University of Pennsylvania

UPenn Biostatistics Working Papers

Year 2015 *Paper* 41

Addressing Confounding in Predictive Models with an Application to Neuroimaging

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

Addressing Confounding in Predictive Models with an Application to Neuroimaging

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

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.

Addressing Confounding in Predictive Models with an Application to Neuroimaging

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

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.

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) [20–22] and voxel-based morphometry (VBM) [1, 12]. 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 [9–11, 13–16, 19, 23, 35–37, 41, 43, 45, 50, 51, 56, 63–65, 67].

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

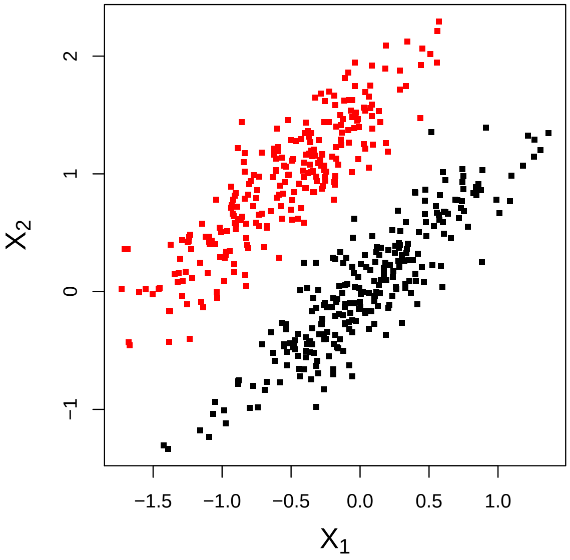

Figure 1: *Marginally, X_1 and X_2 discriminate poorly between the groups, but perfect separability is attained when X_1 and X_2 are considered jointly.*

biomarkers for disease diagnosis, the prediction of disease progression, or prediction of treatment response. Although the MVPA literature often 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. [66]. 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 $(Y_i, \boldsymbol{X}_i^\mathsf{T}, \boldsymbol{A}_i^\mathsf{T})^\mathsf{T}$, $i = 1, \ldots, n$, denote n independent and identically distributed observations of the random vector $(Y, \boldsymbol{X}^T, \boldsymbol{A}^\mathsf{T})^\mathsf{T}$, where $Y \in \{-1, 1\}$ denotes the group label, e.g., control versus disease, $\boldsymbol{X} \in \mathbb{R}^p$ denotes a vectorized image with p voxels, and $\boldsymbol{A} \in \mathbb{R}^r$ denotes a vector of non-image variables such as age and gender. Suppose Y and $\boldsymbol{A}$ both affect $\boldsymbol{X}$. 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 [48]. 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) [7, 62]. This choice is partly motivated by the fact that SVMs are known to work well for high dimension, low sample size data [57]. 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 [61] as well as functional principal components and functional partial least squares [49, 69]. Henceforth, we focus on MVPA using the SVM.

The hard-margin linear SVM solves the contrained 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 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

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

where $C \in \mathbb{R}$ is a tuning parameter that penalizes misclassification, and $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_n)^\mathsf{T}$. For details about solving optimization problems (1) and (2) we refer the reader to Hastie et al. [28].

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 [44]. Thus, MVPA is often applied using the hard-margin linear SVM in (1). For example, this is the approach implemented by: Bendfeldt et al. [3] to classify subgroups of multiple sclerosis patients; Cuingnet et al. [10] and Davatzikos et al. [11] in Alzheimer's disease applications; and Liu et al. [40], Gong et al. [26], and Costafreda et al. [8] for various classification tasks involving patients with depression. This is only a small subset of the relavant 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 [31, 55]. In addition, confounding of the disease–image relationship presents challenges. Figure 2 depicts confounding of the Y–$\boldsymbol{X}$ 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 [39]. 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.

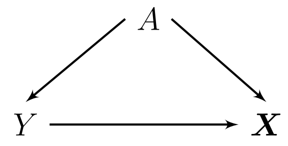

Figure 2: *The relationship between Y (disease) and $\boldsymbol{X}$ (image) is confounded by A (e.g., age), which affects both Y and $\boldsymbol{X}$.*

Let $\boldsymbol{X}_i(y)$ denote the image that would have been observed had subject i been observed with group status $Y_i = y$, possibly contrary to fact. Let $F_{\boldsymbol{X}(-1)}$ and $F_{\boldsymbol{X}(1)}$ denote the distributions of the counterfactual images $\boldsymbol{X}(-1)$ and $\boldsymbol{X}(1)$, respectively. Assume there exists a unique hyperplane in $\mathbb{R}^p$ that maximally separates the counterfactural 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. 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 $\theta^* = S(F_{\boldsymbol{X}(-1)}, F_{\boldsymbol{X}(1)})$.

