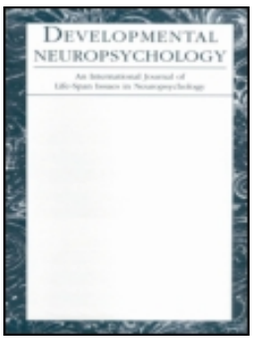

Developmental Neuropsychology

ISSN: 8756-5641 (Print) 1532-6942 (Online) Journal homepage: http://www.tandfonline.com/loi/hdvn20

Novel Machine Learning Methods for ERP Analysis: A Validation From Research on Infants at Risk for Autism

Daniel Stahl , Andrew Pickles , Mayada Elsabbagh , Mark H. Johnson & The BASIS Team

To cite this article: Daniel Stahl , Andrew Pickles , Mayada Elsabbagh , Mark H. Johnson & The BASIS Team (2012) Novel Machine Learning Methods for ERP Analysis: A Validation From Research on Infants at Risk for Autism, Developmental Neuropsychology, 37:3, 274-298, DOI: 10.1080/87565641.2011.650808

To link to this article: http://dx.doi.org/10.1080/87565641.2011.650808

Published online: 30 Apr 2012.

DEVELOPMENTAL NEUROPSYCHOLOGY, *37*(3), 274–298

ISSN: 8756-5641 print / 1532-6942 online
DOI: 10.1080/87565641.2011.650808

Psychology Press
Taylor & Francis Group

Novel Machine Learning Methods for ERP Analysis: A Validation From Research on Infants at Risk for Autism

Daniel Stahl and Andrew Pickles

Department of Biostatistics, Institute of Psychiatry, King's College London, London, United Kingdom

Mayada Elsabbagh

Department of Psychiatry, Faculty of Medicine, McGill University, Montreal, Canada

Mark H. Johnson

Centre for Brain and Cognitive Development, Birkbeck, University of London, London, United Kingdom

The BASIS Team

Machine learning and other computer intensive pattern recognition methods are successfully applied to a variety of fields that deal with high-dimensional data and often small sample sizes such as genetic microarray, functional magnetic resonance imaging (fMRI) and, more recently, electroencephalogram (EEG) data. The aim of this article is to discuss the use of machine learning and discrimination methods and their possible application to the analysis of infant event-related potential (ERP) data. The usefulness of two methods, regularized discriminant function analyses and support vector machines, will be demonstrated by reanalyzing an ERP dataset from infants (Elsabbagh et al., 2009). Using cross-validation, both methods successfully discriminated above chance between groups of infants at high and low risk of a later diagnosis of autism. The suitability of machine learning methods for the use of single trial or averaged ERP data is discussed.

Non-invasive neurophysiological measurement of infants is critical in increasing our understanding of the neuro-cognitive mechanisms that underlie typical and atypical development and, in

The BASIS Team, in alphabetical order: Simon Baron-Cohen (University of Cambridge), Patrick Bolton (King's College London), Tony Charman (Institute of Education, London), Holly Garwood (Birkbeck, University of London), Karla Holmboe (Birkbeck, University of London), Leslie Tucker (Birkbeck, University of London), and Agnes Volein (Birkbeck, University of London).

We acknowledge financial support from UK Medical Research Council PG0701484 and Autism Speaks (1292) to MHJ.

Correspondence should be addressed to Daniel Stahl, Department of Biostatistics, Institute of Psychiatry, King's College London, De Crespigny Park, P.O. Box 20, London, SE5 8AF, United Kingdom. E-mail: daniel.r.stahl@kcl.ac.uk

the long term, for identifying biomarkers of atypicality, possible "red-flags" that signal the need for early intervention. The electroencephalogram (EEG) measures brain electrical activity from electrodes placed on the surface of the scalp. The EEG signals are derived from multiple brain sources and are a measure of the brain's ongoing activity. The event-related potential (ERP) reflects changes in the electrical activity of the brain that are time-locked to the occurrence of a specific event, that is, a response to a discrete external stimulus or an internal mental process (see Fabiani, Gratton, & Federmeier, 2007 for an introduction).

ERP measurements allow non-invasive neurophysiological measurements of infants at a high temporal resolution, enabling us to assess cognitive processes and any dysfunctions that may not be apparent at the behavioral level (Sanei & Chambers, 2007; Woodman, 2010). Furthermore, EEG/ERP techniques are safe and relatively easy to use, which make them of greater application for the study of cognitive processes in infants. Not surprisingly, ERPs are, despite recent advances in functional magnetic resonance and other brain imaging methods, the most commonly used methods to study the neural underpinnings of both typical and atypical cognitive development (de Haan, 2007).

In this article, we describe some of the methodological challenges of analyzing ERP data and discuss potential drawbacks of commonly used analysis methods based on univariate mean group comparisons of averaged ERPs. We then introduce discriminant and machine learning methods, two methods that can potentially reduce some of these problems. We apply those methods to a previously published data set by Elsabbagh et al. (2009), a study that used ERPs to discriminate between groups of infants at high and low-risk for later autism. We discuss the potential of the new methods for improving ERP analysis in similar studies, as well as in different research settings, and the possibility of extending them to single trial data. Further, we briefly explain how the new analytic methods can also be applied in situations where comparing the response of infants under different experimental conditions, or between different age groups, is the aim rather than prognostic classification.

Analyzing ERPs requires translating the wave into measurable components. ERPs are commonly quantified by measuring the amplitude and latency of observable peaks of the signal time-locked stimulus or response event. However, due to the physiological and physical properties of the scalp and the brain, the signal-to-noise ratio of ERP recordings is often low, although the thinner skull early in life may reduce this problem somewhat in infants. Measurements of amplitude and latency of observed peaks differ between trials and the variance will be larger at smaller signal-to-noise ratios. Faced with such complexity, the standard practice is to average the response over a large number of trials (Luck, 2005; de Boer, Scott, & Nelson, 2007). The averaging process (over trials and electrode locations) is intended to filter out noise that is not related to the stimulus presentation and results in an ERP waveform related to the stimulus processing. An averaged ERP wave form is regarded as a series of waves with positive and negative reflections, and the traditional way to analyze this is to measure the amplitude and latencies of a small number of distinctive peaks and troughs that commonly occur (Picton et al., 2000; de Haan, 2007). For example, following presentation of a visual stimulus to infants the following characteristic components are typically observed: P100 (positive potential that usually appears around 100 msec after stimulus presentation over posterior channels), N290 (negative potential appearing around 290 msec over posterior and temporal channels), P350/400 (another positive potential observed around 400 msec, Negative Central (latency around 700 msec), and the late slow wave (latency around 2 sec).

Measuring amplitude and latencies of an electrophysiological peak dates back to Richard Caton in the late nineteenth century (Molfese, Molfese, & Kelly, 2001). Although there was no a priori reason to believe that these peaks represented a biologically interesting aspect of a cognitive process, measuring amplitudes and latencies of specific peaks has been successfully used in hundreds of studies (Gazzaniga, 2004; Luck, 2005; Handy, 2005). In general, amplitude is regarded as a measure of large-scale neural activity associated with the information processing, while latency provides a measure of the rapidity, or sequences, of brain information processing (Molfese et al., 2001). It is now well established that components of an ERP are related to specific cognitive, and other, brain processes and that peaks and latencies are a useful measure of those processes. For a review of using ERPs to study infants' cognitive development see Thierry (2005).

Researchers have more recently focused on improving ERP recording and analysis, particularly within the context of studying atypical or at-risk groups. For instance, an ERP analysis relies on averaging a large number of artifact-free trials. In developmental studies, a large proportion of infants may need to be excluded because they do not provide enough trials per condition to obtain a stable averaged ERP (Stets, Stahl, & Reid, 2012). Furthermore, averaging assumes the consistency of response to the stimulus. Studies by Nikkel and Karrer (1994), Snyder, Webb, & Nelson (2002), and Stets and Reid (2011) showed that the assumption of consistency of response to the stimuli might not be valid due to habituation or other processes. A violation of this assumption could cause a loss of statistical power and a statistical bias due to major alterations in the ERP-components, as discussed by Stahl, Parise, Hoehl, and Striano (2010) and suggested by Stets and Reid (2011).

Averaging also does not take into account possible between trial variability due to cognitive processes. For example, Lazzaro et al. (1997) did not detect significant differences in P300 amplitude or latency between a group of adolescents with attention deficit and hyperactivity disorder (ADHD) and a control group. However, the individuals with ADHD showed significantly increased between trial variability over controls, and this variability was significantly reduced after medication. Further problems of averaging single trial ERPs are discussed in Spencer (2005).

Finally, the standard procedure of averaging trials can result in a multiple testing problem. The analysis of ERP measurements usually involves univariate mean group comparisons, such as comparing responses between a clinical and a control group, or comparing response characteristics of participants between two or more experimental conditions. Group differences are commonly assessed for several amplitudes and latencies at several locations (single electrodes or averages of a set of electrodes), which either causes an increase of family-wise type I error due to multiple testing (p-values are too optimistic and confidence intervals too narrow), or a decrease of power if the alpha error level is adjusted for multiple testing using Bonferroni or similar methods. It would, therefore, be desirable to have statistical methods that improve the statistical power and avoid biased parameter estimates in the analysis of ERP studies.

