CHAPTER

5

Linear methods for classification

Andre F. Marquand[1,2,3], *Seyed Mostafa Kia*[1,2]

[1] Donders Centre for Cognitive Neuroimaging, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, Netherlands; [2] Department of Cognitive Neuroscience, Radboud University Medical Centre, Nijmegen, Netherlands; [3] Department of Neuroimaging, Centre for Neuroimaging Sciences, Institute of Psychiatry, King's College London, London, United Kingdom

5.1 Introduction

Linear classification models are by far the most widely used machine learning technique in brain disorders research and have been used in hundreds of studies to predict diagnosis or clinical outcome (see Arbabshirani, Plis, Sui, and Calhoun (2017), Brown and Hamarneh (2016), Wolfers, Buitelaar, Beckmann, Franke, and Marquand (2015) and Woo, Chang, Lindquist, and Wager (2017) for reviews). Linear models are usually the method of choice because they are fast and interpretable and often offer equivalent or better predictive performance relative to more complex nonlinear models (https://www.biorxiv.org/content/10.1101/473603v2.full). For example, in a recent challenge for predicting autism from structural and functional neuroimaging data (https://paris-saclay-cds.github.io/autism_challenge/), both the winning model and 7 of the top 10 entries used linear models. Many different algorithms fall into this class including penalized logistic regression, linear discriminant analysis, linear Support Vector Machine (SVM) (see Chapter 6), and linear Gaussian process models. While these models differ in many ways, they can all be understood as aiming to find an optimal balance between fitting the data well and restricting the complexity of the derived solution. Classically, this is operationalized either in a penalized regression model, where a penalty on the coefficients of a linear model limits complexity, or in a Bayesian probabilistic model, where a prior distribution (distribution that conveys some

Machine Learning
https://doi.org/10.1016/B978-0-12-815739-8.00005-5

prior knowledge) is applied to the coefficients of a linear model before computing the posterior distribution over the model coefficients using the rules of probability.

An important feature of linear models is that they allow the predictive weights to be visualized in the input space. In the context of brain disorders research, this is of crucial importance for two reasons. First, it can help to understand which aspects of the pattern underlie the classifier decision; for example, in the case of clinical neuroimaging, which regions are providing the greatest contribution to classification. Second, it can be used to help exclude the possibility that any derived accuracy is driven by nuisance variation. For example, in the case of clinical neuroimaging, patients may move more than controls resulting in head motion differences between groups. This has a significant influence on the observed signal and can bias classification (Power, Barnes, Snyder, Schlaggar, & Petersen, 2012). However, the interpretation of the coefficients of linear models in high-dimensional problems, such as neuroimaging data, can still be challenging (Haufe et al., 2014; Kia, Pons, Weisz, & Passerini, 2017; Kraha, Turner, Nimon, Zientek, & Henson, 2012; Weichwald et al., 2015).

In this chapter, we provide a brief introduction to the linear classifiers commonly used in brain disorders research, highlighting their similarities, differences, and relative strengths and weaknesses. Because the vast majority of existing studies have used structural or functional neuroimaging data, we focus on applications of these classifiers in clinical neuroimaging. We first introduce the basic theory behind penalized regression methods and probabilistic approaches, which are the most important classes of methods in clinical neuroimaging. We discuss "discriminative mapping" (Mourao-Miranda, Bokde, Born, Hampel, & Stetter, 2005), which involves mapping the vector of coefficients across brain regions, along with recent extensions that employ penalties on the coefficients of the linear models that impose sparsity (e.g., Grosenick, Klingenberg, Katovich, Knutson, & Taylor, 2013; Michel, Gramfort, Varoquaux, Eger, & Thirion, 2011). The use of such penalties sets many of the coefficients to zero, which can be desirable to restrict the discriminative pattern to a smaller number of brain regions. We briefly review how linear models can be extended to nonlinear prediction and then highlight how these methods are used in the field. As linear classification methods are extremely prevalent in clinical neuroimaging, we illustrate their multiple applications by providing an overview of studies undertaken, before describing in detail two recent studies for discriminating patients with autism spectrum disorder (ASD) (Yahata et al., 2016) and schizophrenia (de Pierrefeu et al., 2018) from healthy participants. We conclude with some general comments and recommendations for the starting practitioner.

5.2 Method description

5.2.1 Fundamentals and notation

We start with a labeled dataset $D = \{\mathbf{x}_i, y_i\}_{i=1}^{m}$, where $i = 1, \ldots, m$ indexes d-dimensional column vectors ($\mathbf{x}_i = [x_1, \ldots, x_d]^{\mathrm{T}}$) containing the input patterns and targets (y_i), which are class labels for classification. In a clinical neuroimaging data classification scenario considered in this chapter, m is the number of subjects, d is the number of measurements, e.g., voxels in functional magnetic resonance imaging (MRI) data, and $\mathbf{x}_i$ and y_i are, respectively, the neuroimaging data and the clinical label of the i-th subject (e.g., diagnostic labels). Here, we assume binary labels, i.e., $y_i \in \{-1, 1\}$, but other conventions are possible (e.g., "one-hot" encoding for multiclass classification). In the classification scenario considered here, the goal is to find (or learn) a function $f : \mathbf{x} \rightarrow y$ among a family of linear functions, where $y = f(\mathbf{x}^{\mathrm{T}}\mathbf{w} + b)$. The outputs are related to the inputs by a linear combination of the inputs and a vector of coefficients or weights $\mathbf{w}$. Here, b is an offset which can be neglected without loss of generality. This function can be learnt by minimizing a loss function in the empirical risk minimization framework or by maximizing the posterior class probabilities in a probabilistic setting. Then, the resulting function can be used to make a prediction (y^*) on a newly seen sample ($\mathbf{x}^*$) for which the label is unknown.