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

$$\begin{array}{lc} (i) & \boldsymbol{X}_i = \dfrac{\boldsymbol{X}_i(1) + \boldsymbol{X}_i(-1)}{2} + Y_i \dfrac{\boldsymbol{X}_i(1) - \boldsymbol{X}_i(-1)}{2}, \\ (ii) & \{\boldsymbol{X}_i(1), \boldsymbol{X}_i(-1)\} \perp\!\!\!\perp Y_i \mid \boldsymbol{A}_i, \end{array}$$

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

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

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

$$\begin{aligned} E\{\mathrm{pr}(X \leq x \mid Y = y, \boldsymbol{A})\} &= E\left\{\mathrm{pr}(X \leq x \mid Y = y, \boldsymbol{A}) \frac{\mathrm{pr}(Y = y \mid \boldsymbol{A})\mathrm{pr}(Y = y)}{\mathrm{pr}(Y = y \mid \boldsymbol{A})\mathrm{pr}(Y = y)}\right\} \\ &= \frac{1}{\mathrm{pr}(Y = y)} \int \mathrm{pr}(X \leq x, Y = y \mid \boldsymbol{A}) \frac{\mathrm{pr}(Y = y)}{\mathrm{pr}(Y = y \mid \boldsymbol{A})} dP_{\boldsymbol{A}} \\ &= \frac{1}{\mathrm{pr}(Y = y)} \int \mathrm{pr}(X \leq x, Y = y \mid \boldsymbol{A}) dP^*_{\boldsymbol{A}} \\ &= F^*_{\boldsymbol{X}|Y=y}, \end{aligned}$$

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

To illustrate the effects of confounding on MVPA, consider a toy example with a single confounder A. Let $\boldsymbol{X}$ consist of two features, $\boldsymbol{X} = (X_1, X_2)^{\mathsf{T}}$, and define the corresponding potential outcomes, $\boldsymbol{X}(Y) = \{X_1(Y), X_2(Y)\}^{\mathsf{T}}$. 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

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

Note that model (3) has the property that Y and A are independent, so that A is not a confounder of the Y–$\boldsymbol{X}$ 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–$\boldsymbol{X}$ 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–$\boldsymbol{X}$ 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 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.

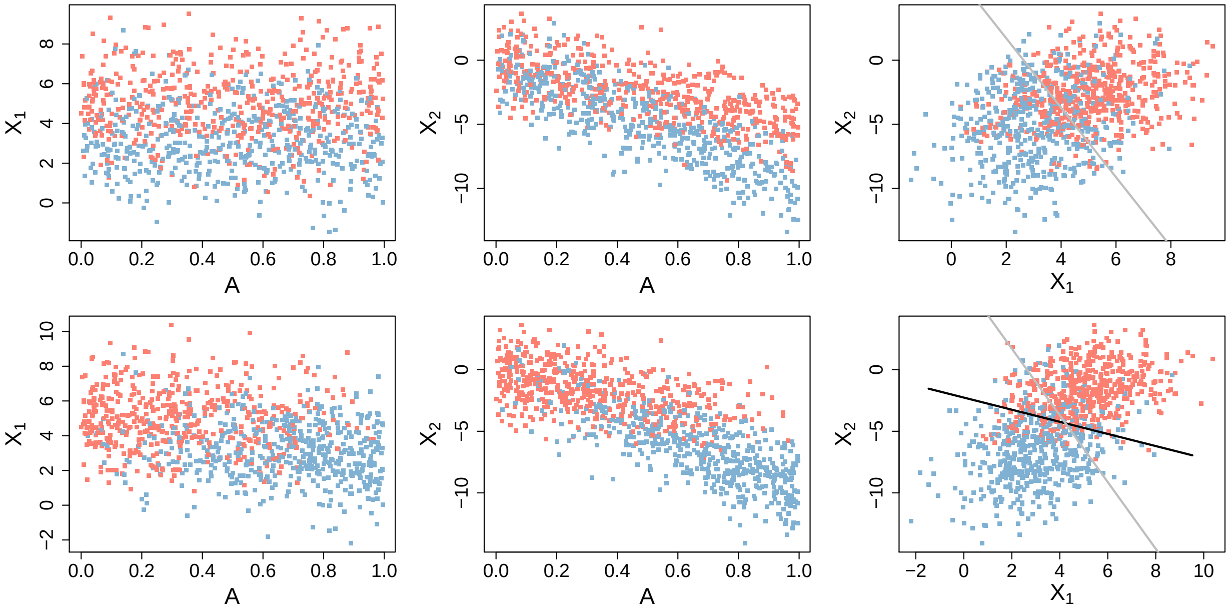

Figure 3: **Top row:** *unconfounded data generated from model (3).* **Bottom row:** *biased-sample data with the Y–$\boldsymbol{X}$ relationship confounded by A. The target parameter is the linear SVM decision rule learned from the data in the top right plot, shown in gray. The black line is the linear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