The aim of this article is twofold. First, we will present new ideas for analyzing ERP data of averaged responses. Specifically, we will introduce new methods of group discrimination, regularized discriminant function analysis and support vector machines in particular. These methods allow the analysis of datasets with a large number of variables relative to sample size, and avoid multiple testing problems of current standard analysis methods. We present cross-validation methods to assess the predictive performance of a derived model, thereby avoiding multiple testing problems. Second, we will apply the two methods of classification, regularized

discriminant function analysis and support vector machines, to a data set of averaged responses from a recently published dataset (Elsabbagh et al., 2009) and compare the performance with standard analysis methods. We will discuss the possibility of applying these methods for the analysis of single trial ERP data, which is not possible with standard statistical methods. Finally, we will explain how the proposed methods can be used not only for classification problems, but also for the analysis of experimental studies or randomized controlled trials using ERP recordings.

MACHINE LEARNING AND SUPERVISED CLASSIFICATION METHODS

Machine learning is the study of how computers can learn patterns in empirical data. Supervised classification (or discrimination) methods are a subset of this field, where predictive models of group membership are built, based on observed characteristics of each case. Traditionally machine learning methods treat the description of the relationship between observed variables and group membership as a "black box," and do not assume a probabilistic data model. These models are mainly concerned with generating algorithms that will be good predictors of future observations. The boundaries between machine learning and statistical modelling have disintegrated in recent years. The two approaches are now combined in statistical learning theory, which studies within a statistical framework the properties of particular learning algorithms (Bousquet, Chapelle, & Hein, 2004; Hastie, Tibshirani, & Friedman, 2009).

Machine learning methods have been successfully applied to a variety of fields that deal with high-dimensional data, often accompanied by small sample sizes (Bishop, 2006). These include speech recognition, automatic character or handwriting recognition, text classification (e.g., spam filters), face recognition, protein fold prediction, analysis of genetic microarray data, automatic segmentation of digital images, computer vision, signal processing, and functional magnetic resonance imaging (fMRI).

There are now numerous machine learning classification methods, and the literature suggests that different methods are suited to different kinds of data (Cristianini & Shawe-Taylor; 2000; Müller et al., 2008; Park & Park, 2008; Bandt, Weymar, Samaga, & Hamm, 2009; Blankertz, Lemm, Treder, Haufe, & Müller, 2010; Das, Giesbrecht, & Eckstein, 2010; Doyle, Temko, Lightbody, Marnane, & Boylan, 2010; Doyle et al., submitted). In recent years, nonlinear discriminant function analysis and machine learning methods, such as support vector machines, have become increasingly popular for the classification of continuous or single trial EEG signals, especially in brain–computer interface studies (Lotte, Congedo, Lécuyer, Lamarche, & Arnaldi, 2007; Müller et al., 2008; Blankertz et al., 2010). Support vector machines have also been successfully applied to the prediction of neurodevelopmental disability outcomes in newborns based on EEG recordings (Doyle et al., submitted).

SUPERVISED CLASSIFICATION AND DISCRIMINATION

Supervised classification and discrimination methods cover a wide range of techniques which aim to model the relationship between a set of training data **x** (in this case ERP signals) and their respective group labels **y** (in this specific case, high or low risk for later autism groups). More precisely, **x** is an n $\times$ p matrix (where there are p measurements on n cases). Each row represents a case (here an infant) and each column a variable ("feature variable"). Table 1a shows a

TABLE 1
(a) Hypothetical Data Set of 10 Subjects With Measurements of Four Feature Variables e1 to e4. (b) The Variance/Covariance Matrix: The Diagonal Shows the Variance of Each of the Four Feature Variables and the Covariances in the Off-Diagonal. For Example, the Variance for e1 is 3.7 and its Covariance With e2 is 2.7. (c) Correlation Matrix of the Example Data Set. The Off-Diagonals Show the Correlation Between the Variables. The Standardized Variance is 1 for All Variables

(a) Data Set (10 × 4 Matrix)

e1	*e2*	*e3*	*e4*
1.9	1.9	1.2	4.8
0.8	1.8	1.9	2.4
0.5	1.0	0.2	1.0
2.2	2.8	2.6	4.3
3.8	4.0	2.0	3.7
0.6	0.5	0.7	3.8
1.8	3.1	3.1	6.1
1.6	1.8	1.4	2.6
1.4	1.8	1.6	4.0
2.2	2.4	0.9	3.2

(b) Variance/Covariance Matrix (or Just Covariance Matrix)

	e1	*e2*	*e3*	*e4*
e1	3.7	2.7	2.8	3.0
e2	2.7	6.6	4.5	8.6
e3	2.8	4.5	5.0	5.6
e4	3.0	8.6	5.6	18.3

(c) Correlation Matrix

	e1	*e2*	*e3*	*e4*
e1	1	0.54	0.65	0.36
e2	0.54	1	0.79	0.78
e3	0.65	0.79	1	0.58
e4	0.36	0.78	0.58	1

hypothetical n × p data matrix of p = 4 feature variables with measures on n = 10 subjects. The information of **x** is often summarized as variances (measures of the variability of data taken by a variable around its mean) and covariances (measuring the extent two variables move together, or covary, in the same or opposite direction). Variance and covariances are displayed as a p × p variance–covariance matrix or simple covariance matrix, where the variances of the feature variables appear along the diagonal and the covariances appear symmetrically on the off-diagonals. Table 1b shows the covariance matrix for the example data set. A standardized covariance is a correlation. The correlation matrix is shown in Table 1c. A sample variance and covariance matrix

is usually used as an estimate of the population covariance matrix. Classification methods such as discriminant analysis and support vector machines are often based on covariance matrices.

The main goal of supervised classification methods is to devise rules derived from measurements of the independent or feature variables that can allocate unseen cases into these groups as accurately as possible. The performance of the model can then be evaluated by "showing" it unseen test data and obtaining a label estimate. Comparing the classification of the test cases with the known group membership allows the true accuracy of the classification model to be estimated. We will introduce and compare the performance of three different supervised classification methods: the traditional linear discriminant analysis, a nonlinear regularized discriminant analysis and support vector machine.

VALIDATION OF THE CLASSIFIER

The future performance of a classification rule or "classifier" can be assessed in several ways. Supervised learning techniques used by a classifier are inductive, meaning that they extract the form of their function from data (Fielding, 2007). It is necessary to assess the predictive power of a classifier using validation procedures. Predicting cases using data from which the classification rule is derived (resubstitution) is misleading because it overestimates the true hit rate (Molinaro, Simon, & Pfeiffer, 2006). External methods of classification quality are needed, that test accuracy in a sample that has not been used in the construction of the classification rule. If sample size is large enough, the data set can be randomly split into training and test sub-samples. The classification rule will be derived from the training sub-sample, while the accuracy will be assessed using the test set ("hold-out method"). The true accuracy rate of the classifier is estimated by the percentage of test set cases that are correctly classified.

However, sample size is often not large enough to split the sample in this way. A classifier that uses all available data will perform better than a classifier based on a subset of the data. The hold-out method therefore makes insufficient use of the data. Error estimation using the hold-out method tends to perform poorly, as the classification error estimate will overestimate the unknown generalization error for a classifier built on the full data set. A better alternative for small data sets is k-fold cross-validation (Kohavi, 1995; Goutte, 1997). The classification function is derived from the whole data set while estimating the "true" prediction error of the classifier based on prediction of external data. In cross-validation the data set is randomly split in k subsets (folds) of approximately similar size. One set is used as a test set and the remaining k-1 sets form the training set. This procedure is repeated so that each subset is used as a test set. The cross-validated estimate of accuracy is then based on the average over the k test sets. To improve the stability of the cross-validation, the procedure can be repeated several times, taking new random subsamples and averaging the results. The extreme of k-fold cross-classification is the leave-one-out method (k = sample size n) where each case is used as a single test set and the remaining n-1 cases are used for estimating the rule. However, simulation studies have shown that this method can be unreliable and underestimates the true predictive error. It has been shown that 5-fold to 10-fold cross-validation can work better than leave-one-out, especially if the sets are stratified such that each set is representative of the whole data set (Breiman & Spector, 1992; Kohavi, 1995; Martens & Dardenne, 1998). Cross-validation produces nearly unbiased estimates of accuracy but can still be highly variable, and bootstrap methods have been developed to smooth this variability

(Efron, 1983; Efron & Tibshirani, 1997). The available data are repeatedly sampled with replacement in order to mimic the drawing of future random sampling. Each bootstrap sample is used to estimate classification error rates and the classifier is tested on the remaining non-selected data. Bootstrapping also allows us to estimate confidence intervals for the prediction error (Efron, 2004; Efron & Tibshirani, 1997; Jiang & Simon, 2007; Jiang, Varma, & Simon, 2008).

ERP DATA PRE-PROCESSING: SIGNAL EXTRACTION METHODS

In principle, one might leave the method of resolving the best predictor from the entire raw ERP record to supervised classification methods. However, in practice, it is usually helpful to undertake a preliminary data pre-processing and reduction step to form a set of data features. After removing trials with artefacts, feature extraction techniques including temporal and spatial filters such as bandpass, notch, or Laplace filters to reduce the signal-to-noise ratio (Luck, 2005; de Boer et al., 2007; Fujioka, Mourad, & Trainor, 2011; Hoehl and Wahl, 2012).

