterized by cognitive decline that does not generally interfere with normal daily function and is distinct from Alzheimer's disease [57]. However, individuals with MCI are considered to be at risk for progression to Alzheimer's disease. Thus, studying the development of MCI and factors associated with progression to Alzheimer's disease is of critical scientific importance. In this analysis, we study the effects of confounding on the identification of multivariate patterns of atrophy in the brain that are associated with MCI.

We apply the IPW-SVM to structural MRIs from the ADNI database. Before performing group-level analyses, each subject's MRI is passed through a series of preprocessing steps that facilitate between-subject comparability. We implemented a multi-atlas segmentation pipeline [58] to estimate the volumes of $p = 137$ 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, and these volumes are used as features for SVM training. The data we use here consist of $n = 551$ subjects, where $n_{-1} = 224$ are healthy controls and $n_1 = 327$ are patients diagnosed with MCI between the ages of 69 and 90. Neurodegenerative diseases are associated with atrophy in the brain, and thus the MCI group has smaller volumes on average in particular ROIs compared to the control group.

Although the ADNI study was approximately matched on age and gender, a logistic regression of disease group on age in our sample returns an estimated odds ratio of 1.06 with 95% confidence interval (1.02, 1.09), indicating that age is a possible confounder of the disease-image relationship. In this analysis, our focus is on identifying multivariate patterns in the brain that represent differences between the MCI and control groups, rather than predictive performance of the MVPA classifier. We perform four separate multivariate pattern analyses: (i) an unadjusted SVM, (ii) the adjusted SVM described in Section 3.1, (iii) the control-adjusted SVM described in Section 3.1, and the IPW-SVM described in Section 3.2 with estimated weights. We compare the results from these four methods to the estimated weight pattern from a SVM trained on a one-to-one age-matched subsample of the data. Figure 6 displays the top 10 weighted SVM features from the one-to-one age-matched data. Blue (red) regions correspond to negative (positive) weights. From top to bottom, Figure 7 displays the top 10 weighted SVM features from the IPW-SVM, unadjusted SVM, control-adjusted SVM, and adjusted SVM.

Figure 6: Top 10 weighted SVM features from the one-to-one age-matched data. Blue (red) regions correspond to negative (positive) weights.

In general, all four methods perform similarly and return patterns that closely resemble the pattern learned from the matched data. Table 1 gives the L_2 distance between the estimated patterns and the matched-data SVM weight pattern. The IPW-SVM results in the least-biased weight pattern, and the regression-based adjustments demonstrate improvement over the unadjusted SVM.

It should be noted that although there is a significant disease-age relationship in the observed data, it is unlikely representative of the true disease-age relationship in the population because the

Figure 7: Top 10 weighted SVM features from the (top to bottom) IPW-SVM, unadjusted SVM, control-adjusted SVM, and adjusted SVM. Blue (red) regions correspond to negative (positive) weights.

Table 1: L_2 distance between the estimated patterns and the matched-data SVM weight pattern.

Method	Distance
IPW-SVM	0.52
Unadjusted SVM	0.76
Control-Adjusted SVM	0.58
Adjusted SVM	0.56

MCI cases are over-sampled. Thus, MVPA classifiers trained to study disease patterns in the brain may demonstrate suboptimal performance when classifying new subjects in the population. Dataset shift methods, or models that integrate imaging biomarkers with knowledge of the true disease-age relationship in the target population, may be applied to improve any MVPA imaging biomarkers derived from the ADNI data.

6 Discussion

We have proposed a framework for addressing confounding in MVPA that weights individual subjects by the conditional probability of observed class given confounders, i.e., inverse probability weighting (IPW). When the goal of MVPA is to estimate complex disease patterns in the brain, using IPW to address confounding is more principled than the current practice of “regressing out” confounder effects separately at each voxel without regard to the correlation structure of the data. When machine learning predictive models such as the SVM are used to perform MVPA, the IPW approach can recover underlying patterns in the brain associated with disease in the presence of measured confounding.

