Binary classification with pFDR-pFNR losses

Thorsten Dickhaus[*,1], **Benjamin Blankertz**[2,3], and **Frank C. Meinecke**[4]

[1] Department of Mathematics, Humboldt-University Berlin, Unter den Linden 6, D-10099 Berlin, Germany
[2] Neurotechnology Group, Berlin Institute of Technology, Marchstrasse 23, D-10587 Berlin, Germany
[3] Bernstein Center for Computational Neuroscience Berlin, Philippstrasse 13, Haus 6, D-10115 Berlin, Germany
[4] Machine Learning/Intelligent Data Analysis Group, Berlin Institute of Technology, Marchstrasse 23, D-10587 Berlin, Germany

Received 28 February 2012; revised 27 November 2012; accepted 30 November 2012

Connecting multiple testing with binary classification, we derive a false discovery rate-based classification approach for two-class mixture models, where the available data (represented as feature vectors) for each individual comparison take values in $\mathbb{R}^d$ for some $d \geq 1$ and may exhibit certain forms of autocorrelation. This generalizes previous findings for the independent case in dimension $d = 1$. Two resulting classification procedures are described which allow for incorporating prior knowledge about class probabilities and for user-supplied weighting of the severity of misclassifying a member of the "0"-class as "1" and vice versa. The key mathematical tools to be employed are multivariate estimation methods for probability density functions or density ratios. We compare the two algorithms with respect to their theoretical properties and with respect to their performance in practice. Computer simulations indicate that they can both successfully be applied to autocorrelated time series data with moving average structure. Our approach was inspired and its practicability will be demonstrated by applications from the field of brain-computer interfacing and the processing of electroencephalography data.

Keywords: Autocorrelated data; Brain-computer interfaces; Density ratio estimation; False discovery rate; Weak dependence.

1 Multiple testing and binary classification

Over the last two decades, multiple testing has become a "hot topic" in mathematical and applied statistics, see the bibliometric overview by Benjamini (2010). Due to rapid technical developments in many scientific fields, the number m of tests to be performed simultaneously can nowadays become extremely large. In genetic association studies, for example, common values are $m \approx 5 \times 10^5$ or $m \approx 10^6$. When such massive multiplicity is encountered, control of the false discovery rate (FDR) is a popular approach. Instead of controlling the family-wise error rate, i.e., the probability of one or more type I errors, Benjamini and Hochberg (1995) proposed to control the expected proportion of false positives among all significant findings, which typically implies to allow for a few type I errors. This idea, together with the simple and intuitive linear step-up procedure, proved to be useful and attractive to practitioners such that the "Benjamini–Hochberg correction"

*Corresponding author: e-mail: dickhaus@math.hu-berlin.de, Phone: +49-30-2093-5841, Fax: +49-30-2093-5848

meanwhile can be found in many major statistics software packages. Nowadays, the literature on FDR methodology and practice is exponentially increasing over time and some very sophisticated FDR-controlling multiple test procedures have recently been introduced under a variety of perspectives, cf., e.g., Storey et al. (2004), Genovese and Wasserman (2004), Finner et al. (2009), Blanchard and Roquain (2008, 2009). Although the FDR was originally introduced as a purely frequentist statistical tool, also (empirical) Bayesian interpretations and methods have been developed, cf. Storey (2003), Efron (2003, 2008), Do et al. (2005), Müller et al. (2004, 2007). In this work, we exploit some of these Bayesian perspectives toward the FDR and utilize them for binary classification problems.

Classification, also referred to as pattern recognition, denotes the problem of assigning objects to one of a finite number of classes. This problem is often addressed by statistical learning techniques, see, for instance, Hastie et al. (2009) and Vapnik (1998) for introductions. Recently, statistical learning methods for classification have also found their way into medical applications, see Freidlin et al. (2010) and Lipkovich et al. (2011), for example. Binary classification and multiple testing are related statistical fields. The decision pattern of a multiple test for a family of m hypotheses has the same structure as the output of a binary classifier for m test data points, namely, a vector in $\{0, 1\}^m$ indicating the m binary decisions. Moreover, in both problems realizations $\mathbf{x}_i$, $1 \leq i \leq m$, of random vectors $\mathbf{X}_i$ with values in $\mathbb{R}^d$ build the basis for the decision rule (the multiple test or the classifier), which is thus chosen according to statistical criteria. In the testing context, $\mathbf{x}_i$ has the interpretation of a data sample for the i-th individual test, while $\mathbf{x}_i$ is referred to as the i-th feature vector to be classified in the classification terminology.

On the other hand, usual loss functions for binary classification differ from the ones that are typically utilized in multiple testing. In particular, multiple tests controlling a type I error rate like the FDR at a fixed significance level are in general not good classifiers, because they treat null hypotheses and alternatives asymmetrically. Exceptions are sparse cases where class probabilities are highly unbalanced. Under sparsity, multiple tests can, at least asymptotically ($m \to \infty$), achieve optimal classification risks as shown by Bogdan et al. (2011). In the context of model selection, minimax properties of FDR thresholding under sparsity have been derived by Abramovich et al. (2006). For a different classification approach in high dimensions under sparsity assumptions which is related to multiple testing, see Donoho and Jin (2008).