5.2.2 Empirical risk minimization

A straightforward strategy to optimize the predictive performance of linear classification models is to minimize the error or "loss" that the classification function $f(\mathbf{x})$ obtains. This can be quantified using a "loss function" $\mathscr{L}(y, f(\mathbf{x}))$, which measures the loss associated with the prediction when the true value of the target is y. The simplest loss function is the "0–1 loss" where $\mathscr{L}(y, f(\mathbf{x})) = 1$ if the classifier makes a mistake and zero otherwise. Many other loss functions have been proposed including the logistic loss that forms the basis for logistic regression and the "hinge loss" used in the SVM classifier described in Chapter 6 (see Bishop, 2006; Scholkopf & Smola, 2002). These are summarized in Table 5.1.

The log loss is one of the most common loss functions in statistical machine learning and is well known in the statistical literature in the context of generalized linear modeling (McCullagh & Nelder, 1983). Logistic regression assumes a linear model to describe the log odds ratio of the two classes, i.e., $\mathbf{w}^{\mathrm{T}}\mathbf{x} = \log[p(y = 1|\mathbf{x}) / p(y = -1|\mathbf{x})]$. By also observing that in the binary case the class predictions must sum to 1 (i.e.,

TABLE 5.1 Some basic loss functions.

Name	Definition $L(y_i, f(\mathbf{x}_i))$	Notes
Zero-one loss	$\delta[\text{sign}(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i) \neq y_i]$	• Simple to interpret • Not continuous, not practical to optimize
Hinge loss	$\max[1 - y_i\mathbf{w}^{\mathrm{T}}\mathbf{x}_i, 0]$	• Used in the Support Vector Machine • Penalizes linearly outside a margin • Convex but not smooth
Log loss or logistic loss	$\log[1 + \exp(-y_i\mathbf{w}^{\mathrm{T}}\mathbf{x}_i)]$	• Used in logistic regression • Provides outputs scaled between 0 and 1 • Convex and smooth

$p(y = 1|\mathbf{x}, \mathbf{w}) = 1 - p(y = -1|\mathbf{x}, \mathbf{w})$), this provides a probabilistic prediction such that a data sample is assigned to class 1 with probability:

$$p(y=1|\mathbf{x},\mathbf{w}) = \frac{1}{1+e^{-\mathbf{w}^{\mathrm{T}}\mathbf{x}}} = \sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}\right)$$

The logistic activation function, $\sigma()$, has the effect of constraining the linear predictor to the unit interval and allowing the prediction to be interpreted as a probability. Logistic regression also has a straightforward extension to multiclass classification, where it is referred to as multinomial logistic regression (see elsewhere for details (Bishop, 2006; Hastie, Tibshirani, & Friedman, 2001)). Many other loss functions are possible; for example, the loss function may accommodate asymmetric misclassification costs. This could be desirable in clinical applications where false negative errors (failing to detect a disease when the subject really has it) are potentially much more costly than false positive errors (predicting that a subject has a disease when really they do not).

Given a classification function, $f(\mathbf{x})$, and a loss function, $\mathscr{L}(y, f(\mathbf{x}))$, the overall loss is measured by integrating over the joint distribution of targets and inputs to yield the "expected risk":

$$R[f] = \int L(y, f(\mathbf{x}))p(y, \mathbf{x})dyd\mathbf{x}$$

In practice, the joint distribution $p(y, \mathbf{x})$ is usually unknown so the true expected risk is approximated by the "empirical risk," defined over the training set:

$$R_{\text{emp}}[f] = \int L(y, f(\mathbf{x}))p_{\text{emp}}(y, \mathbf{x})dyd\mathbf{x}$$

$$= \frac{1}{m}\sum_{i=1}^{m} L(y, f(\mathbf{x}))$$

Many classifiers such as penalized linear models (and also the SVM described in Chapter 6) involve minimizing an objective function consisting of the empirical risk and an additional regularization term (see below).

5.2.3 Regularization

In neuroimaging, the dimensionality of the image data is often extremely high relative to the number of data points typically available, i.e., $d \gg m$. In a typical neuroimaging experiment, there are potentially hundreds of thousands to millions of dimensions (e.g., voxels or vertices), whereas the number of samples is typically of the order of hundreds to thousands, implying that neuroimaging classification problems are extremely ill-posed (i.e., have more than one solution) (Lautrup et al., 1995). A common approach to address this involves penalizing model complexity, making it possible to perform accurate prediction even when the problem dimensionality is substantially greater than the number of data points available. This is referred to as regularization and can be traced at least as far back as Tikhonov and Arsenin (1977) who proposed probably the simplest example, known as "ridge regression" (or "Tikhonov regularization"). Formally this achieved by modifying the empirical risk functional defined above in the following way:

$$R_{\text{emp}}[f] = \frac{1}{m}\sum_{i=1}^{m} L(y, f(\mathbf{x})) + \lambda\Omega(\mathbf{w})$$

Here, $\Omega(\mathbf{w})$ is a penalty term that restricts the magnitude of the coefficients and can take many forms; λ is a hyperparameter that controls the amount of regularization applied or, in other words, balances the trade-off between fitting the data accurately and minimizing complexity. Some common regularization penalties are given in Table 5.2.