We believe there are several advantages to addressing confounding in MVPA using IPW. First, as demonstrated by simulation results, IPW better estimates the target parameter of interest, which is the disease pattern that would be present under no confounding. In cases where a matched study is too expensive or otherwise infeasible, IPW methods will enable researchers to perform MVPA and obtain correct, reproducible results. Finally, IPW is simple and intuitive, and the general idea is well-established in the causal inference and statistics communities. Thus, future research aiming to perform inference on the estimated disease patterns can rely on existing theory. We are currently working on extending existing inference methods for MVPA [14, 59] to account for confounding.

Further exploring the effects of confounding on high-dimensional classification models is imperative for neuroimaging research and may greatly impact current practice in the field. An interesting avenue for future research would be to develop dimension reduction techniques that could be applied before or concurrently with MVPA that account for possible confounding in the data. Developing sensitivity analysis methods for assessing the role of confounding in MVPA also merits attention in future work.

Although we have focused on the use of SVMs for binary classification problems, the idea of subject-level weighting to address confounding applies more generally to machine learning techniques for a variety of classification problems. In practice, incorporating subject-level weights into black box machine learning methods may not always be straightforward, and implementation of IPW might require specific tailoring to each problem. For example, generalized versions of the propensity score exist for exposures with more than two groups and continuous exposures [60, 61]. Intuitively, it seems that applying generalized propensity score methods to multiclass classification problems or support vector regression for a continuous exposure is a natural extension of the methods proposed in this work. We believe these extensions are non-trivial and warrant focused attention in future research.

Funding: 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.

References

1. Frackowiak R, Friston K, Frith C, Dolan R, Mazziotta J, (eds.). Human brain function. San Diego, CA: Academic Press USA, 1997.
2. Friston KJ, Frith C, Liddle P, Frackowiak R. Comparing functional PET images: the assessment of significant change. *J Cereb Blood Flow Metab* 1991;11:690–9.
3. Friston KJ, Holmes AP, Worsley KJ, Poline J-P, Frith CD, Frackowiak RS. Statistical parametric maps in functional imaging: a general linear approach. *Hum Brain Mapp* 1994;2:189–210.
4. Ashburner J, Friston KJ. Voxel-based morphometry – the methods. *Neuroimage* 2000;11:805–21.
5. Davatzikos C, Genc A, Xu D, Resnick SM. Voxel-based morphometry using the RAVENS maps: methods and validation using simulated longitudinal atrophy. *Neuroimage* 2001;14:1361 – 1369.
6. Craddock RC, Holtzheimer PE, Hu XP, Mayberg HS. Disease state prediction from resting state functional connectivity. *Magn Reson Med* 2009;62:1619–28.