In recent years, special interest has been shown in more sophisticated methods such as blind source separation using independent component analysis (Johnson et al., 2001; Mehta, Jerger, Jerger, & Martin 2009; Blankertz et al., 2010; Reynolds and Guy, 2012). Independent component analysis (ICA) decomposes the sensory single trial data derived from several sources into a linear combination of temporally independent components according to modality and the spatial location of the signals (Johnson et al., 2001; Luck, 2005; Michel et al., 2004; Makeig, Jung, Ghahremani, Bell, & Sejnowski, 1997; Makeig et al. 2002; Makeig, Debener, Onton, & Delorme, 2004). These components are assumed to reflect the activity of different brain areas in a trial. Pre-processing using ICA thus facilitates studying intersubject and intertrial variability of task responses. However, ICA methods often need larger sample sizes than are available, and the underlying brain processes that produce both spontaneous and event-related potentials recorded at the different electrodes is still little known in infants. A detailed discussion about ICA is beyond the scope of this article, but for an introduction see Onton, Westerfield, Townsend, and Makeig (2006) or Stone (2004). By identifying a set of candidate response features these methods commonly shorten the **x** data vector. It nonetheless commonly remains long in comparison to the number of subjects.

LINEAR DISCRIMINANT FUNCTION ANALYSIS

We will first introduce linear discriminant function analysis (LDA), which is necessary to understand the extension to regularized discriminant function analysis and support vector machines. Linear discriminant function analysis is still one of the most commonly used methods for discriminating between two or more groups of subjects. It classifies participants into groups, based on a number of measurements or features. If the features within each group are multivariate normally distributed, and the variances and covariances of the feature variables are the same between groups (equal covariance matrix assumption), then discriminant function analysis allows optimal separation between a priori groups of cases (based on a specific criterion, such as “at risk”), and provides the best classification rule of future cases. Unlike many other classification methods, LDA also allows us to assess the contribution of each variable to discrimination, enabling a relatively easy description of the major differences between groups.

There are two main methods of classification in LDA—Fisher's and Mahalanobis'. Fisher's LDA aims to find an optimal linear combination of the features **x**'s that provides best separation of two (or more) classes under the assumption of homogeneity of covariance matrices ($\mathbf{\Sigma}1 = \mathbf{\Sigma}2 = \mathbf{\Sigma}$). The linear function has the form

$$f(x) = w_0 + w_1x_1 + w_2x_2 \ldots + w_px_p$$

where w_i is the weight (or regression coefficient) for variable x_i and w_0 is the constant.

For each case i the linear function f(x) can estimate a discriminant score. The weights are estimated by maximizing the ratio of the between-group variation in discriminant scores, as given by the discriminant function, to the within-group variance.

Given two groups of data, the aim of LDA is to find a linear projection from a multidimensional data set onto a line in a way that the projected cases of the two groups are optimally separated. The classification of a case is based on the discriminant score and its position relative to the midpoint between the two class means of discriminate scores. A case is assigned to the group depending on which side of the midpoint it is. The method by which the linear discriminant function is calculated takes the correlation between the feature variables into account. Specifically, the coefficients of class j are found by taking the inverse of the pooled covariance matrix $\mathbf{\Sigma}$ and multiply it by the vector with the feature variable means of this class. If there are only two groups a case is classified into a class depending whether the discriminant score is above or below 0.

Fisher's LDA can be extended to more than two groups. In the case of j groups a set of j-1 independent or orthogonal discriminant functions will be derived, each an optimal linear combination of the feature variables. The first function provides the most overall discrimination between groups, the second provides second most, and so on. The projections reduce the dimensionality from p dimensions to j-1. It can be shown that in the two group scenario, Fisher's discriminant function approach is mathematically equivalent to the Mahalanobis approach, which calculates the covariance-adjusted distance from a data point to the centroid of each group. A case is assigned to the group with the smallest distance to the centroid. When the data input space is two-dimensional, then the decision boundary would be a line that separates the two groups. Again, only the means of the variables and pooled covariance matrix are needed for calculations.

LDA allows us to assess the contribution of each feature variable to discrimination; variables with larger standardised weights are more important. Another advantage of LDA compared to other classification methods is that it is possible to estimate the probability of being in a particular group. This allows us to include prior probabilities (the probability of an observation coming from a particular group) or the costs of a misclassification in the classification process (McLachlan, 2004).

For known multivariate normally distributed data with different means but the same covariance matrix in all classes, LDA is the optimal classifier (Duda, Hart, & Stork, 2001). An optimal classifier is a function that minimizes the rate of misclassification for new samples drawn from the same population (Kohavi, 1995). LDA performs well if the ratio of sample size to the number of features is large (>20, Stevens, 2009), even if the assumptions of a normal distribution and equal covariances are moderately violated. If the assumption of equal covariance matrices cannot be assumed then quadratic discriminant function analysis (DFA) can be used. The quadratic DFA allows us to discriminant between groups with nonlinear, quadratic boundaries and is therefore

more flexible. However, the cost of this flexibility is that data is easily over-fitted, and thus generalizability to new data is low if sample size is small (Stevens, 2009). A method which reduces the risk of the overfitting of the quadratic DFA, and the possible bias of simple LDA, is the regularized DFA, which we will now describe.

REGULARIZED DISCRIMINANT FUNCTION ANALYSIS

Means and covariances of the population distribution need to be estimated from the data for linear DFA. If the sample size is small compared to the number of variables, the covariance estimates become unstable and highly variable and not all of the parameters are estimable. This situation leads to an overfit of the training sample and hence leads to a poor generalization and high classification error (so called eigenvector bias, see McLachlan, 2004 for details). Furthermore, the covariance matrix may be singular and cannot be inverted (no solution can be obtained) and the robustness to departures from the equal covariance assumption can no longer be assumed. Linear DFA is therefore of limited use in defining a classifier for high-dimensional, small sample sized ERP data sets. The problems are considerably increased for quadratic DFA because variances and covariances need to be estimated for each class separately.

Friedman (1989) proposed the use of regularized DFA for such a circumstance. The aim of regularization is to improve the estimates and to stabilize the covariance matrix estimates. The regularized DFA needs two parameters to be specified, lambda (λ) and gamma (γ), which determine the covariance matrices of the two groups. Lambda controls the regularization (or shrinkage) of the covariance parameter estimates from a unique covariance matrix for each group ($\lambda = 0$, equivalent to quadratic DFA) toward a single common covariance matrix ($\lambda = 1$, equivalent to linear DFA). The second regularization parameter γ ($0 \leq \gamma \leq 1$) controls the shrinkage of the estimated covariance parameters towards a multiple of the identity matrix (a square p $\times$ p matrix which has a 1 for each element on the main diagonal and 0 for all other elements) for a given λ value. This means that the covariances between variables "shrink" toward 0 and the variances toward equal variances. If γ is 1 then there is no correlation between feature variables ("conditional independence"). Shrinkage improves the stability of the covariance matrix and reduces the overfit of the training sample data (reduction of the eigenvalue bias problem, for details see McLachlan, 2004). This is similar to how ridge-regression tackles the problem of colinearity among predictors in ordinary regression. Regularized DFA provides a wide range of regularization alternatives. Table 2 shows how four regularization extremes of λ and γ affect the covariance matrices of two groups.

Optimal shrinkage parameters that jointly maximize the accuracy of future predictions can be derived by grid searches for parameter values that minimize classification error, or through specialised analytical methods (Vidaurre, Schlögl, Cabeza, Scherer, & Pfurtscheller, 2005; Blanckertz et al., 2010). Blankertz et al. (2010) describe in detail how to use a version of regularized DFA in single trial analyses and classification of ERP components. It can be shown that for high dimensional data with small sample sizes, assuming independence between covariates and replacing off-diagonal elements of the sample covariance matrices with zeros (for example by setting λ to 0), often works very well for classification (Pang, Tong, & Zhao, 2009).

Unlike linear DFA, regularized (or quadratic) DFA does not provide a simple general method to assess the impact of the variables on classification.

TABLE 2
Effect of Changing the Regularization Parameter Lambda (λ) and Gamma (γ) in Regularized Discriminant Function Analysis (DFA)

(1) Unregularized (observed) Variance/Covariance matrices:

Matrix 1: $\begin{pmatrix} 2 & 0.5 \\ 0.5 & 1 \end{pmatrix}$ and Matrix 2: $\begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}$

Different regularizations:

(2) $\lambda = 0$ and $\gamma = 0$ (equivalent to quadratic DFA)

Matrix1: $\begin{pmatrix} 2 & 0.5 \\ 0.5 & 1 \end{pmatrix}$ and Matrix 2: $\begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}$

(3) $\lambda = 1$ and $\gamma = 0$ (equivalent to linear DFA)

Matrix1: $\begin{pmatrix} 1.5 & 0.3 \\ 0.3 & 1 \end{pmatrix}$ and Matrix 2: $\begin{pmatrix} 1.5 & 0.3 \\ 0.3 & 1 \end{pmatrix}$

(4) $\lambda = 0$ and $\gamma = 1$ (conditional independence and equal variances withon a group)

Matrix1: $\begin{pmatrix} 1.5 & 0 \\ 0 & 1.5 \end{pmatrix}$ and Matrix 2: $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

(5) $\lambda = 1$ and $\gamma = 1$ (conditional independence and common equal variances)

Matrix1: $\begin{pmatrix} 1.25 & 0 \\ 0 & 1.25 \end{pmatrix}$ and Matrix 2: $\begin{pmatrix} 1.25 & 0 \\ 0 & 1.25 \end{pmatrix}$.