TABLE 5.2 Common penalty terms for regularizing linear models.

Name	Definition $\Omega(\mathbf{w})$	Notes
Ridge or ℓ_2	$\mathbf{w}^{\mathrm{T}}\mathbf{w} = \lVert\mathbf{w}\rVert_2^2 = \sum_{j=1}^{d} w_j^2$	• Differentiable and convex • Dense (all variables have nonzero coefficients)
Lasso or ℓ_1	$\lVert\mathbf{w}\rVert_1 = \sum_{j=1}^{d} \lvert w_j\rvert$	• Not differentiable everywhere, convex (but not strictly convex) • Sparse (sets some coefficients to exactly zero)
Elastic net	$\alpha\lVert\mathbf{w}\rVert_1 + (1-\alpha)\lVert\mathbf{w}\rVert_2^2 \;\alpha \in [0,1]$	• Combines ℓ_1 and ℓ_2 • Sparse but also allows correlated variables to be included in the model

These penalties have different properties: the ridge penalty is widely applied and is the default choice for many widely used algorithms (e.g., the SVM, described in detail in Chapter 6) and provides the advantages of being easy to optimize and providing stable, often relatively accurate predictions. Sparse penalties that push coefficients to zero are also popular and can assist the interpretation of discriminative patterns such that only a subset of the variables (e.g., voxels or brain regions) contribute to the final predictions. The Lasso penalty (Hastie, Tibshirani, & Friedman, 2009) is the method of choice in many application domains but is not well suited to many neuroimaging problems because it does not accommodate collinear predictor variables and therefore tends toward solutions that are overly sparse and unstable with respect to small perturbations in the data. The elastic net (Zou & Hastie, 2005) combines the lasso and ridge penalties and helps to address these problems to a certain extent in that it provides sparsity and allows correlated variables to enter the model. The elastic net has been relatively widely applied to neuroimaging data (Carroll, Cecchi, Rish, Garg, & Rao, 2009; Marquand et al., 2012). In addition to these generic penalty functions, there is also an increasing interest in applying "structured" and "grouped" sparse penalties that encode additional assumptions, for example, in that the pattern of regression coefficients is spatially sparse but locally smooth (de Brecht & Yamagishi, 2012; Grosenick et al., 2013; Kia, Pedregosa, Blumenthal, & Passerini, 2017; Michel et al., 2011; van Gerven, Hesse, Jensen, & Heskes, 2009). These can be beneficial in modeling the local smoothness of neuroimaging data.

5.2.4 Optimization

Finding the optimal coefficients for linear classification models involves minimizing the empirical risk with respect to the coefficients, or in other words solving the following optimization problem:

$$\min_{\mathbf{w}} \frac{1}{m} \sum_{i=1}^{m} L(y, f(\mathbf{x})) + \lambda \Omega(\mathbf{w})$$

There are many different optimization algorithms and an extended discussion of different strategies for optimization is beyond the scope of this chapter (see Nocedal and Wright (2006) for a detailed treatment). Briefly, the choice of algorithm depends on the choice of loss function and regularizer. For example, smooth functions are easier to optimize than nonsmooth functions and convex functions have the desirable property that they have a unique minimum.

In addition to estimating the coefficients, $\mathbf{w}$, it is also necessary to estimate an optimal value for the regularization hyperparameter, λ, which can have a crucial influence on the derived solution, particularly for sparse models. A standard way to achieve this is to perform nested cross-validation over a grid of hyperparameters (see Chapter 2).

5.2.5 Probabilistic classification

Probabilistic approaches provide an alternative to penalized regression methods and are based on Bayesian probability theory. This can be motivated by the heuristic that to minimize the probability of making an error one should predict choosing the class with the maximum posterior class probability (Bishop, 2006). To provide a simple example, Bayesian logistic regression is a discriminative approach that extends logistic regression. It involves directly modeling the posterior class probability using the logistic function specified in the first equation in Section 5.2.2, i.e., $p(y = 1|\mathbf{x}_i) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i)$. One then places a prior distribution over the model coefficients $p(\mathbf{w}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes any hyperparameters on which the prior depends and computes the posterior distribution over the coefficients using Bayes rule, i.e.,

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}) = \frac{p(\mathbf{w}|\boldsymbol{\theta}) \prod_{i=1}^{m} \sigma\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i\right)}{p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X})}$$