7. Cuingnet R, Rosso C, Chupin M, Lehicry S, Dormont D, Benali H, et al. Spatial regularization of SVM for the detection of diffusion alterations associated with stroke outcome. *Med Image Anal* 2011;15:729–37, Special Issue on the 2010 Conference on Medical Image Computing and Computer-Assisted Intervention.
8. Davatzikos C, Bhatt P, Shaw LM, Batmanghelich KN, Trojanowski JQ. Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiol Aging* 2011;32:2322.e19–2322.e27.
9. Davatzikos C, Resnick S, Wu X, Parmpi P, Clark C. Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. *NeuroImage* 2008;41:1220 – 1227.
10. Davatzikos C, Ruparel K, Fan Y, Shen D, Acharyya M, Loughhead J, et al. Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. *Neuroimage* 2005;28:663–8.
11. Davatzikos C, Xu F, An Y, Fan Y, Resnick SM. Longitudinal progression of Alzheimer's-like patterns of atrophy in normal older adults: the spare-ad index. *Brain* 2009;132:2026–35.
12. De Martino F, Valente G, Staeren N, Ashburner J, Goebel R, Formisano E. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *Neuroimage* 2008;43:44–58.
13. Fan Y, Shen D, Gur RC, Gur RE, Davatzikos C. Compare: classification of morphological patterns using adaptive regional elements. *IEEE Trans Med Imaging* 2007;26:93–105.
14. Gaonkar B, Davatzikos C. Analytic estimation of statistical significance maps for support vector machine based multi-variate image analysis and classification. *NeuroImage* 2013;78:270–283.
15. Klöppel S, Stonnington CM, Chu C, Draganski B, Scahill RI, Rohrer JD, et al. Automatic classification of MR scans in Alzheimer's disease. *Brain* 2008;131:681–9.
16. Koutsouleris N, Meisenzahl EM, Davatzikos C, Bottlender R, Frodl T, Scheuerecker J, et al. Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. *Arch Gen Psychiatry* 2009;66:700–12.
17. Langs G, Menze BH, Lashkari D, Golland P. Detecting stable distributed patterns of brain activation using gini contrast. *NeuroImage* 2011;56:497–507.
18. Mingoia G, Wagner G, Langbein K, Maitra R, Smesny S, Dietzek M, et al. Default mode network activity in schizophrenia studied at resting state using probabilistic ICA. *Schizophrenia research* 2012;138:143–9.
19. Moura'o-Miranda J, Bokde AL, Born C, Hampel H, Stetter M. Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data. *NeuroImage* 2005;28:980–95.
20. Pereira F. Beyond brain blobs: machine learning classifiers as instruments for analyzing functional magnetic resonance imaging data. ProQuest, 2007.
21. Reiss PT, Ogden RT. Functional generalized linear models with images as predictors. *Biometrics* 2010;66:61–9.
22. Richiardi J, Eryilmaz H, Schwartz S, Vuilleumier P, Van De Ville D. Decoding brain states from fMRI connectivity graphs. *Neuroimage* 2011;56:616–26.
23. Sabuncu MR, Van Leemput K. The relevance voxel machine (rvxm): a bayesian method for image-based prediction. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI* 2011. Springer, 2011: 99–106.
24. Vemuri P, Gunter JL, Senjem ML, Whitwell JL, Kantarci K, Knopman DS, et al. Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. *Neuroimage* 2008;39:1186–97.
25. Venkataraman A, Rathi Y, Kubicki M, Westin C-F, Golland P. Joint modeling of anatomical and functional connectivity for population studies. *IEEE Trans Med Imaging* 2012;31:164–82.
26. Wang Z, Childress AR, Wang J, Detre JA. Support vector machine learning-based fMRI data group analysis. *NeuroImage* 2007;36:1139–51.
27. Xu L, Groth KM, Pearlson G, Schretlen DJ, Calhoun VD. Source based morphometry: the use of independent component analysis to identify gray matter differences with application to schizophrenia. *Hum Brain Mapping* 2009;30:711–24.
28. Raz N, Rodrigue KM. Differential aging of the brain: patterns, cognitive correlates and modify. *Neurosci Biobehav Rev* 2006;30:730–48.
29. Cortes C, Vapnik V. Support-vector networks. *Machine Learning* 1995;20:273– 297.
30. Vapnik V. The nature of statistical learning theory. Springer: Springer Science & Business Media, 2000.
31. Schölkopf B, Tsuda K, Vert J-P. Kernel methods in computational biology. Cambridge, MA: MIT press, 2004.
32. Sun D, van Erp TG, Thompson PM, Bearden CE, Daley M, Kushan L, et al. Elucidating a magnetic resonance imaging-based neuroanatomic biomarker for psychosis: classification analysis using probabilistic brain atlas and machine learning algorithms. *Biol Psychiatry* 2009;66:1055–60.
33. Reiss PT, Ogden RT. Functional principal component regression and functional partial least squares. *J Am Stat Assoc* 2007;102:984–96.
34. Zipunnikov V, Caffo B, Yousem DM, Davatzikos C, Schwartz BS, Crainiceanu C. Functional principal component model for high-dimensional brain imaging. *NeuroImage* 2011;58:772–84.
35. Friedman J, Hastie T, Tibshirani R. *The elements of statistical learning*. Vol. 1. Springer, Berlin: Springer series in statistics, 2001.
36. Orru G, Pettersson-Yeo W, Marquand AF, Sartori G, Mechelli A. Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neurosci Biobehav Rev* 2012;36:1140–52.