Note. The variances and covariances of measurements of two variables in two groups (1 and 2) are summarized in the two variance/covariance matrices shown at the top of the table (1). The effect of four extreme regularization parameters are shown: If λ and γ are both 0, the individual covariance matrices are used, this is equivalent to quadratic DFA (2). If γ remains 0 and λ is changed to 1, a common, pooled covariance is used and the regularized DFA simplifies to linear DFA (3). If λ remains 0 and γ is set to 1, the covariances between the variables is set to 0 (no correlation, conditional independence) and the variances within each group are equal (4). If both parameters are set to 1, the covariances are 0 and the variances are equal in both groups (5).

LOGISTIC REGRESSION

An alternative to linear DFA is logistic regression, which does not require the independent variables to be multivariate and normally distributed. Variables can be categorical and equal covariance matrices are not required (Tabachnick & Fidell, 2007). However, Michie, Spiegelhalter, and Taylor (1994) found little practical differences between linear DFA and logistic regression. Similar to regularized DFA, a regularized (or penalized) logistic regression exists for high-dimensional data (Harrell, 2001; Dettling & Bühlmann, 2004).

SUPPORT VECTOR MACHINES

Support vector machine (SVM) methods are a computer-intensive classification method introduced by Boser, Guyon, and Vapnik (1992). Widely used in bioinformatics for the automatic classification of microarray gene expression profiles, they can deal with high-dimensional data sets and are able to model diverse sources of data with complex covariance structures. A non-technical introduction is given by Noble (2006).

SVM performs classification by constructing a (p-1)-dimensional hyperplane that optimally separates the data into two categories. For example in the simple case we have two groups with measurements on two features. Data points can be plotted on a 2-dimensional plane. In this case SVM analysis attempts to find a line (a 1-dimensional hyperplane) that optimally separates the two groups. However, there are an infinite number of possible lines (hyperplanes), and the question is which hyperplane is the optimal choice (Figure 1).

SVM solves this problem by choosing the hyperplane that is maximally far away from any data point and thus maximizes the margin (distance between the closest point and the hyperplane, see Figure 2). This particular hyperplane maximizes the prediction accuracy of the classification of previously unseen cases. Each data point can be seen as a p-dimensional vector between the origin (0,0) and that point. Vectors that constrain the width of the margin are called support vectors, and give the method its name. Other data points are not involved in determining the decision boundary that is selected.

If measurements are available for three features, data can be plotted in a three-dimensional cube and a plane would separate data points into in two groups. If n features (or attributes as they are often called) are available, data can be plotted in an n-dimensional space and can be separated by a p-1 dimensional hyperplane. However, linear planes cannot always separate the data into two groups. In this case SVM applies kernel methods to model nonlinear relationships by projecting the input data into a high-dimensional feature space, where an n-1 dimensional linear hyperplane

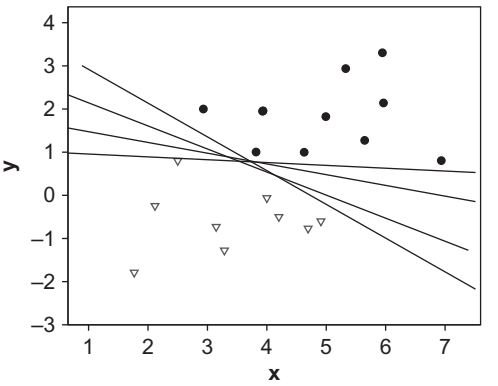

FIGURE 1 Hypothetical scatterplot for two feature variables, x and y, measured on cases of two groups (filled circles and open triangles). There is an indefinite number of possible lines (one dimensional hyperplanes) that could perfectly separate the two groups. Four possible lines the separating cases of the two groups are shown.

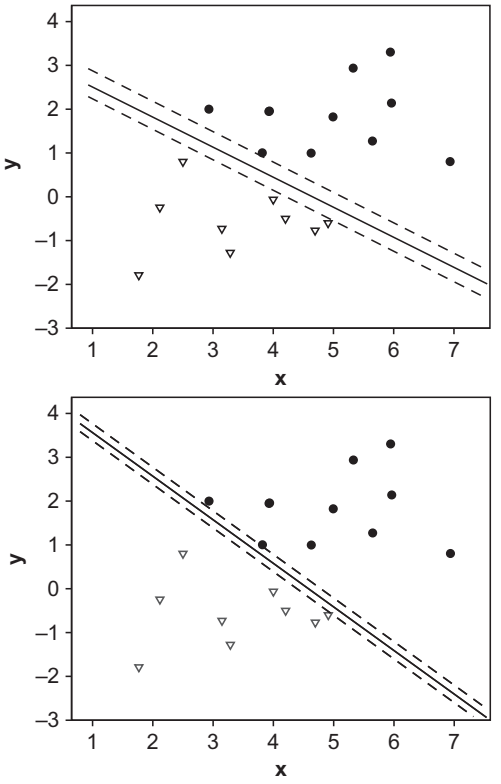

FIGURE 2 Maximum margin: Support vector machine (SVM) chooses the hyperplane (solid line), which maximizes the margin (distance between closest point and the hyperplane). The dashed lines parallel to the hyperplane represent the margin and the data (vectors) that constrain the width of the margins are the support vectors. The margin of the figure on the top is wider than compared to the one on the bottom and the hyperplane on the top is the preferred one.

can be selected to separate the two groups. This so-called kernel trick of SVM models allows separations to be performed even with complex boundaries (Schölkopf, Smola, & Müller, 1998). Commonly used kernels are polynomial and radial basis functions. By using kernel functions it is often possible to find a hyperplane that perfectly separates the cases into two non-overlapping groups. However, such perfect separation may result in a model that does not generalize well to other data, and classification error of unseen data would be large. To avoid overfitting (or in case the two groups cannot be separated even with kernels), the SVM algorithm is modified by allowing some cases to fall on the "wrong side" of the hyperplane. This soft margin permits some misclassifications in the training set. The larger the soft margin, the more classification errors are allowed and the better the generalizability of the model. A cost parameter C, which controls the soft margin, needs to be specified. Larger values of C decrease the soft margin and reduce the number of misclassifications in the training set, but the cost is that it may not generalize well—it

may perform less well with an entirely new sample. Therefore an optimal value for C needs to be found, which minimizes misclassification in a new sample (e.g., the k-fold cross-classification). If kernels are used, optimal parameters for the kernel functions, to determine the complexity of the boundary, also need to be selected.

The choice of the parameters of the soft margin and the kernel functions is a trade-off between accuracy of the model and minimizing the generalization error. Computer intensive grid searches are used to find an optimal combination of parameters, and external methods such as k-fold cross-validation are used during the search process of the training set to select the optimal parameters. Similar to regularized DFA, simple general methods to assess the impact of variables on classification are not available and SVM does not provide a (reliable) predictive probabilistic classification (e.g., the probability a case belonging to a group) but only a threshold value. For an accessible introduction to support vector machines, see Hamel (2009) and Bishop (2006) for pattern recognition and machine learning in general.

AN APPLICATION: ERP CORRELATES OF EYE GAZE PROCESSING IN INFANTS AT-RISK FOR AUTISM

Background

To illustrate the performance of the proposed methods we applied them to data recently collected as part of the British Autism Study of Infant Siblings (BASIS; Elsabbagh et al., 2009). This study is among the first attempts to investigate brain development in infants at-risk for autism. A number of studies reveal very early differences in neural processing in infants at-risk prior to the age at which overt behavioral symptoms of autism emerge in a subgroup of these infants. Large-scale studies are now in progress to investigate whether these early differences in brain development could offer useful predictors of clinical diagnosis, therefore allowing for the possibility of early screening and intervention. Moreover, understanding early brain development in this group could explain variability of outcomes in infants at-risk when they reach toddlerhood (see Elsabbagh & Johnson, 2010 for a review). As such, advancing techniques for analysis of ERP data is critical for both basic and translational science.

In their first initial study, Elsabbagh et al. (2009) investigated ERP correlates of eye gaze processing in infants aged approximately 10 months old. This is because atypical eye gaze processing is one of the key features of individuals with autism. EEG was recorded using a 64-channel geodesic net while infants sat on their parent's lap and watched a monitor. On each trial a static colourful fixation stimulus was presented followed by a colour image of a female face, with gaze directed either toward ("direct gaze") or away ("averted gaze") from the infant.

After artefact rejection, ERPs recorded from four channel groups selected on the basis of previous research were averaged for each infant. The channel groups were a posterior group, a right temporal group, a left temporal group, and an anterior central group. Mean amplitude and latency of three gaze-sensitive components in infancy (P1: 100–199 msec, N290: 200–319 msec, and P400: 320–539 msec) were averaged for each of the channel groups.

The analysis of Elsabbagh et al. (2009) revealed significant differences between the two groups in the latency of the averaged occipital P400 ERP measure in response to direct gaze. No significant differences between the groups were detected in response to averted gaze. The focus of

the present study is the demonstration of statistical methods for the analysis of ERP data in general. It is not our intention to provide any new findings about the neural correlates of eye gaze processing.

Methods

In this current analysis, 18 averaged measurements per experimental condition were used to assess the usefulness of three supervised classification methods in classifying infants: classical discriminant function analysis, regularized discriminant function analysis (Friedman, 1987) and support vector machines (Vapnik, 1995). The data are averaged ERP signals and the number of feature variables relative to the sample size is, with a ratio of 1, relatively large. The methodological task is how best to classify infants as to their risk group (at-risk group versus control group).