Here, $\mathbf{X}$ is a matrix containing all training data, $p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta})$ is the posterior distribution over the weights and denominator, and $p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{X})$ is the model evidence or marginal likelihood and is obtained by integrating out the weights. Analogous to the penalty functions described above, the prior regularizes the solution in that it constrains the magnitude of the weights. While a Gaussian prior is common and is analogous to the ridge regression penalty, many alternative forms are possible. Inference then proceeds by estimating the posterior distribution over the weights, using deterministic or stochastic approximations or on the basis of Markov chain Monte Carlo methods. For the purposes of the present discussion, there are two important differences between probabilistic approaches and the penalized regression approaches described above: first, probabilistic approaches take into account uncertainty in the parameter estimation and propagate this uncertainty to the predictions. This is crucial in any application domain where quantification of uncertainty is important. Second, making predictions in penalized linear models is effectively a one-step process. In contrast, an important feature of the probabilistic approach is that prediction can be performed in two stages: an *inference* stage, where the posterior class distributions are computed, and a *decision* stage, where the classifier output is used to make optimal class assignments and decide upon an action (see Bishop, 2006). This confers several advantages: (1) the same inference stage can be reused with different loss functions, which allows the risk to be easily reassessed if the misclassification costs change, (2) it permits the machine learning practitioner to include prior information about the function and/or noise model, (3) it may be easily used to compensate for class priors (e.g., if one class is much more common than the other), and (4) it permits a reject option, which means that the classifier only makes a decision if the predictive confidence is high, otherwise it defers the decision to a human or a more sophisticated classifier. Another important feature of probabilistic models is that they provide a convenient method for optimizing the hyperparameters of the model without resorting to cross-validation. This is achieved by maximizing the marginal likelihood of the probabilistic model. The reader is referred elsewhere for details (Bishop, 2006; Rasmussen & Williams, 2006) and examples in neuroimaging (Huertas et al., 2017; Marquand et al., 2010). Finally, and similar to the regularization-based models described above, it is also possible to employ priors that result in sparse solutions. For example, a Bayesian variant of sparse logistic regression has been proposed for neuroimaging data (Yamashita, Sato, Yoshioka, Tong, & Kamitani, 2008). A detailed treatment of this method is outside the scope of this chapter, but briefly, this method is equivalent to the well-known "relevance vector machine" (Tipping, 2001) in that it employs prior that encourages sparsity (see Bishop (2006) and Neal (1996) for further details).

5.2.6 Mapping the discriminating pattern

As noted above, an important feature of linear models is that they allow the model coefficients to be extracted and visualized in the input (e.g., voxel) space. This forms the basis for discriminative mapping (Mourao-Miranda et al., 2005), which can be used to infer the relative contribution of different brain regions to the classifier predictions. As noted above, this can be combined with sparse penalties or priors that result in only a subset of variables having nonzero coefficients (Carroll et al., 2009; Grosenick et al., 2013; Michel et al., 2011). However, the features in neuroimaging data are often highly collinear, which can seriously complicate the interpretation of the coefficients of linear models (Haufe et al., 2014; Kia, Pons, et al., 2017; Kraha et al., 2012; Weichwald et al., 2015). This interpretation is further compounded by the fact that there are often many possible linear combinations of features that can yield the same predictions. There are two main problems when predictor variables are highly collinear: (i) regression coefficients can have a high variance and, therefore, can be unstable with respect to slight perturbations in the data; (ii) care must be taken in the interpretation of high magnitude coefficients because a high magnitude coefficient can arise because a covariate is directly useful in predicting the data or because it acts as a "suppressor" variable (Kraha et al., 2012) in that it helps to cancel out noise or mismatch in other variables. Therefore, to properly interpret the discriminating pattern, it is important to examine alternative methods to understand the importance of different features in predicting class labels. For example, classical statistical methods are often useful to understand the group difference between classes, for example, structure coefficients (Kraha et al., 2012), and other methods proposed specifically for neuroimaging data (e.g., Haufe et al., 2014; Marquand, Brammer, Williams, & Doyle, 2014).

5.2.7 Extensions to nonlinear models

All the linear models described in this chapter can be easily extended to model nonlinear relationships between the data points, for example, using the "kernel trick" (Aizerman, Braverma, & Rozonoer, 1965). This involves transforming the data to an inner product space (which may be higher dimensional), using some nonlinear mapping $\phi\ :\ X \rightarrow F$. X is referred to as the "input space," where the pattern vectors (i.e., the $\mathbf{x}_i$) reside, and F as the "feature space," which is the space in which kernel algorithms actually operate. The basic idea behind the kernel trick is that a classification problem that is not linearly separable in the input space (e.g., voxel space) may become separable in the feature space. However, according to statistical learning theory, the data are always linearly separable in the input space if $d > m$, or in other words linear methods are sufficient to "shatter" the problem (Vapnik, 1995). This is usually the case for structural and functional neuroimaging data; this means that, in clinical neuroimaging, nonlinear kernels generally do not improve predictive accuracy over linear models and also do not permit exact estimation of the model coefficients (although approximations exist; see Bernhard Scholkopf et al. (1999)). Therefore, they will not be considered further here. The reader is referred elsewhere for details (Scholkopf & Smola, 2002).

5.3 Applications to brain disorders

Linear models have been applied in hundreds of studies and to many psychiatric and neurological disorders (Arbabshirani et al., 2017; Wolfers et al., 2015; Woo et al., 2017). Of the methods employed, the linear SVM (described in Chapter 6) has been highly dominant, although it is important to point out that the differences between different methods are relatively minor in terms of predictive accuracy. Many promising results have been obtained: for example, in neurology, high accuracies have been reported across many studies for predicting disease state in dementia (e.g., Davatzikos, Resnick, Wu, Parmpi, & Clark, 2008; Kloppel et al., 2008; Zhang et al., 2011) (see Arbabshirani et al. (2017) for a review), stroke (Saur et al., 2010), Parkinsonian disorders (Filippone et al., 2012; Marquand et al., 2013), and others. Likewise, in psychiatry, many studies have separated controls from patients with major depression, autism, schizophrenia, obsessive-compulsive disorder, bipolar disorder, and attention-deficit hyperactivity disorder (see Orru, Pettersson-Yeo, Marquand, Sartori, and Mechelli (2012) and Wolfers et al. (2015) for reviews). Despite these successes, a number of recent reviews have highlighted common shortcomings to many of these studies (Arbabshirani et al., 2017; Varoquaux, 2018; Wolfers et al., 2015; Woo et al., 2017). They are mostly on small samples (less than 100 patients), contrasting only two conditions (usually completely healthy controls and patients) and they are only validated within a single scanning site (e.g., under cross-validation) and do not validate the models on different sites. Of most concern, the mean accuracy decreases with increasing sample size across all disorders, which suggests strongly that small sample estimates of accuracy are overly optimistic due to "vibration" effects resulting from multiple models being evaluated and only the maximum accuracy obtained being reported (Fig. 5.1).