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 contols [18]. In an age-matched study, age is not a confounder, and adjusting for its relationship with $\boldsymbol{X}$ is unnecessary. To address confounding, one approach proposed in the neuroimaging literature is to "regress-out" the effects of confounders from the image $\boldsymbol{X}$. 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 [18, 22].

The resulting "residual image" is then used in MVPA. Formally, the following model is fit using least squares, separately for each $j = 1, \ldots, p$:

$$X_j = \beta_{0,j} + \boldsymbol{\beta}_{1,j}^{\mathsf{T}} \boldsymbol{A} + \epsilon_j, \tag{4}$$

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

$$\widetilde{X}_j = X_j - (\widehat{\beta}_{0,j} + \widehat{\boldsymbol{\beta}}_{1,j}^{\mathsf{T}} \boldsymbol{A}).$$

Combining all residuals gives the vector $\widetilde{\boldsymbol{X}} = (\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_p)$ 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 [18]. We refer to this approach as the *control-adjusted MVPA*. Let $\widehat{\beta}^c_{0,j}$ and $\widehat{\boldsymbol{\beta}}^c_{1,j}$ denote the least squares estimates of $\beta_{0,j}$ and $\boldsymbol{\beta}_{1,j}$ when model (4) is fit using only control-group data. The control-group adjusted features used in the MVPA classifier are then $\widetilde{\boldsymbol{X}}^c = (\widetilde{X}^c_1, \widetilde{X}^c_2, \ldots, \widetilde{X}^c_p)$, where $\widetilde{X}^c_j = X_j - (\widehat{\boldsymbol{\beta}}^{c\mathsf{T}}_{0,j} + \widehat{\boldsymbol{\beta}}^{c\mathsf{T}}_{1,j} \boldsymbol{A})$.

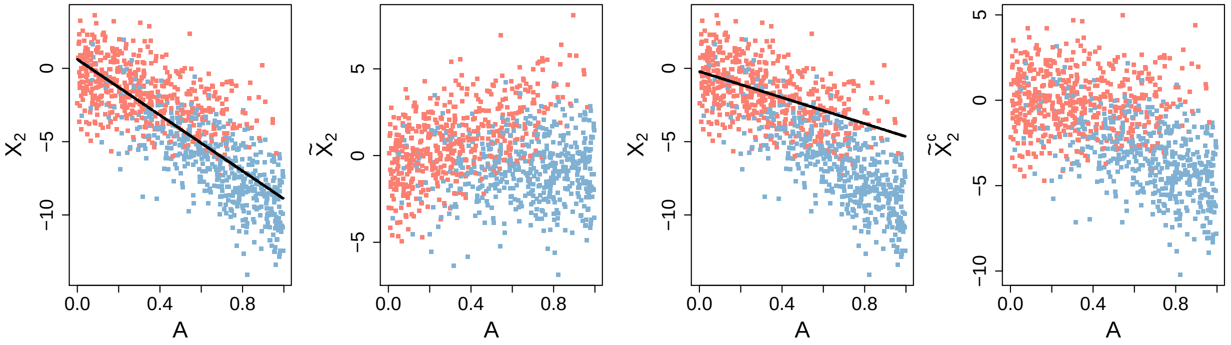

Figure 4: *Comparison of adjusted and control-adjusted MVPA features. Left to right: original X_2 with estimated age effect; residuals, $\widetilde{X}_2$; original X_2 with contol-group estimated age effect; residuals, $\widetilde{X}^c_2$. Lines are the least squares fit of X_2 on A using the full and control-group data, respectively.*

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, $\widetilde{X}_2$. Although the residuals $\widetilde{X}_2$ are orthogonal to A by definition of least squares residuals, separability of the classes by $\widetilde{X}_2$ alone is much less than marginal separabilty 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 contol-adjusted MVPA fails to remove the association between X_2 and A. Higher $\widetilde{X}^c_2$ values correspond to lower values of A and lower values of $\widetilde{X}_2^c$ 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.

3.1.1 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) [6, 29, 52, 53]. We have already shown that weighting observations by the inverse probability of Y given $\boldsymbol{A}$ relates the observed data to the counterfactual distributions $F_{\boldsymbol{X}(-1)}$ and $F_{\boldsymbol{X}(1)}$. 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: 42, 46, 68].