Statistical Analysis

The infant ERPs were classified using discriminant function analysis, regularized function analysis, and support vector machines. The regularized DFA needs two parameters to be specified, lambda and gamma, which influence the estimated covariance matrices of the two groups. For the support vector machine the cost parameter C, which controls the penalty for misclassifying a training point and thus the complexity of the classification function, needs to be specified: A larger value of C will result in smaller soft-margins and fewer misclassified cases. This will, therefore, create a more complex function at the cost of lower generalizability for future predictions. Linear SVM classifiers do not need any further parameters to be set, while a nonlinear SVM needs additional parameters to be specified depending on the kernel type. In this instance, a nonlinear kernel did not improve the prediction error of the classifier and results, and so we present only the linear model findings.

Parameter selection of the regularized DFA and SVM were done manually using a grid search, and selection was based on minimizing the prediction error of 6-fold cross-fold classification. The classification techniques were first applied to the full data set and then for each experimental condition (direct and averted gaze) separately.

The results of a classification procedure is presented as a 2×2 confusion matrix, which displays the number of correct and incorrect predictions made by models compared with the actual classifications in the test data (see the result Table 3). The rows of the matrix present the number of instances in a predicted group while the columns present the instances in an actual group. The estimated prediction error of the classifier is calculated as number of correctly classified cases divided by the total number of cases. In the case of unequal group sizes prediction error should not be compared to chance classification (50%) but to a majority rule classifier ($19/36 = 52.7\%$). The *sensitivity* of the classifier refers to how many infants with an autistic sibling can be detected (correctly classified siblings/[correctly classified siblings + control cases classified as siblings]), while *specificity* refers to the proportion of control infants that are correctly identified (number of correctly classified control infants divided by the sum of true correctly and falsely classified control infants). A classifier with a high specificity has a low type I error rate while a classifier with a high sensitivity has a low type II error rate. In addition, the positive predictive value (PPV), the

TABLE 3
Classification Results of Discriminant Function Analysis, Regularized Discriminant Function Analysis, and Support Vector Machine Based on Latencies and Amplitudes of Event-Related Potentials of (a) Complete Data Set, (b) Direct Gaze, and (c) Averted Gaze Condition

(a) Complete Data Set

		TRUE			
	Groups	*Sibling*	*Control*	*Accuracy*	
Discriminant Function Analysis					
Predicted	Sibling	9	6	0.60	Sensitivity
	Control	10	11	0.52	Specificity
	Accuracy	0.47	0.35	0.56	
		PPV	NPV	Overall	
Regularized Discriminant Function Analysis					
Predicted	Sibling	11	6	0.65	Sensitivity
	Control	8	11	0.58	Specificity
	Accuracy	0.58	0.35	0.61	
		PPV	NPV	Overall	
Support Vector Machine					
Predicted	Sibling	14	8	0.64	Sensitivity
	Control	5	9	0.64	Specificity
	Accuracy	0.74	0.47	0.64	
		PPV	NPV	Overall	

(b) Direct Gaze

		TRUE			
	Groups	*Sibling*	*Control*	*Accuracy*	
Discriminant Function Analysis					
Predicted	Sibling	11	7	0.61	Sensitivity
	Control	8	10	0.56	Specificity
	Accuracy	0.58	0.41	0.58	
		PPV	NPV	Overall	
Regularized Discriminant Function Analysis					
Predicted	Sibling	12	5	0.71	Sensitivity
	Control	7	12	0.63	Specificity
	Accuracy	0.63	0.29	0.67	
		PPV	NPV	Overall	
Support Vector Machine					
Predicted	Sibling	13	6	0.68	Sensitivity
	Control	6	11	0.65	Specificity
	Accuracy	0.68	0.35	0.67	
		PPV	NPV	Overall	

Note. PPV = Positive predictive power; NPV = negative predictive power; Overall = overall correct classification (accuracy).

(*Continued*)

TABLE 3
(Continued)

(c) Averted Gaze					
		TRUE			
	Groups	*Sibling*	*Control*	*Accuracy*	
Discriminant Function Analysis					
Predicted	Sibling	9	10	0.47	Sensitivity
	Control	10	7	0.41	Specificity
	Accuracy	0.47	0.59	0.44	
		PPV	NPV	Overall	
Regularized Discriminant Function Analysis					
Predicted	Sibling	9	10	0.47	Sensitivity
	Control	10	7	0.41	Specificity
	Accuracy	0.47	0.59	0.44	
		PPV	NPV	Overall	
Support Vector Machine					
Predicted	Sibling	8	11	0.42	Sensitivity
	Control	11	6	0.35	Specificity
	Accuracy	0.42	0.65	0.39	
		PPV	NPV	Overall	

proportion of correctly classified at-risk infants, and negative predictive value (NPV), the proportion of correctly classified control infants, are presented. Finally, we present average % of correct classification based on 100 6-fold cross-classification using nested cross-validation for SVM; that is, the cost parameter C is estimated within each fold.

Analyses were done using R 2.12 (R Development Core Team, 2011). Linear discriminant function analysis was done using the function *lda* of the library *MASS* (Venables & Ripley, 2002). The function *rda* of the library *klaR* (Weihs, Ligges, Luebke, & Raabe, 2005) was used for regularized discriminant function analysis. For SVM classification the function *svm* of the library *e1071* (Dimitriadou, Hornik, Leisch, Meyer, & Weingessel, 2010) was used. *Svm* provides a rigid interface to the software *libsvm* (Chang & Lin, 2010).

RESULTS

Analysis of Direct and Averted Gaze Conditions Together

Of the three classification methods using all available data, SVM and regularized DFA gave classifiers that predicted similarly well (64% and 61%, respectively) while the classification accuracy of the LDA was 56%, just slightly above the majority classifier of 52.7% (Table 3a). The average of 100 cross-validations showed similar results (LDA: 54.6%, regularized DFA: 61.9 and SVM: 60.9% correct classification). The sensitivity was similar for both regularized DFA and SVM procedures (65% and 64%) while specificity was only slightly better for the classifier derived from SVM (64% and 58% for regularized DFA). Sensitivity and specificity were lowest for the standard LFA (60% and 52%). The regularization parameters for the regularized DFA were $\lambda = 0$ and

$\gamma = 1$, which means that the covariances were set to 0 and variances were equal within each group.

Analysis of Direct and Averted Gaze Conditions Separately

The error prediction accuracy improved slightly for all three classification methods if only the ERP recordings of the direct gaze trial were used (DFA: 58%, regularized DFA: and SVM: 67%, Table 3b) while it reduced below 50%, and therefore below chance level, for all three procedures if the data of the averted gaze were analyzed alone (Table 3c).

Standard Logistic Regression Analysis

For comparison, we compare our results with a model based directly on previous findings, estimating a logistic regression with group as the dependent variable and the latency of the averaged occipital P400 ERP response. The prediction error for the model was estimated using the same 6-fold classification method, but because the *selection* of this variable was based on group comparisons of the whole data set (see Elsabbagh et al., 2009) and not on 6-fold cross-validation, the classification error will be overestimated and cannot be directly compared with the prediction error of the other classification methods. Infants with longer latencies to direct gaze were more likely to be classified as an at risk sibling (OR = 1.03 (95% C.I. 1.007–1.058, $p = .01$). The overall cross-validated classification accuracy was 61% and better than a majority classifier of 52.7% (Table 4).

DISCUSSION

SVM and regularized DFA were able to discriminate between infants at risk for autism and control infants with above chance classification using 6-fold cross-fold validation. Unlike standard

TABLE 4
Classification Results of Logistic Regression Analysis With Occipital P400 Latency Event-Related Potential Response as Dependent Variable and Group as Predictor Variable During Direct Gaze Condition

		Logistic Regression			
		TRUE			
	Groups	*Sibling*	*Control*	*Accuracy*	
Predicted	Sibling	12	7	0.63	Sensitivity
	Control	7	10	0.59	Specificity
	Accuracy	0.63	0.41	0.61	
		PPV	NPV	Overall	

Note. PPV = Positive predictive power; NPV = negative predictive power; Overall = overall correct classification (accuracy).

analyses, which commonly compare a series of means between groups, there is no multiple testing problem because the models were assessed on data which were not used for the assessment of the accuracy. In this illustration, the cross-validation performance of the logistic regression is similar to regularized DFA and SVM analyses but is likely to be overestimated because the occipital P400 latency variable was selected based on the results of several group comparisons of the whole data set. When assessing reports claiming prognostic discrimination, care needs to be taken to check that all aspects of multiple-testing bias have been removed from the reported estimates. Both regularized DFA and SVM methods can be relatively easily performed using libraries from the open source software R (R Development Core Team, 2011) and are therefore a promising tool for discrimination and prediction tasks for ERP data sets, where the sample size is small in comparison to the number of feature variables.

The analysis also confirmed the result that the infants could only be discriminated in the direct gaze condition, but not in the averted gaze condition. From a methodological point of view the results suggest that feature selection is still important because the accuracy of the classifier was somewhat better if only the direct gaze dataset was used, the condition for which more differences between the two groups were expected for theoretical reasons.