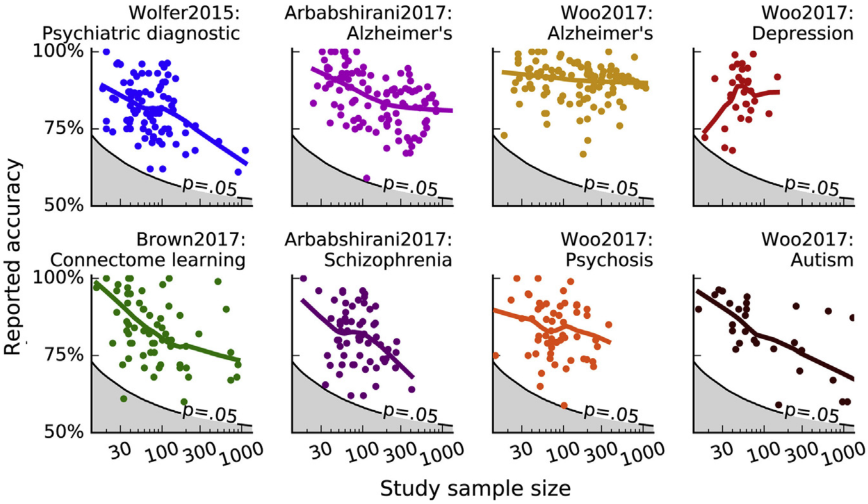

FIGURE 5.1 Figure showing that accuracy reported across different studies decreases with sample size for all disorders. *Taken from Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars.* NeuroImage, 180, *68–77.*

5.3.1 Classification of autism spectrum disorder

In view of the considerations above, it is clear that validation of prospective classification models is very important. For example, many studies have aimed to predict ASD diagnosis using structural and functional MRI (e.g., Ecker, Marquand, et al., 2010; Ecker, Rocha-Rego, et al., 2010); see Orru et al. (2012) and Wolfers et al. (2015) for reviews. However, until recently such studies were estimated on small samples derived from a single scanning site, leaving it uncertain how accurately such classifiers can be generalized to more realistic and heterogenous samples. To help to answer this question, Yahata and colleagues (Yahata et al., 2016) recently proposed an approach to discriminate patients with ASD from typically developed controls (TDC) on the basis of functional connectivity features. This classifier was trained on a cohort of 174 Japanese participants (74 with ASD and 107 TDC) from three scanning sites. This was then applied to make predictions on two independent validation cohorts, one from Japan (27 ASD and 27 TDC) and one from North America (44 ASD and 44 TDC).

The authors employed a classification pipeline based on sparse Bayesian logistic regression (Yamashita et al., 2008), combined with a novel variable selection procedure based on sparse canonical correlation analysis. This variable selection procedure aims to identify variables (i.e., connections between brain regions) that are not confounded by nuisance variables (e.g., site differences). The authors reported an accuracy of 85% (area under the receiver operating characteristic curve (AUC) = 0.93) for discriminating ASD and TDC participants under leave-one-subject-out cross-validation of the training sample, which involves repeatedly retraining the classifier, excluding each subject once. When applied to the validation cohorts, this classifier obtained an accuracy of 75% (AUC = 0.76) for the North American sample and an accuracy of 70% (AUC = 0.77) for the second Japanese sample. The authors also report that this classifier was driven by a small number of connectivity features (15 ± 0.7 connections = 0.2% of the total number) and further showed that these features were also predictive of symptom scores measured by the autism diagnostic schedule A (Lord et al., 1989) and the Autism Diagnostic Interview-Revised (Lord, Rutter, & Lecouteur, 1994). Finally, the authors also applied this classifier to other psychiatric disorders to evaluate the generalization of these classifiers to other disorders and to identify those which share similar features in terms of brain connectivity. Among the disorders evaluated (schizophrenia, attention-deficit disorder, and major depression), only schizophrenia showed a similar profile, which the authors interpret as evidence for partially shared underlying mechanisms.

5.3.2 Classification of schizophrenia

The second example we consider aims to discriminate patients with schizophrenia from healthy control subjects using penalized linear models (de Pierrefeu et al., 2018). Like the first study, this study used a multisite dataset containing 526 participants across four sites where cross-validation was performed using a leave-one-site-out (LOSO) procedure whereby the classifier was repeatedly trained on all sites except one, before making predictions for the withheld site. In addition, the authors evaluate the accuracy of the estimated classifiers on a separate validation dataset containing 133 participants acquired at a different site, therefore providing a realistic estimate of generalization capacity. This is important because—like bipolar disorder—a wide range of accuracies have been reported in the literature for discriminating schizophrenia patients from controls (Wolfers et al., 2015).

The authors of this study evaluated the predictive capability of multiple features derived from structural MRI data, including estimates of voxel-wise gray matter derived from voxel-based morphometry (Ashburner & Friston, 2000), vertex-wise cortical thickness features derived from FreeSurfer, and region-of-interest estimates of brain volume. A key feature of this work was that it used linear models with penalties that give rise to a sparse pattern that restricts the discriminating pattern to a subset of brain regions. For this, the authors investigated two different penalties: an elastic net penalty that mixes the ridge and lasso penalties (see Table 5.2) and a "total variation" penalty that provides structured sparsity and encourages spatial sparsity and local piecewise constant smoothness. As a reference, the authors also evaluated a linear SVM. The optimization of the total variation objective is quite challenging because the penalty is not smooth. In this work, the authors make use of a custom optimization algorithm for this purpose (Hadj-Selem et al., 2018).