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 $\boldsymbol{A}$, also known as the propensity score [2, 54]. Logistic regression is commonly used to model the propensity score, however, more flexible approaches using machine learning have also received attention [38]. Using logistic regression, the model would be specified as

$$\text{logit}[\text{pr}(Y = 1 \mid \boldsymbol{A})] = \gamma_0 + \boldsymbol{A}^\mathsf{T}\gamma_1.$$

Then, the estimated inverse probability weights would follow as

$$\widehat{w}_i^{-1} = [\mathbb{1}_{Y_i=1}\text{expit}(\widehat{\gamma}_0 + \boldsymbol{A}_i^\mathsf{T}\widehat{\gamma}_1) + \mathbb{1}_{Y_i=0}\{1 - \text{expit}(\widehat{\gamma}_0 + \boldsymbol{A}_i^\mathsf{T}\widehat{\gamma}_1))\}]^{-1},$$

where $\text{expit}(x)$ is the inverse of the logit function, $\text{expit}(x) = e^x/(1 + e^x)$.

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 psuedo-population which consists of T_i copies of each original subject's data, $i = 1, \dots, n$. This psuedo-population has $n^* = \sum_{i=1}^n T_i$ observations. The soft-margin SVM in the psuedo-population is then

$$\begin{aligned} &\arg\min_{\boldsymbol{v},b,\boldsymbol{\xi}^*} \frac{1}{2}||\boldsymbol{v}||^2 + C\sum_{j=1}^{n^*}\xi_j^* \\ &\text{such that:} \\ &y_j^*(\boldsymbol{v}^\mathsf{T}\boldsymbol{x}_j^* + b) \geq 1 - \xi_j^* \;\; \forall j = 1, \dots, n^*, \\ &\xi_j^* \geq 0 \;\; \forall j = 1, \dots, n^*. \end{aligned} \tag{5}$$

However, in the approximately balanced psuedo-population, some of the $(y_j^*, \boldsymbol{x}_j^*)$ pairs are identical copies which implies some of the constraints are redundant. For example, if $(y_1^*, \boldsymbol{x}_1^*)$

and $(y_2^*, \boldsymbol{x}_2^*)$ are identical copies that correspond to $(y_1, \boldsymbol{x}_1)$ 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

$$\begin{aligned} &y_1^*(\boldsymbol{v}^\mathsf{T}\boldsymbol{x}_1^* + b) \geq 1 - \xi_1^*, \\ &y_2^*(\boldsymbol{v}^\mathsf{T}\boldsymbol{x}_2^* + b) \geq 1 - \xi_2^*, \\ &\xi_1^* \geq 0, \\ &\xi_2^* \geq 0, \end{aligned}$$

in (5) are equivalent to

$$\begin{aligned} &y_1(\boldsymbol{v}^\mathsf{T}\boldsymbol{x}_1 + b) \geq 1 - \xi_1, \\ &\xi_1 \geq 0. \end{aligned}$$

In fact, assuming all observations in the original n samples are unique, there are n unique constraints of the form $y_i(\boldsymbol{v}^\mathsf{T}\boldsymbol{x}_i + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$, corresponding to the original $i = 1, \ldots, n$ 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:

$$\begin{aligned} &\arg\min_{\boldsymbol{v},b,\boldsymbol{\xi}} \frac{1}{2}||\boldsymbol{v}||^2 + C\sum_{i=1}^{n} T_i \xi_i \\ &\text{such that:} \\ &y_i(\boldsymbol{v}^\mathsf{T}\boldsymbol{x}_i + b) \geq 1 - \xi_i \;\; \forall i = 1, \ldots, n, \\ &\xi_i \geq 0 \;\; \forall i = 1, \ldots, n. \end{aligned} \tag{6}$$

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 [47] that enables weighting the slack variables at the subject level does not exist. Subject-level weighting is available in the popular library libSVM [5]. 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). Development code is available at github.com/kalinn

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:

$$\begin{aligned} &\arg\min_{\boldsymbol{v},b} \frac{1}{2}||\boldsymbol{v}||^2 \\ &\text{such that:} \\ &y_1(\boldsymbol{v}^\mathsf{T}\boldsymbol{x}_1 + b) = 1, \\ &y_2(\boldsymbol{v}^\mathsf{T}\boldsymbol{x}_2 + b) = 1. \end{aligned}$$

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.