37. Bendfeldt K, Klöppel S, Nichols TE, Smieskova R, Kuster P, Traud S, et al. Multivariate pattern classification of gray matter pathology in multiple sclerosis. *Neuroimage* 2012;60:400–8.
38. Liu F, Guo W, Yu D, Gao Q, Gao K, Xue Z, et al. Classification of different therapeutic responses of major depressive disorder with multivariate pattern analysis method based on structural MR scans. *PLoS One* 2012;7:e40968.
39. Gong Q, Wu Q, Scarpazza C, Lui S, Jia Z, Marquand A, et al. Prognostic prediction of therapeutic response in depression using high-fi mr imaging. *Neuroimage* 2011;55:1497–503.
40. Costafreda SG, Chu C, Ashburner J, Fu CH. Prognostic and diagnostic potential of the structural neuroanatomy of depression. *PLoS One* 2009;4:e6353.
41. Holland PW. Statistics and causal inference. *J Am Stat Assoc* 1986;81:945–60.
42. Rubin DB. Estimating causal effects of treatments in randomized and non-randomized studies. *J Educ Psychol* 1974;66:688.
43. Li L, Rakitsch B, Borgwardt K. ccsvm: correcting support vector machines for confounding factors in biological data classification. *Bioinformatics* 2011;27:i342– i348.
44. Dukart J, Schroeter ML, Mueller K. Age correction in dementia – matching to a healthy brain. *PLoS ONE* 2011;6:e22193.
45. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol* 2008;168:656–64.
46. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Commun Health* 2006;60:578–86.
47. Robins JM. Marginal structural models. In: *Proceedings of the Section on Bayesian Statistical Science*. Alexandria, VA: American Statistical Association, 1998: 1–10.
48. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000;11:550–60.
49. Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla NV, Herrera F. A unifying view on dataset shift in classification. *Pattern Recognit* 2012;45:521– 530.
50. Quionero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND. *Dataset shift in machine learning*. Cambridge, MA: The MIT Press, 2009.
51. Zadrozny B. Learning and evaluating classifiers under sample selection bias. In: *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004:114.
52. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research* 2011;46:399–424.
53. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;70:41–55.
54. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat Med* 2010;29:337–46.
55. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2014.
56. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. *ACM Trans Intell Syst Technol* 2011;2:27:1–27:27.
57. Gauthier S, Reisberg B, Zaudig M, Petersen RC, Ritchie K, Broich K, et al. Mild cognitive impairment. *The Lancet* 2006;367:1262–70.
58. Doshi J, Erus G, Ou Y, Gaonkar B, Davatzikos C. Multi-atlas skull-stripping. *Acad Radiol* 2013;20:1566–76.
59. Gaonkar B, Shinohara RT, Davatzikos C, Initiative ADN, et al. Interpreting support vector machine models for multivariate group wise analysis in neuroimaging. *Med Image Anal* 2015;24(1):190–204.
60. Hirano K, Imbens GW. The propensity score with continuous treatments. *Appl Bayesian Model Causal Inference Incomplete-Data Perspect* 2004;226164:73–84.
61. Imbens GW. The role of the propensity score in estimating dose-response functions. *Biometrika* 2000;87:706–10.
62. Weichwald S, Meyer T, Özdenizci O, Schölkopf B, Ball T, Grosse-Wentrup M. Causal interpretation rules for encoding and decoding models in neuroimaging. *NeuroImage* 2015;110:4859.