Under LOSO cross-validation, all classifiers performed approximately equivalently, yielding moderate accuracy for discriminating schizophrenia (SVM: 64%–72%, depending on the features used; elastic net: 61%–71%; total variation: 66%–69%). Importantly, this was approximately equivalent to the accuracy obtained on the data from the validation site (SVM: 64%–71%, total variation: 61%–73%). An additional outcome from this study is a set of discriminating weights for each of the methods, which illustrate the effect of different penalties on the regression coefficients (see Fig. 5.2). The SVM is a dense method and so has nonzero coefficients for every voxel or vertex input into the method. The elastic net bases its prediction on a set of scattered brain regions, whereas the total variation penalty prefers a few larger regions. While it is tempting to prefer the pattern of coefficients provided by total

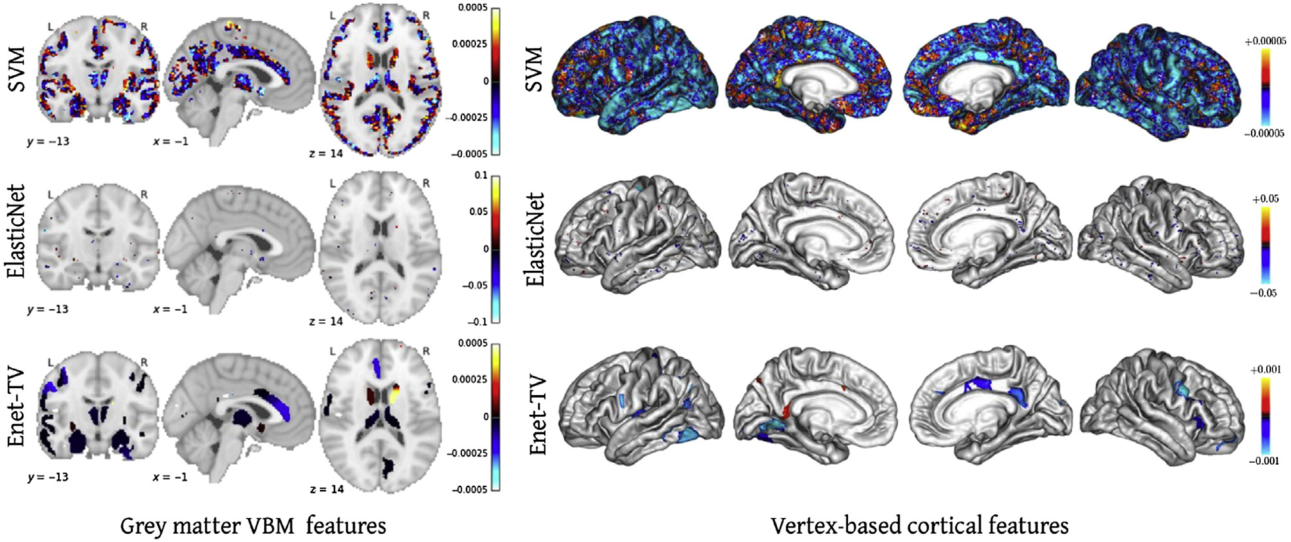

FIGURE 5.2 The effect of different regularization penalties for classifiers discriminating patients with schizophrenia from controls. *Enet-TV*, elastic net with total variation; *SVM*, Support Vector Machine. The color scale indicates the relative magnitude of different components. Warm colors reflect relatively greater values for cortical thickness in patients, whereas cool colors reflect relatively greater values for controls. *Taken from de Pierrefeu, A., Lofsted, T., Hadj-Selem, F., Bourgin, J., Hajek, T., Spaniel, F., et al. (2018). Identifying a neuroanatomical signature of schizophrenia, reproducible across sites and stages, using machine learning with structured sparsity.* Acta Psychiatrica Scandinavia, 138*(6), 571—580.*

variation over the others due to its resemblance to mass-univariate statistic images and its concentration of predictive weight in a few brain regions, it is important to remember that all patterns are equally accurate and—at least on the basis of accuracy—provide alternative and equally appropriate methods by which the classes can be discriminated. It may be possible, however, to adjudicate between the methods using other metrics (e.g., reproducibility of the discriminative patterns across cross-validation folds).

5.4 Conclusion

In this chapter, we introduced some of the most important methods for linear classification of brain imaging data. Linear methods are by far the most widely used machine learning approach in clinical neuroimaging, and when properly validated (e.g., on unseen samples) they often show performance that equals or exceeds that of more complex nonlinear models. We gave a brief overview of the most important methods for linear classification—penalized regression methods and probabilistic methods—and illustrated how many methods can be considered to be members of one of these classes, for example, by combining a logistic loss function with a penalty over the coefficients. We illustrated this discussion by first providing an overview of the studies published so far, followed by highlighting two recent applications for discriminating ASD and schizophrenia from healthy controls. Taken together, these studies provide a sobering context for the use of pattern recognition methods in neuroimaging and highlight that the field has many challenges to overcome. First, they illustrate that many of the small samples that dominate the literature provide optimistic estimates of predictive accuracy and that in larger, multisite studies discrimination accuracy is considerably lower (in the range of 70% or even lower for discriminating healthy controls from patients with psychiatric disorders). This strongly suggests that even simple linear models can overfit in small samples, therefore more complex models (e.g., nonlinear kernel methods and deep learning) should also be carefully scrutinized to ensure overfitting has not taken place. Second, the accuracies obtained on validation samples are well below levels that might be considered clinically useful and are even lower in the few studies that aim to come closer to the clinical reality than the proof-of-concept studies that discriminate between patients and healthy controls. For example, few studies have attempted to perform a differential diagnosis (e.g., between different psychiatric disorders) and those that have generally report even lower accuracy than for discriminating patients from controls (Wolfers et al., 2015). These considerations notwithstanding, linear classification methods are likely to remain important in clinical neuroimaging and can be used to measure the separation between groups on the basis of a multivariate pattern of effects, for example, to quantify how accurately a candidate biomarker separates clinical groups and for making predictions of disease course or outcome, which are difficult to make in other ways.