In contrast to the latter works, we consider nonsparse cases and draw the connection between multiple testing and binary classification not through the analysis of specific procedures, but through the interrelation of particular error rates. We will show that the Bayes risk of a classifier can be closely related to the weighted average of the positive FDR (pFDR) and its type II analogue, the positive false nondiscovery rate (pFNR). This relationship has already been mentioned by Storey (2003) for the case of $d = 1$ and independent and identically distributed (iid.) $\mathbf{X}_i$. We will generalize the results of Storey (2003) to treat the cases of $d \geq 1$ and stationary, but (nontrivially) autocorrelated feature vectors. For the construction of classifiers based on this connection, we restrict our attention to situations where classification has to be performed sequentially and immediately after having observed $\mathbf{x}_i \in \mathbb{R}^d$, meaning that the information $\mathbf{x}_{i+1}$ is not accessible. This holds true especially in cases where on-line feedback is required as in the application presented in Section 5. Consequently, we will focus on fixed rejection regions or, in other words, single-step techniques.

The paper is organized as follows. In Section 2, theoretical results from Bayesian classification theory are recalled and the outlined relationship between the Bayes risk and the multiple testing error rates pFDR and pFNR is established. Section 3 is then concerned with two algorithms for binary classification by minimizing a weighted average of pFDR and pFNR. By means of computer simulations, we demonstrate the accuracy of these algorithms in case of stationary, but autocorrelated moving average time series in Section 4. Finally, Section 5 discusses the applicability of our approach in practice by analyzing real-life data from the field of brain-computer interfacing. We conclude with a discussion in Section 6.

2 Bayesian classification rules

We formalize the binary classification problem as follows. Assume that training data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are at hand. In this, the $\mathbf{x}_\ell$ are feature vectors with values in $\mathbb{R}^d$ as described before and each $y_\ell \in \{0, 1\}$ is a class indicator which we will call a label in the remainder of this work. Furthermore, assume that test data points $\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N$ with unknown and random labels $Y_{n+1}, \ldots, Y_N$ are given or about to be encountered. Formally, we can now describe the classification task by the pairs of hypotheses

$$H_{0,i} : Y_{n+i} = 0 \text{ versus } H_{1,i} : Y_{n+i} = 1 \text{ for } i = 1, \ldots, m = N - n.$$

The number m of test data points to be classified may be known or unknown in advance. Let prior probabilities $\pi_{0,i} = \pi(H_{0,i})$ and $\pi_{1,i} = 1 - \pi_{0,i} = \pi(H_{1,i})$ for the validity of the i-th null and alternative hypothesis, respectively, be given. Equivalently, we can state that we have prior distributions $Y_{n+i} \sim \text{Bernoulli}(\pi_{1,i})$ for the unknown labels which are regarded as the "parameters" in the classification model. The data-generating process is modeled by a (joint) probability measure $\mathbb{P}$, where some systematic relationship between Y_{n+i} and $\mathbf{X}_{n+i}$ for $1 \le i \le m$ is assumed. The random vectors $\mathbf{X}_{n+1}, \ldots, \mathbf{X}_N$ are assumed to be continuously distributed with class-conditional cumulative distribution functions (cdfs) given by $F_{j,i}(\mathbf{x}) = \mathbb{P}(\mathbf{X}_{n+i} \le \mathbf{x} | H_{j,i})$ for $j = 0, 1$, $i = 1, \ldots, m$, and $\mathbf{x} \in \mathbb{R}^d$ (where orderings are component-wise) and with corresponding probability density functions (pdfs) $f_{j,i}$.

Application of Bayes' theorem for densities yields that the posterior probability for $H_{0,i}$ given the observed data $\mathbf{x}_{n+i}$ can be expressed by

$$\mathbb{P}(H_{0,i} | \mathbf{X}_{n+i} = \mathbf{x}_{n+i}) = \frac{\pi_{0,i} f_{0,i}(\mathbf{x}_{n+i})}{\pi_{0,i} f_{0,i}(\mathbf{x}_{n+i}) + \pi_{1,i} f_{1,i}(\mathbf{x}_{n+i})} = 1 - \mathbb{P}(H_{1,i} | \mathbf{X}_{n+i} = \mathbf{x}_{n+i}). \tag{1}$$

A decision rule for classification can consequently be based on the test statistics $T_i = \mathbb{P}(H_{0,i} | \mathbf{X}_{n+i})$, $1 \le i \le m$. Moreover, as mentioned in the introduction, we consider the case of single trial sequential classification (where each "trial" corresponds to one value of i) and hence apply a single-step technique, i.e., we employ a critical region Γ_i (say) for T_i, $i = 1, \ldots, m$, the construction of which only depends on properties of $\mathbb{P}$ and the training data, but not on any data point $\mathbf{x}_{n+i'}$ within the test data set for $i' \ne i$. For the determination of Γ_i, we consider for a given cost parameter $c \in (0, 1)$ minimization of the Bayes risk associated with the action $a_i = \mathbf{1}_{\Gamma_i}(\mathbf{x}_{n+i})$, i.e.,

$$R_\pi(\Gamma_i) = (1 - c)\, \mathbb{P}(\mathbf{X}_{n+i} \in \Gamma_i, Y_{n+i} = 0) + c\, \mathbb{P}(\mathbf{X}_{n+i} \notin \Gamma_i, Y_{n+i} = 1).$$