3.2 Remarks on Biomarker Development

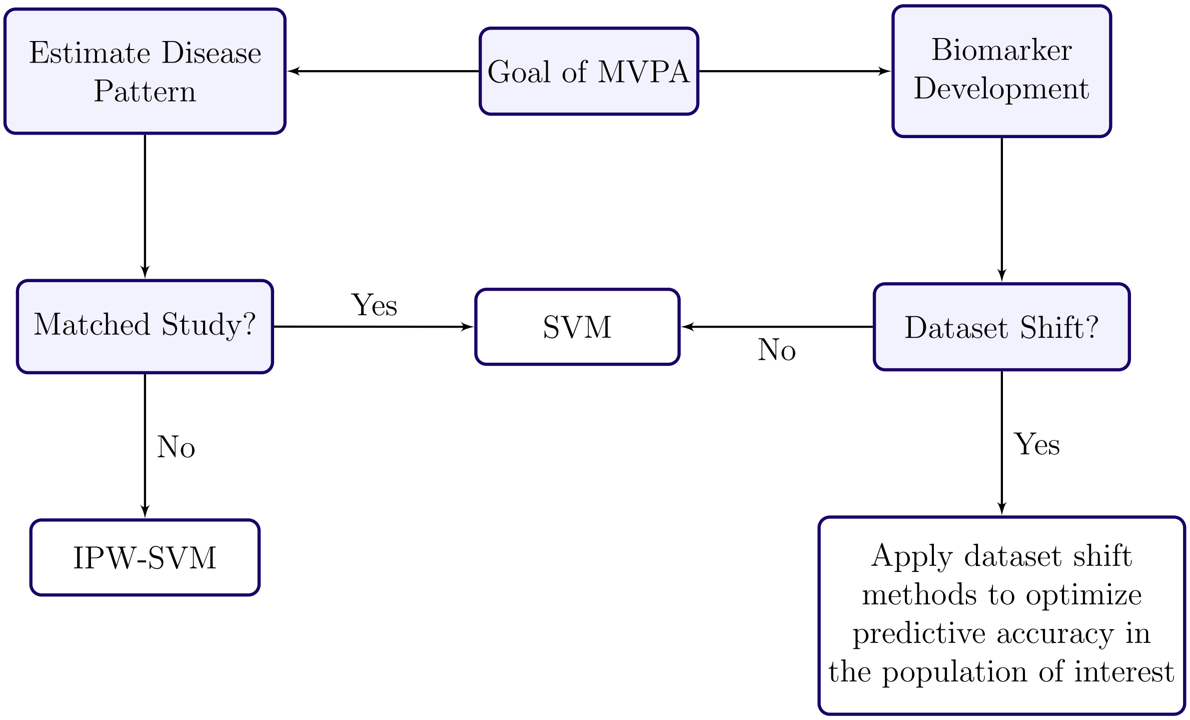

Figure 5: *Recommended anaylsis plan for MVPA in the presence of measured confounding or dataset shift.*

Non-imaging variables play a different, and sometimes advantageous, role when MVPA is used to develop biomarkers for disease diagnosis, progression, or treatment response. It is unlikely that the true underlying distribution of non-imaging variables such as age is balanced with respect to disease. That is, it is unlikely a matched study is representative of the population to which a derived biomarker will be applied. As a result, matched studies or IPW methods that create balance with respect to the disease and non-imaging variables may not result in the optimal classifier or biomarker. This observation has been made previously in different contexts, including the in the statistical literature [34] and the machine learning literature [42, 46].

In machine learning, *dataset shift* is the phenomenon where the joint distribution of training data differs from the data distribution where the classifier will be applied [42, 46]. *Covariate shift* is a special case of dataset shift which corresponds to a shift in the feature distribution used to obtain predictions from the classification model. Solutions usually involve some version of observation weighting or moment matching to make the training and test feature distributions more comparable [4, 27, 32, 58–60]. Applying dataset shift methods to neuroimaging data has the potential to improve biomarker effectiveness and generalizability. For example, suppose a biomarker is developed using imaging and demographic data from a matched study. That is, patients have been selected so that there are equal numbers of cases and controls for all values of the demographic variables. However, suppose it is known that in the general population the disease is more prevalent in older patients. Then, the matched study data and the population to which the biomarker will be applied come from different joint distributions. Dataset shift methods enable prior knowledge of the population distribution to be leveraged to attain optimal predictive performance of the biomarker.

To summarize the main points in this section, Figure 5 provides a decision tree with recommended analysis plans for given data structures and scientific aims. We believe tools such as Figure 5 may be useful to help initiate or guide discussion with collaborators about the design and analysis of future neuroimaging studies.