A main disadvantage of regularized DFA and SVM is that they do not provide information about the importance of each variable, and consequently do not allow feature variable selection for the training of the classifier. In classification problems, an important task is often to identify temporal and spatial subsets of ERP components that best discriminate the classes in order to understand the underlying mechanism of class differences. Graphical methods can be used to evaluate and explain the obtained results (Poulet, 2004), but these methods are only useful if the number of feature variables is not too large. In high-dimensional data sets with many irrelevant, unreliable and often correlated variables, including all variables may undermine the success of the machine learning process. In ERP recordings we can assume highly correlated measurements due to spatial and temporal dependencies if single trials are used. Furthermore, ERP response to experimental stimuli is expected to be restricted to specific channel groups, and therefore only some features of the ERP signal are expected to contain reliable information for discrimination. It is therefore advisable to eliminate irrelevant feature variables prior to classification analysis in order to increase predictive accuracy and enhance the generalization performance of the classification learning algorithm (Rakotomamonjy, 2003). Zuber and Strimmer (2009) and Ahdesmäki and Strimmer (2010) developed a feature selection method for regularized DFA of high-dimensional data, which ranks correlation adjusted features according to their importance on the group means (centroids). They further proposed to select features based on a False-non-discovery rate. In order to avoid over-optimistic prediction error estimates, feature selection process and regularized discriminant analysis need to be embedded within the cross-validation process. A second, computationally more expensive feature selection approach was proposed by Guo, Hastie, and Tibshirani (2007) and Guo (2010). They used a regularized discriminant function analysis with a so-called shrunken centroid estimator which performs classification and feature selection simultaneously. The "shrunken centroids regularized discriminant analysis" (SCRDA) effectively removes most non-contributing variables for future prediction. Finally, Tai and Pan (2007) suggested the use of a regularized DFA with shrinkage, which incorporates prior knowledge of the variables and considers known group relationships among variables. All three methods improved the classification error rate of high-dimensional gene expression data. A regularization technique called the elastic net, which simultaneously does automatic variable

selection for support vector machines, was introduced by Zou and Hastie (2005) and seems to perform similarly well.

A preliminary analysis using the feature ranking method and R library "*sda*" of Ahdesmäki and Strimmer (2009, 2010) based on correlation-adjusted *t*-scores derived from the regularized covariance matrix revealed that the prolonged latency of the occipital P400 ERP component in response to direct gaze was ranked top and the decorrelated *t*-score of 3.3 was the only score greater than 2. This confirms the result of Elsabbagh et al. (2009) who also reported that the P400 component differentiated the two groups, with the sib-ASD group showing prolonged latency in responding to direct gaze. This P400 component is known to be sensitive to face processing in infants (Halit, de Haan, & Johnson, 2003). Elsabbagh et al. (2009) suggested that the broader autism phenotype, including an atypical response to eye gaze, may be detectable early in infancy. They discussed that the neural patterns associated with later attentional modulation and those sensitive to referential aspects of eye gaze are comparable to those reported in children and adults diagnosed with autism. The observed early atypical response in the latency of the occipital P400 ERP component to eye gaze is likely to combine with other risk factors, resulting in a later diagnosis for some individuals (Elsabbagh & Johnson, 2007).

Feature ranking and selection is therefore not only useful in improving prediction, but provides a shortlist of feature variables relevant for discriminating the classes of interest that can be an important tool in explaining the mechanism of the underlying discriminatory process between the two infant groups. Applying machine learning methods with feature selection on EEG/ERP data does not only offer better possibilities in the early detection of abnormalities in EEG signals and the development of a biomarker for early detection of developmental cognitive disorders but also it can improve our understanding of the neurodevelopmental pathway of autism and how intervening early might make a difference.

A weakness of SVM compared with regularized DFA is that, given an infant sample, the former only predicts group membership, but does not provide a reliable estimate of probabilistic classification for an individual infant, only a threshold value. A probabilistic classification does not only deliver estimates of class membership probabilities distinguishing between at-risk and control infants, but also differentiates among those at-risk siblings those who go on to show clear symptoms and possible autism diagnosis and those who do not. Such a probabilistic classification would enable a separation of infants that can be confidently classified into autism risk groups from more ambiguous cases, thereby allowing separate classes of likely heterogeneous groups of infants who will receive a clinical diagnosis. These probabilities should prove to be of value for translational research testing where similar measures can be used in clinical practice. Regularized DFA has the advantage of additionally providing an estimate of the underlying probability. A further method that provides probabilities are Gaussian process models (Rassmussen & Williams, 2006).

Finally, both methods are based on regularization of the covariance matrix and suffer from the problem of how to set penalty parameters to avoid overfitting, another important element for establishing clinical utility. External methods, such as k-fold cross-validation during the search process to select the optimal parameters, need to be applied and these computer intensive methods may not be time efficient with larger datasets. In this study, the sample size in both groups was very similar. However, it is common that data sets are—often severely—unbalanced and the construction of the classifier may be undermined and the estimation of the accuracy subverted, for example by classifying all cases to the larger class (Malley, Malley, & Pajevic, 2011). A simple

solution to reduce the problem for regularized DFA is to change the threshold of classification and to choose a threshold to obtain an appropriate balance between the proportions misclassified from each of the classes (see Weiss, 2008, for an overview).

A possible alternative to commonly used discrimination and machine learning techniques, which overcomes the described problems, may be Gaussian process (GP) models (Rassmussen & Williams, 2006). Marquand et al. (2010) successfully applied GP classification procedures to fMRI decoding and emphasized the advantage of their predictive capability and their ability to capture information encoded by spatially correlated voxels. GP models have been shown to perform competitively when applied to EEG signals for brain computer interfacing studies (Zhong, Lotte, Girolami, & Lécuyer, 2008; Wang, Wan, Mak, Mak, & Vai, 2009). However, their application to either averaged or single trial ERP studies has not been investigated.

Single trial ERP data are not commonly used for the analysis of infant ERP studies (but see, for example, Bishop and Hardiman, 2010, who analyzed single trial ERPs with older children). However, the rather narrow set of features that are commonly used in the study of averaged ERP responses of infants may present not only a very restricted view of the process of the development of the infant brain, but may be inappropriate for identifying discrepancies that provide early indicators of atypical development. For example, instead of the amplitude and latency, it may be the variability across cycles or spatial concentration, or the dispersion across locations, that may be developmentally salient and prognostic. However, candidate responses are almost countless and standard statistical methods do not allow us to preserve details of waveforms arising from single stimulus presentations.

Such complex data are high-dimensional, that is, the number of variables (or features) easily exceeds the number of infants by many times over. In recent years machine learning methods were advocated for such high-dimensional data where $N>>n$, such as in bioinformatics (Larrañaga et al., 2006; Inza et al., 2010), fMRI studies (Mourao-Miranda, Bokde A., Born, Hampel, & Stetter, 2005; Pereira, Mitchell, & Botvinick, 2009; Ecker et al., 2010), and more recently EEG studies (Dornhege, Krauledat, Müller, & Blankertz, 2007; Müller et al., 2008; Khodayari-Rostamabad, Hasey, Maccrimmon, Reilly, & de Bruin, 2010; Bosl, Tiemey, Tager-Flusberg, & Nelson, 2011). Recently, methods were developed for single trial data from adult ERP studies using mainly discriminant function and machine learning methods (Machine learning: Yamagishi, Tsubone, & Wada, 2008; Rakotomamonjy & Guigue, 2008; Salvaris & Sepalved, 2009; Müller, Candrian, Kropotov, Ponomarev, & Baschera, 2010; Discriminant function analysis: Bandt, Weymar, Samaga, & Hamm, 2009; Blankertz et al, 2010). Applying machine learning and feature variable selection methods to single trial ERP data would allow the assessment of a large number of signals, including possible interactions and correlations between wavelet curves, and could shed new insight on the neural underpinning of cognitive processes. Further research is desirable to assess the potential of machine learning methods using single trial ERPs in infants.

Although classification of infants at-risk and future prediction of their clinical outcomes is the main aim of the study considered here, the methods have wider applications. The methods can potentially improve standard methods of ERP analysis in experimental studies investigating various cognitive processes. For example, Stahl et al. (2010) studied emotional face processing. In this experimental study, infants were presented with a female face showing a neutral or angry facial expression, using either direct or averted gaze. Instead of using standard random effects model analyses, the difference in ERP signals of an infant between an angry and neutral face could be calculated for each gaze condition, and analyzed using DFA or SVM. An above chance

classification would reveal an interaction between emotional expression and gazing direction, which could be further analyzed by feature ranking or additional DFA or SVM analyses. The simultaneous analyses of many feature variables using regularized DFA or other machine learning methods may allow us to increase the power of such studies, and research in the applicability of DFA and machine learning methods in this area would be desirable.

The same methods can also be used in the context of clinical trials. For example, Feighner and Sverdlov (2001) analyzed a randomized clinical trial with a new antidepressant using linear DFA to separate drug-treated from placebo populations. They used the change from baseline for 21 items of the Hamilton Depression scale at 11 time points for their analysis and successfully separated the treatment groups. The methods can be applied to group comparisons of experimental tasks aiming to study cognitive processes.

In this article we provide some supportive evidence that classification methods such as regularized DFA or SVM can increase the discriminative power of ERP measurements. Cross-validation or other methods in assessing the accuracy of the classification avoid or at least reduce the problem of multiple testing that frequently occurs by using a long series of univariate group comparisons. The analysis methods have wide applications: distinguishing risk or clinical groups from control groups, predicting subgroups who will receive a diagnosis within an at-risk group, or distinguishing treatment groups from placebo. Moreover, the same methods can potentially improve standard ERP analysis of experimental conditions in typical populations. The methods have been successfully applied to discriminate between groups using single trial ERPs in adults and further research is needed to assess the potential of machine learning methods using either averaged or single trial ERPs in infants.