5.5 Key points

- Linear models are a widely used and powerful approach for clinical neuroimaging.
- They are fast, accurate, and interpretable in that they allow model coefficients to be extracted and visualized.
- Many widely used algorithms are members of this class including SVM and penalized logistic regression.
- Penalized regression models are defined by the choice of loss function and regularizer.
- Probabilistic models are derived by applying a prior to the regression coefficients then computing the posterior distribution.
- There have been many applications to brain disorders; however, many have been based on small and single-site datasets.
- Validation in large and multisite samples is essential for all methods to accurately estimate generalizability.

References

Aizerman, M. A., Braverma, E. M., & Rozonoer, L. I. (1965). Theoretical foundations of potential function method in pattern recognition learning. *Automation and Remote Control, 25*(6), 917–936.

Arbabshirani, M. R., Plis, S., Sui, J., & Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage, 145*, 137–165.

Ashburner, J., & Friston, K. J. (2000). Voxel-based morphometry - the methods. *NeuroImage, 11*(6), 805–821.

Bishop, C. (2006). *Pattern recognition and machine learning*. New York: Springer.

de Brecht, M., & Yamagishi, N. (2012). Combining sparseness and smoothness improves classification accuracy and interpretability. *NeuroImage, 60*(2), 1550–1561.

Brown, C. J., & Hamarneh. (2016). *Machine learning on human connectome data from mri*. ArXiV, 1611.08699.

Carroll, M. K., Cecchi, G. A., Rish, I., Garg, R., & Rao, A. R. (2009). Prediction and interpretation of distributed neural activity with sparse models. *NeuroImage, 44*(1), 112–122.

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

Ecker, C., Marquand, A., Mourao-Miranda, J., Johnston, P., Daly, E. M., Brammer, M. J., et al. (2010). Describing the brain in autism in five dimensions-magnetic resonance imagingassisted diagnosis of autism spectrum disorder using a multiparameter classification approach. *Journal of Neuroscience, 30*(32), 10612−10623.

Ecker, C., Rocha-Rego, V., Johnston, P., Mourao-Miranda, J., Marquand, A., Daly, E. M., et al. (2010). Investigating the predictive value of whole-brain structural MR scans in autism: A pattern classification approach. *NeuroImage, 49*(1), 44−56.

Filippone, M., Marquand, A., Blain, C., Williams, S., Mourao-Miranda, J., & Girolami, M. (2012). Probabilistic prediction of neurological disorders with a statistical assessment of neuroimaging data modalities. *Annals of Applied Statistics, 6*, 1883−1905.

van Gerven, M., Hesse, C., Jensen, O., & Heskes, T. (2009). Interpreting single trial data using groupwise regularisation. *NeuroImage, 46*(3), 665−676.

Grosenick, L., Klingenberg, B., Katovich, K., Knutson, B., & Taylor, J. E. (2013). Interpretable whole-brain prediction analysis with GraphNet. *NeuroImage, 72*(0), 304−321.

Hadj-Selem, F., Lofstedt, T., Dohmatob, E., Frouin, V., Dubois, M., Guilemot, V., et al. (2018). Continuation of Nesterov's smoothing for regression with structured sparsity in high-dimensional neuroimaging. *IEEE Transactions on Medical Imaging, 1*, 2403−2413.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, prediction and inference*. New York: Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). New York: Springer.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., et al. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage, 87*(0), 96−110.

Huertas, I., Oldehinkel, M., van Oort, E. S. B., Garcia-Solis, D., Mir, P., Beckmann, C. F., et al. (2017). A Bayesian spatial model for neuroimaging data based on biologically informed basis functions. *NeuroImage, 161*, 134−148.

Kia, S. M., Pedregosa, F., Blumenthal, A., & Passerini, A. (2017). Group-level spatio-temporal pattern recovery in MEG decoding using multi-task joint feature learning. *Journal of Neuroscience Methods, 285*, 97−108.

Kia, S. M., Pons, S. V., Weisz, N., & Passerini, A. (2017). Interpretability of multivariate brain maps in linear brain decoding: Definition, and heuristic quantification in multivariate analysis of MEG time-locked effects. *Frontiers in Neuroscience, 10*.

Kloppel, S., Stonnington, C. M., Chu, C., Draganski, B., Scahill, R. I., Rohrer, J. D., et al. (2008). Automatic classification of MR scans in Alzheimers disease. *Brain, 131*, 681−689.

Kraha, A., Turner, H., Nimon, K., Zientek, L. R., & Henson, R. K. (2012). Tools to support interpreting multiple regression in the face of multicollinearity. *Frontiers in Psychology, 3*, 44.