4 Simulation Study

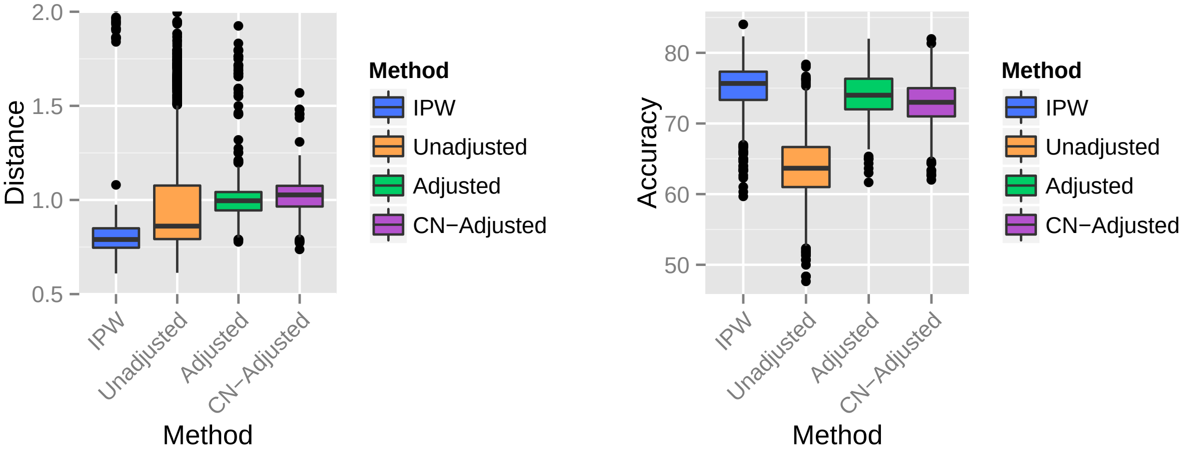

Figure 6: **Left:** *Distribution of L_2 distance between the true and estimated weight vectors resulting from the IPW-SVM, Unadjusted SVM, Adjusted SVM, and Control-Adjusted SVM.* **Right:** *Percent test accuracy of the estimated SVM decision rules on an unconfounded test set.*

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 $p = 100$:

$$A \sim \text{Unif}(0,1), \qquad X_j(Y) = \begin{cases} 1 - Y + A\epsilon_{1,j} + \epsilon_{2,j} & j = 1,2 \\ 5 - 11.25 * A - .75YA + A\epsilon_{1,j} + \epsilon_{2,j} & j = 3,\ldots,p \end{cases}$$
$$\epsilon_1 \sim \text{Normal}(0_{p\times 1}, \Sigma_1), \qquad \epsilon_2 \sim \text{Normal}(0_{p\times 1}, \Sigma_2), \tag{7}$$

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 $(A, \boldsymbol{X}(-1)^{\intercal}, \boldsymbol{X}(1)^{\intercal})^{\intercal}$ from model (7). We train a SVM using the features $\boldsymbol{X}_i(-1)$ and $\boldsymbol{X}_i(1)$, $i = 1, \ldots N$, and take the resulting SVM weights to be the "true" weight vector. Next, we simulate confounding by setting $\boldsymbol{X} = \boldsymbol{X}(Y^{obs})$, where $Y^{\text{obs}} = 2Y^* - 1$, $Y^* =$ 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 $\boldsymbol{X} = \boldsymbol{X}(Y^{test})$, where $Y^{\text{test}} = 2Y^* - 1$, $Y^* =$ 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 $Y^{\text{obs}} = -1$ (CN-Adjusted). Section 3.1 gives details about the regression-based adjustment methods. We use L_2 distance between the true and estimated weight vectors as one criterion for comparison. Figure 6 displays boxplots of the average test accuracy and average L_2 distance from the true weights for $M = 1,000$ iterations. The IPW-SVM performs the best with respect to median L_2 distance and attains the highest median test accuracy. The left plot of Figure 6 has been zoomed-in to better compare the interquartile ranges. The IPW-SVM and Unadjusted SVM resulted in more outliers than the regression-based adjustment methods. The IPW-SVM seems sensitive to very large weights which occured by chance in several iterations. Using stabilized weights provided modest improvement improvement (results not presented here) in this simulation study.

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 [25]. 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.