REFERENCES

Ahdesmäki, M., & Strimmer, K. (2009). sda: Shrinkage discriminant analysis and feature selection. R package version 1.1.0. Retrieved from http://CRAN.R-project.org/package=sda

Ahdesmäki, M., & Strimmer, K. (2010). Feature selection in omics prediction problems using cat scores and false non-discovery rate control. *Annals of Applied Statistics*, *4*, 503–519.

Bandt, C., Weymar, M., Samaga, D., & Hamm, A. O. (2009). A simple classification tool for single-trial analysis of ERP components. *Psychophysiology*, *46*(4), 747–757.

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

Bishop, D. V. M., & Hardiman, M. J. (2010). Measurement of mismatch negativity in individuals: A study using single-trial analysis. *Psychophysiology*, *47*, 697–705.

Blankertz, B., Lemm, S., Treder, M. Haufe, S., & Müller, K.-R. (2010). Single-trial analysis and classification of ERP compoents—A tutorial. *NeuroImage*. doi:10.1016/j.neuroimage.2010.06.048

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on computational learning theory (COLT′92)* (pp. 144–152). New York, NY: ACM.

Bosl, W., Tiemey, A., Tager-Flusberg, H., & Nelson, C. (2011). EEG complexity as a biomarker for autism spectrum disorder risk. *BMC Medicine*, *9*(18). doi:10.1186/1741-7015-9-18

Bousquet, O., Chapelle, O., &. Hein, M. (2004). Measure based regularization. In S. Thrun, L. Saul, & B. Schoelkopf (Eds.), *Advances in neural information processing systems 16 (NIPS 2003)* (pp. 1–8). Cambridge, MA: MIT Press.

Breiman, L., & Spector, P. (1992). Submodel selection and evaluation in regression: The X-random case. *International Statistical Review*, *60*, 291–319.

Caragea, D., Cook, D., Wickhamand, H., & Honavar, V. (2008). Visual methods for examining SVM classifiers. In S. J. Simoff, M. H. Bohlen, & A. Mazeika (Eds.), *Visual data mining: Theory, techniques and tools for visual analytics* (pp. 136–153). New York, NY: Springer.

Chang, C.-C., & Lin, C.-J. (2010). LIBSVM: A library for Support Vector Machines. Version 3.0. Retrieved from http://www.csie.ntu.edu.tw/~cjlin/libsvm

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge, England: Cambridge University Press.

Das, K., Giesbrecht, B., & Eckstein, M. P. (2010). Predicting variations of perceptual performance across individuals from neural activity using pattern classifiers. *Neuroimage*, *51*(4), 1425–1437.

de Boer, T., Scott, L. S., & Nelson, C. A. (2007). Methods for acquiring and analysing infant event-related potentials. In M. de Haan (Ed.), *Infant EEG and event-related potentials* (pp. 5–37). Hove, England: Psychology Press.

de Haan, M. (2007). *Infant EEG and event-related potentials*. Hove, England: Psychology Press.

Dettling, M., & Bühlmann, P. (2004). Finding predictive gene groups from microarray data. *Journal of Multivariate Analysis*, *90*, 106–131.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., & Weingessel, A. (2010). e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.5-24. http://CRAN.R-project.org/package=e1071

Dornhege, G., Krauledat, M., Müller, K.-R., & Blankertz, B. (2007). General signal processing and machine learning tools for BCI. In G. Dornhege, J. Millán, T. Hinterberger, D. J. McFarland, & K.-R. Müller (Eds.), *Towards brain–computer interfacing* (pp. 207–233). Cambridge, MA: MIT Press.

Duda, R. O., P. E. Hart, & D. G. Stork (2001). *Pattern classification*, *2nd ed*. New York, NY: Wiley-Interscience.

Doyle, O. M., Temko, A., Lightbody, G., Marnane, W., & Boylan, G. B. (2010). Heart rate based automatic seizure detection in the newborn. *Medical Engineering & Physics*, *32*(8), 829–839.

Doyle, O. M., Temko, A., Murray, D. M., Marname, W., Lightbody, G., & Boylan, G. B. (submitted) Combination of multimodel data for the prediction of neurodevelopmental outcome in newborns with hypoxic-ischemic encephalopathy.

Ecker, C., Rocha-Rego, V., Johnston, P., Mourao-Miranda, J., Marquand, A., Daly, E. M., . . . Murphy, D. G. (2009). Investigating the predictive value of whole-brain structural MR scans in autism: A pattern classification approach. *Neuroimage*, *49*, 44–56.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78316–78331.

Efron, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99619–642.

Efron, B., & Tibshirani, R. J. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association, 92*, 92548–92560.

Elsabbagh, M., & Johnson, M. H. (2007). Infancy and autism: Progress, prospects, and challenge. *Prog Brain Research*, *164*, 355–383.

Elsabbagh, M., Volein, A., Csibra, G., Holmboe, K., Garwood, H., Tucker, L., . . . Johnson, M. H. (2009). Neural correlates of eye gaze processing in the infant broader autism phenotype. *Biological Psychiatry*, *65*, 31–38.

Fabiani, M., Gratton, G., & Federmeier, K. D. (2007). Event-related brain potentials: Methods, theory, and applications. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology*, *3rd ed*. (pp. 55–83). Cambridge, England: Cambridge University Press.

Feighner, J. P., & Sverdlov, L. (2001). The use of discriminant analysis to separate a study population by treatment subgroups in a clinical trial with a new pentapeptide antidepressant. *Journal of Applied Research in Clinical and Experimental Therapeutics*, *2*(1), 50–57.

Fielding, A. H. (2007). *Cluster and classification techniques for the biosciences*. Cambridge, England: Cambridge University Press.

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, *84*, 165–175.

Fujioka, T., Mourad, N., & Trainor, L. J. (2011). Development of auditory-specific brain rhythm in infants. *European Journal of Neuroscience*, *33*, 521–529.

Gazzaniga, M. S. (2004). *The cognitive neurosciences III*. Cambridge, MA: MIT Press.

Goutte, C. (1997). Note on free lunches and cross-validation. *Neural Computation*, *9*, 1211–1215.

Guo, J. (2010). Simultaneous variable selection and class fusion for high-dimensional linear discriminant analysis. *Biostatistics*, *11*(4), 599–608.

Guo, Y., Hastie, T., & Tibshirani, R. (2007). Regularized linear discriminant analysis and its applications in microarrays. *Biostatistics*, *8*(1), 86–100.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Halit, H., de Haan, M., & Johnson, M. H. (2003) Cortical specialisation for face processing: Face-sensitive event-related potential components in 3- and 12-month-old infants. *Neuroimage*, *19*, 1180–1193.

Hamel, L. H. (2009). *Knowledge discovery with support vector machines*. Hoboken, NJ: Wiley

Handy, T. C. (2005). *Event-related potentials: A methods handbook*. Cambridge, MA: The MIT Press.

Harrell, F. E. (2001). *Regression modeling strategies with applications to linear models, logistic regression, and survival analysis*. New York, NY: Springer.

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

Hoehl, S., & Wahl, S. (2012). Recording infant ERP data for cognitive research. *Developmental Neuropsychology*, *37*, 187–209.

Inza, P., Calvo, B., Armañanzas, R., Bengoetxea, E., Larrañaga, P., & Lozano, J. A. (2010). Machine learning: An indispensable tool in bioinformatics. In R. Matthiesen (Ed.), *Bioinformatics methods in clinical research* (pp. 25–48). New York, NY: Springer.

Jiang, W., & Simon, R. (2007). A comparison of bootstrap methods and an adjusted bootstrap approach for estimating the prediction error in microarray classification. *Statistics in Medicine*, *26*, 5320–5334.

Jiang, W., Varma, S., & Simon, R. (2008). Calculating confidence intervals for prediction error in microarray classification using resampling. *Statistical Applications in Genetics and Molecular Biology*, *7*(1), 1–19.

Johnson, M. H., de Haan, M., Oliver, A., Smith, W., Hatzakis, H., Tucker, L. A., & Csibra, G. (2001). Recording and analyzing high density ERPs with infants using the Geodesic Sensor Net. *Developmental Neuropsychology*, *19*, 295–323.

Khodayari-Rostamabad A., Hasey G. M., Maccrimmon D. J., Reilly J. P., & de Bruin, H. (2010). A pilot study to determine whether machine learning methodologies using pre-treatment electroencephalography can predict the symptomatic response to clozapine therapy. *Clinical Neurophysiology*, *121*(12), 1998–2006.

Knapp, M., Romeo, R., & Beecham, J. (2009). The economic cost of autism in the UK. *Autism*, *13*, 317–336.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, *2*(12), 1137–1143.

Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., . . . Robles, V. (2006). Machine learning in bioinformatics. *Briefings in Bioinformatics*, *7*(1), 86–112.

Lazzaro, I., Anderson, J., Gordon, E., Clarke, S., Leong, J., & Meares, R. (1997). Single trial variability within the P300 (250–500 ms) processing window in adolescents with attention deficit hyperactivity disorder. *Psychiatry Research*, *73*(1–2), 91–101.

Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., & Arnaldi, B. (2007). A review of classification algorithms for EEG-based brain-computer interfaces. *Journal of Neural Engineering*, *4*, R1–R13. doi: 10.1088/1741-2560/4/2/R01

Luck, S. J. (2005). *An introduction to the event-related potential technique*. Cambridge, MA: The MIT Press.