Lautrup, B., Hansen, L. K., Law, I., Morch, N., Svarer, C., & Strother, S. C. (1995). *Massive weight sharing: A cure for extremely ill-posed problems. Paper presented at the proceedings of the workshop on supercomputing in brain research: From tomography to neural networks, Julich*.

Lord, C., Rutter, M., Goode, S., Heemsbergen, J., Jordan, H., Mawhood, L., et al. (1989). Autism diagnostic observation schedule - a standardized observation of communicative and social-behavior. *Journal of Autism and Developmental Disorders, 19*(2), 185−212.

Lord, C., Rutter, M., & Lecouteur, A. (1994). Autism diagnostic interview-revised - a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders, 24*(5), 659−685.

Marquand, A. F., Brammer, M., Williams, S. C. R., & Doyle, O. M. (2014). Bayesian multi-task learning for decoding multi-subject neuroimaging data. *NeuroImage, 92*, 298−311.

Marquand, A. F., Filippone, M., Ashburner, J., Girolami, M., Mourao-Miranda, J., Barker, G. J., et al. (2013). Automated, high accuracy classification of Parkinsonian disorders: A pattern recognition approach. *PLoS One, 8*(7). e69237−e69237.

Marquand, A., Howard, M., Brammer, M., Chu, C., Coen, S., & Mourao-Miranda, J. (2010). Quantitative prediction of subjective pain intensity from whole-brain fMRI data using Gaussian processes. *NeuroImage, 49*(3), 2178−2189.

Marquand, A. F., O'Daly, O. G., Simoni, S. D., Alsop, D. C., Maguire, R. P., Williams, S. C. R., et al. (2012). Dissociable effects of methylphenidate, atomoxetine and placebo on regional cerebral blood flow in healthy volunteers at rest: A multi-class pattern recognition approach. *NeuroImage, 60*(2), 1015−1024.

McCullagh, P., & Nelder, J. (1983). *Generalized linear models*. London: Chapman and Hall.

Michel, V., Gramfort, A., Varoquaux, G.l., Eger, E., & Thirion, B. (2011). Total variation regularization for fMRI-based prediction of behavior. *IEEE Transactions on Medical Imaging, 30*(7), 1328−1340.

Mourao-Miranda, J., Bokde, A. L., Born, C., Hampel, H., & Stetter, M. (2005). Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional MRI data. *NeuroImage, 28*(4), 980−995.

Neal, R. (1996). *Probabilistic inference using Markov-chain Monte Carlo methods*. Technical Report, University of Toronto.

Nocedal, J., & Wright, S. (2006). *Numerical optimization*. Springer.

Orru, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G., & Mechelli, A. (2012). Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review. *Neuroscience and Biobehavioural Reviews, 36*, 1140−1152.

de Pierrefeu, A., Lofsted, T., Hadj-Selem, F., Bourgin, J., Hajek, T., Spaniel, F., et al. (2018). Identifying a neuroanatomical signature of schizophrenia, reproducible across sites and stages, using machine learning with structured sparsity. *Acta Psychiatrica Scandinavia, 138*(6), 571−580.

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage, 59*(3), 2142−2154.

Rasmussen, C. E., & Williams, C. (2006). *Gaussian processes for machine learning*. MIT Press.

Saur, D., Ronneberger, O., Kummerer, D., Mader, I., Weiller, C., & Kloppel, S. (2010). Early functional magnetic resonance imaging activations predict language outcome after stroke. *Brain, 133*, 1252−1264.

Scholkopf, B., Mika, S., Burges, C. J. C., Knirsch, P., Müller, K. R., Rätsch, G., et al. (1999). Input space versus feature space in kernel-based methods. *IEEE Transactions on Neural Networks, 10*(5), 1000−1017.

Scholkopf, B., & Smola, A. (2002). *Learning with kernels. Support vector machines, regularization, optimization and beyond*. MIT Press.

Tikhonov, A. N., & Arsenin, V. Y. (1977). *Solution of ill-posed problems*. Washington, DC: Winston.

Tipping, M. E. (2001). Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research, 1*, 211−244.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer.

Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage, 180*, 68−77.

Weichwald, S., Meyer, T., Ozdenizci, O., Scholkopf, B., Ball, T., & Grosse-Wentrup, M. (2015). Causal interpretation rules for encoding and decoding models in neuroimaging. *NeuroImage, 110*, 48−59.

Wolfers, T., Buitelaar, J. K., Beckmann, C., Franke, B., & Marquand, A. F. (2015). From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neuroscience and Biobehavioral Reviews, 57*, 328−349.

Woo, C. W., Chang, L. J., Lindquist, M. A., & Wager, T. D. (2017). Building better biomarkers: Brain models in translational neuroimaging. *Nature Neuroscience, 20*(3), 365−377.

Yahata, N., Morimoto, J., Hashimoto, R., Lisi, G., Shibata, K., Kawakubo, Y., et al. (2016). A small number of abnormal brain connections predicts adult autism spectrum disorder. *Nature Communications, 7*.

Yamashita, O., Sato, M., Yoshioka, T., Tong, F., & Kamitani, Y. (2008). Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. *NeuroImage, 42*(4), 1414−1429.

Zhang, D. Q., Wang, Y. P., Zhou, L. P., Yuan, H., Shen, D. G., & Alzheimers Dis Neuroimaging, I. (2011). Multimodal classification of Alzheimer's disease and mild cognitive impairment. *NeuroImage, 55*(3), 856−867.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 67*, 301−320.