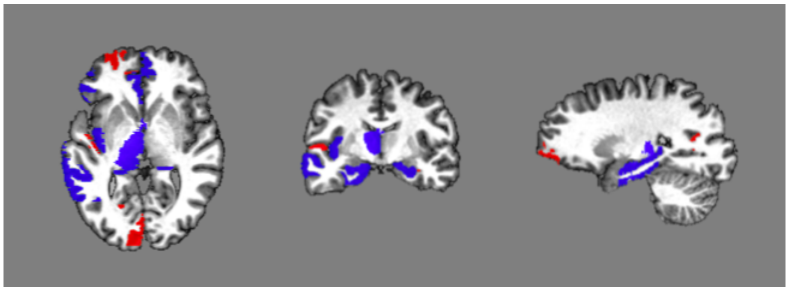

Figure 7: *Top 10 weighted SVM features from the one-to-one age-matched data. Blue (red) regions correspond to negative (positive) weights.*

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 [17] to estimate 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 we use here consist of 224 healthy controls and 327 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 a significant p-value (p=0.003), 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 3.1 with estimated, non-truncated 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 7 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 8 displays the top 10 weighted SVM features from the IPW-SVM, unadjusted SVM, control-adjusted SVM, and adjusted SVM.

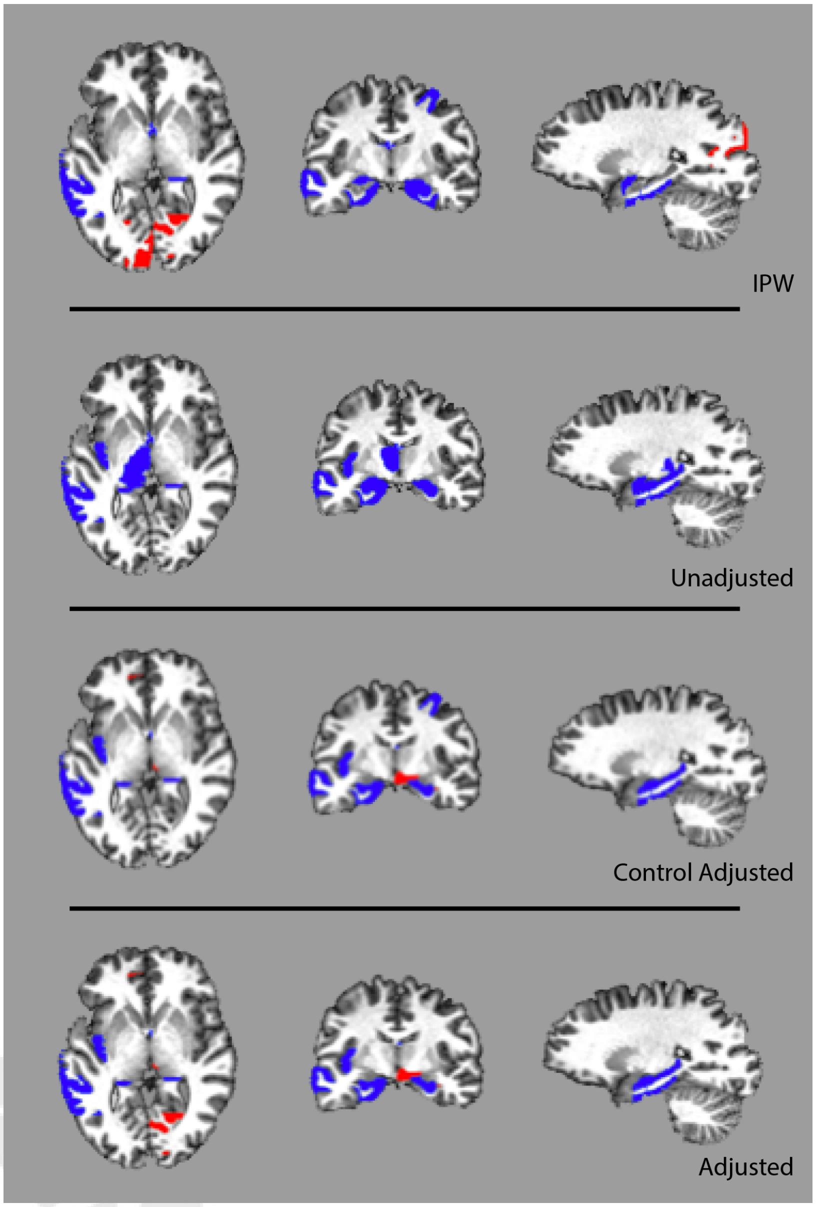

Figure 8: *Top 10 weighted SVM features from the (top to bottom) IPW-SVM, unadjusted SVM, control-adjusted SVM, and adjusted SVM. Blue (red) regions correspond to negative (positive) weights.*

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.