Makeig, S., Debener, S., Onton, J., & Delorme, A. (2004). Mining event-related brain dynamics. *Trends in Cognitive Science*, *8*(5), 204–210.

Makeig, S., Jung, T.-P., Ghahremani, D., Bell, A. J., & Sejnowski, T. J. (1997). Blind separation of auditory event-related brain responses into independent components. *Proceedings of the National Academy of Sciences*, *94*, 10979–10984.

Makeig, S., Jung, T.-P., Ghahremani, D., & Sejnowski, T. J. (2000). Independent component analysis of simulated ERP data. In T. Nakada (Ed.), *Integrated human brain science* (pp. 1–24). New York, NY: Elsevier.

Makeig, S., Westerfield, M., Jung, T.-P., Enghoff, S., Townsend, J., Courchesne, E., & Sejnowki, T. J. (2002). Dynamic brain sources of visual evoked responses. *Science*, *295*, 690–694.

Malley, J. D., Malley, K. G., & Pajevic, S. (2011). *Statistical learning for biomedical data*. Cambridge, England: Cambridge University Press.

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

Martens, H. A., & Dardenne, P. (1998). Validation and verification of regression in small data sets. *Chemometrics and Intelligent Laboratory Systems*, *44*, 99–121.

McLachlan, G. J. (2004). *Discriminant analysis and statistical pattern recognition*. Hoboken, NJ: John Wiley & Sons.

Mehta, J., Jerger, S., Jerger, J., & Martin, J. (2009). Electrophysiological correlates of word comprehension: event-related potential (ERP) and independent component analysis (ICA). *International Journal of Audiology*, *48*(1), 1–11.

Michel, C. M., Murray, M. M., Lantz, G., Gonzalez, S., Spinelli, L., & Grave de Peralta, R. (2004). EEG source imaging. *Clinical Neurophysiology*, *115*(10), 2195–2222.

Michie, D., Spiegelhalter, D., & Taylor, C. (1994). *Machine learning, neural and statistical classification*. Upper Saddle River, NJ: Ellis Horwood.

Molfese, D. L., Molfese, V. J., & Kelly, S. (2001). The use of brain electrophysiology techniques to study language: a basic guide for the beginning consumer of electrophysiology information. *Learning Disability*, *24*, 177–188.

Molinaro, A. M., Simon, R., & Pfeiffer, R. M. (2005). Prediction error estimation: A comparison of resampling methods. *Bioinformatics*, *21*(15), 3301–3307.

Mourao-Miranda, J., Bokde, A. L. W., Born, C., Hampel, H., & Stetter, M. (2005). Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on fMRI data. *Neuroimage*, *28*, 980–995.

Müller A., Candrian G., Kropotov J. D., Ponomarev V. A., & Baschera, G. M. (2010). Classification of ADHD patients on the basis of independent ERP components using a machine learning system. *Nonlinear Biomedical Physics 4*(*Suppl 1*), S1. doi: 10.1186/1753-4631-4-S1-S1

Müller, K. R., Tangermann, M., Dornhege, G., Krauledat, M., Curio, G., & Blankertz, B. (2008). Machine learning for real-time single-trial EEG-analysis: From brain-computer interfacing to mental state monitoring. *Journal of Neuroscience Methods*, *167*, 82–90.

Nikkel, L., & Karrer, R. (1994). Differential effects of experience on the EEG and behavior of 6-month-old infants: Trends during repeated stimulus presentations. *Developmental Neuropsychology*, *10*, 1–11.

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, *24*, 1565–1567.

Onton, J., Westerfield, M., Townsend, J., & Makeig, S. (2006). Imaging human EEG dynamics using independent component analysis. *Neuroscience & Biobehavioral Reviews*, *30*(6), 808–822.

Pang, H., Tong, T., & Zhao, H. (2009). Shrinkage-based diagonal discriminant analysis and its applications in high-dimensional data. *Biometrics*, *65*(4), 1021–1029.

Pardoe, I., Yin, X., & Cook, R. D. (2007). Graphical tools for quadratic discriminant analysis. *Technometrics*, *49*(2), 172–183.

Park, C. H., & Park, H. (2008). A comparison of generalized linear discriminant analysis algorithms. *Pattern Recognition*, *41*(3), 1083–1097.

Pereira, F., Mitchell, T., & Botvinick, M. (2008). Machine learning classifiers and fMRI: A tutorial overview. *Neuroimage*, *5*(1), S199–209.

Picton, T. W., Bentin, S., Berg, P., Donchin, E., Hillyard, S. A., Johnson, R. Jr, . . . Taylor, M. J. (2000). Guidelines for using human event-related potentials to study cognition: Recording standards and publication criteria. *Psychophysiology*, *37*, 127–152.

Poulet, F. (2004). SVM & graphical algorithms: A cooperative approach. *Proceedings of the Fourth IEEE International Conference on Data Mining*. Washington, DC: IEEE Computer Society.

R Development Core Team. (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Retrieved from http://www.R-project.org

Rakotomamonjy, A. (2003). Variable selection using SVM-based criteria. *Journal of Machine Learning Research*, *3*, 1357–1370.

Rakotomamonjy, A., & Guigue, V. (2008). Competition III: Dataset II- ensemble of SVMs for BCI P300 speller. *IEEE Transactions on Biomedical Engineering*, *55*, 1147–1154.

Rassmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Cambridge, MA: MIT Press.

Reynolds, G. D., & Guy, M. W. (2012). Brain–behavior relations in infancy: Integrative approaches to examining infant looking behavior and event-related potentials. *Developmental Neuropsychology*, *37*, 210–225.

Salvaris, M., & Sepulveda, F. (2009). Visual modifications on the P300 speller BCI paradigm. *Journal of Neural Engineering*, 6:046011, 1–8. doi: 10.1088/1741-2560/6/4/046011

Sanei, S., & Chambers, J. A. (2007). *EEG signal processing*. New York, NY: John Wiley & Sons.

Schölkopf, B., Smola, A., & Müller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, *10*, 1299–1319.

Snyder, K. A., Webb, S. J., & Nelson, C. A. (2002). Theoretical and methodological implications of variability in infant brain response during a recognition memory paradigm. *Infant Behavior and Development*, *25*, 466–449.

Spencer, K. M. (2005). Averaging, detection, and classification of single-trial ERPs. In T. C. Handy (Ed.), *Event-related potentials: A methods handbook* (pp. 209–227). Cambridge, MA: The MIT Press.

Stahl, D., Parise, E., Hoehl, S., & Striano, T. (2010). Eye contact and emotional face processing in 6-month-old infants: Advanced statistical methods applied to event related potentials. *Brain and Development*, *32*(4), 305–317.

Stets, M., & Reid, V. M. (2011). Infant ERP amplitudes change over the course of an experimental session: Implications for cognitive processes and methodology. *Brain and Development*, *33*(7), 558–568.

Stets, M., Stahl, D., & Reid, V. M. (2012). A meta-analysis investigating factors underlying attrition rates in infant ERP studies. *Developmental Neuropsychology*, *37*, 226–252.

Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences*, 5th ed. New York, NY: Routledge Academic.

Stone, J. V. (2004). *Independent component analysis: A tutorial introduction*. Cambridge, MA: MIT Press.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*, *5th ed*. Boston, MA: Allyn and Bacon.

Tai, F., & Pan, W. (2008). Incorporating prior knowledge of gene functional groups into regularized discriminant analysis of microarray data. *Bioinformatics*, *23*(23), 3170–3177.

Thierry, G. (2005). The use of event-related potentials in the study of early cognitive development. *Infant and Child Development*, *14*, 85–94.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York, NY: Springer.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S. Fourth edition*. New York, NY: Springer.

Vidaurre, C., Schlögl, A., Cabeza, R., Scherer, R., & Pfurtscheller, G. (2005). Adaptive on-line classification for EEG-based brain computer interfaces with AAR parameters and band power estimates. *Biomedical Engineering*, *50*(11), 350–354.

Wang, B., Wan, F., Mak, P. U., Mak, P. I., & Vai, M. I. (2009). EEG signals classification for brain computer interfaces based on Gaussian process classifier. In P. Shum, M. I. Vai, & L. Wang (Eds.), *Proceedings of the 7th international conference on information, communications and signal processing (ICICS′09)* (pp. 784–788). Piscataway, NJ: IEEE Press.

Weihs, C., Ligges, U., Luebke, K., & Raabe, N. (2005). klaR analyzing German business cycles. In D. Baier, R. Decker, & L. Schmidt-Thieme (Eds.), *Data analysis and decision support* (pp. 335–343). Berlin, Germany: Springer-Verlag.

Weiss, G. (2008). Mining with rarity: A unifying framework. *SIGKDD Explorations*, *6*(1), 7–19.

Woodman, G. F. (2010). A brief introduction to the use of event-related potentials in studies of perception and attention. *Attention, Perception and Psychophysics*, *72*(8), 2013–2046.

Yamagishi, Y., Tsubone, T., & Wada, Y. (2008). Possibility of reinforcement learning based on event-related potential. *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, 654–657. doi: 10.1109/IEMBS.2008.4649237

Zhong, M., Lotte, F., Girolami, M., & Lécuyer, A. (2008). Classifying EEG for brain computer interfaces using Gaussian processes. *Pattern Recognition Letters*, *29*, 354–359.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, *67*(2), 301–320.

Zuber, V., & Strimmer, K. (2009). Gene ranking and biomarker discovery under correlation. *Bioinformatics*, *25*(20), 2700–2707.