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

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

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 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). In addition, we have distinguished the goal of optimizing biomarker performance from the goal of studying multivariate disease-related changes in the brain. In the former case, IPW may not be appropriate since the true disease prevalance in the population of interest is likely not balanced across the distribution of the confounders. Instead, weighting methods designed to address dataset shift should be applied to optimize and evaluate the performance of the classifier. However, when the goal of MVPA is to estimate complex disease patterns in the brain, using IPW to address confounding is more principled that 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 [23, 24] 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 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, generalizied versions of the propensity score exist for exposures with more than two groups and continuous exposures [30, 33]. 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.

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.

References

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

[2] Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424.

[3] 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.

[4] Bickel, S., Brückner, M., and Scheffer, T. (2009). Discriminative learning under covariate shift. *The Journal of Machine Learning Research*, 10:2137–2155.

[5] 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.

[6] Cole, S. R. and Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American journal of epidemiology*, 168(6):656–664.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] Doshi, J., Erus, G., Ou, Y., Gaonkar, B., and Davatzikos, C. (2013). Multi-atlas skull-stripping. *Academic radiology*, 20(12):1566–1576.

[18] Dukart, J., Schroeter, M. L., Mueller, K., and Initiative, T. A. D. N. (2011). Age correction in dementia – matching to a healthy brain. *PLoS ONE*, 6(7):e22193.

[19] 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.

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

[21] 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.

[22] 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.

[23] 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.

[24] Gaonkar, B., Shinohara, R. T., Davatzikos, C., Initiative, A. D. N., et al. (2015). Interpreting support vector machine models for multivariate group wise analysis in neuroimaging. *Medical Image Analysis*.

[25] Gauthier, S., Reisberg, B., Zaudig, M., Petersen, R. C., Ritchie, K., Broich, K., Belleville, S., Brodaty, H., Bennett, D., Chertkow, H., et al. (2006). Mild cognitive impairment. *The Lancet*, 367(9518):1262–1270.

[26] 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.

[27] Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. (2009). Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5.

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

[29] Hernán, M. A. and Robins, J. M. (2006). Estimating causal effects from epidemiological data. *Journal of epidemiology and community health*, 60(7):578–586.

[30] Hirano, K. and Imbens, G. W. (2004). The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, 226164:73–84.

[31] Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

[32] Huang, J., Gretton, A., Borgwardt, K. M., Schölkopf, B., and Smola, A. J. (2006). Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pages 601–608.

[33] Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710.

[34] Janes, H. and Pepe, M. S. (2008). Matching in studies of classification accuracy: implications for analysis, efficiency, and assessment of incremental value. *Biometrics*, 64(1):1–9.

[35] 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.

[36] 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.

[37] 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.

[38] Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in medicine*, 29(3):337–346.

[39] Li, L., Rakitsch, B., and Borgwardt, K. (2011). ccsvm: correcting support vector machines for confounding factors in biological data classification. *Bioinformatics*, 27(13):i342–i348.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

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

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

[47] R Core Team (2014). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

[48] Raz, N. and Rodrigue, K. M. (2006). Differential aging of the brain: patterns, cognitive correlates and modifiers. *Neuroscience & Biobehavioral Reviews*, 30(6):730–748.

[49] Reiss, P. T. and Ogden, R. T. (2007). Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association*, 102(479):984–996.

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

[51] 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.

[52] Robins, J. M. (1998). Marginal structural models. In *Proceedings of the Section on*

Bayesian Statistical Science, Alexandria, VA: American Statistical Association, pages 1–10.

[53] Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

[54] Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

[55] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

[56] 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.

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

[58] Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244.

[59] Sugiyama, M., Krauledat, M., and Müller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *The Journal of Machine Learning Research*, 8:985–1005.

[60] Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P. V., and Kawanabe, M. (2008). Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in neural information processing systems*, pages 1433–1440.

[61] Sun, D., van Erp, T. G., Thompson, P. M., Bearden, C. E., Daley, M., Kushan, L., Hardt, M. E., Nuechterlein, K. H., Toga, A. W., and Cannon, T. D. (2009). Elucidating a magnetic resonance imaging-based neuroanatomic biomarker for psychosis: classification analysis using probabilistic brain atlas and machine learning algorithms. *Biological psychiatry*, 66(11):1055–1060.

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

[63] 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.

[64] 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.

[65] 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.

[66] Weichwald, S., Meyer, T., Özdenizci, O., Schölkopf, B., Ball, T., and Grosse-Wentrup, M. (2015). Causal interpretation rules for encoding and decoding models in neuroimaging. *NeuroImage*, 110:4859.

[67] 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.

[68] Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, page 114. ACM.

[69] Zipunnikov, V., Caffo, B., Yousem, D. M., Davatzikos, C., Schwartz, B. S., and Crainiceanu, C. (2011). Functional principal component model for high-dimensional brain imaging. *NeuroImage*, 58(3):772–784.