Similarly as in the Neyman–Pearson fundamental lemma, a set of best "rejection regions" is then given by (see, for instance, Section 5.3.3 in Berger, 1985)

$$\Gamma_i \equiv \Gamma_{i,c} = \left\{ \mathbf{x} : \frac{\pi_{0,i} f_{0,i}(\mathbf{x})}{\pi_{0,i} f_{0,i}(\mathbf{x}) + \pi_{1,i} f_{1,i}(\mathbf{x})} \le c \right\}. \tag{2}$$

However, the fact that $m > 1$ data points have to be classified in our setting provokes the question if multiple testing error rates are appropriate classification criteria in this context, too. Interestingly, as pointed out by Storey (2003) for the case of $d = 1$, the following connection to multiple testing theory can be drawn: If all $\pi_{0,i}$ are identical ($\pi_{0,i} \equiv \pi_0$) and the $\mathbf{X}_i$ are (marginally) iid. with mixture pdf $\pi_0 f_0 + \pi_1 f_1$, then among the sets considered in (2) there is also a rejection region that minimizes the weighted average of the pFDR and the pFNR, i.e., it exists a constant $c(w)$ such that

$$(1 - w) \cdot \text{pFDR}(\Gamma_{c(w)}) + w \cdot \text{pFNR}(\Gamma_{c(w)}) = \min_{\Gamma} (1 - w) \cdot \text{pFDR}(\Gamma) + w \cdot \text{pFNR}(\Gamma). \tag{3}$$

The quantities $\mathrm{pFDR}(\Gamma)$ and $\mathrm{pFNR}(\Gamma)$ in (3) are given by

$$\mathrm{pFDR}(\Gamma) = \mathbb{P}(H_0|\mathbf{X} \in \Gamma) = \mathbb{E}\left[\frac{V(\Gamma)}{R(\Gamma)}|R(\Gamma) > 0\right],$$
$$\mathrm{pFNR}(\Gamma) = \mathbb{P}(H_1|\mathbf{X} \notin \Gamma) = \mathbb{E}\left[\frac{T(\Gamma)}{W(\Gamma)}|W(\Gamma) > 0\right],$$

and we dropped the index i at every occurrence according to the iid. assumptions. Moreover, $\mathrm{pFDR}(\Gamma)$ and $\mathrm{pFNR}(\Gamma)$ do not even depend on m, a well-known scalability property in FDR research. The multiple testing quantities V, R, T, and W denote the number of type I errors, the total number of rejections, the number of type II errors and the total number of nonrejections, respectively. The weighted average $A(w) = (1-w)\cdot \mathrm{pFDR}(\Gamma_{c(w)}) + w \cdot \mathrm{pFNR}(\Gamma_{c(w)})$ from Eq. (3) can therefore be interpreted either as a weighted Bayesian classification risk measure or as a weighted combination of frequentist multiple testing error measures. The severity of misclassifying a member of the "0"-class as "1" compared to misclassifying a member of the "1"-class as "0" can be specified by choosing the weight parameter w. If the problem is totally symmetric, one would choose $w = 1/2$. In analogy to (1), $\mathrm{pFDR}(\Gamma)$ and $\mathrm{pFNR}(\Gamma)$ can be calculated by

$$\mathrm{pFDR}(\Gamma) = \frac{\pi_0 I_0(\Gamma)}{\pi_0 I_0(\Gamma) + \pi_1 I_1(\Gamma)} \quad \text{and} \quad \mathrm{pFNR}(\Gamma) = \frac{\pi_1[1 - I_1(\Gamma)]}{\pi_1[1 - I_1(\Gamma)] + \pi_0[1 - I_0(\Gamma)]} \tag{4}$$

with $I_j(\Gamma) = \int_\Gamma f_j(\mathbf{u})\lambda^d(d\mathbf{u})$, $j = 0, 1$, showing that the Bayes risk $R_\pi(\Gamma)$ can be regarded as a local version of $A(w)$.

Some analytical results on FDR-controlled classification can be found in the works by Scott et al. (2009) and Genovese and Wasserman (2004). In the present work, we focus on practical considerations, namely, implementability of Algorithm 3.1 below and alternative computational approaches based on direct estimation of density ratios. Throughout the remainder, we will make the assumption that (at least for large values of N) the (class-conditional) distribution of every test data point can be described well by stationary pdfs f_0 and f_1, but we will relax the independence assumption made by Storey (2003).

3 Implementing the classification rule

From (1), (2), and (4) it follows that the prior probability π_0 and the pdfs f_j, $j = 0, 1$, together with the test data, entirely determine the classification procedure. However, since f_0 and f_1 are typically unknown in practice, some kind of inference for these quantities will be required in most practically relevant situations. We will use the training data for this purpose and pursue the classification task from an empirical Bayes perspective. With respect to the labels, many experimental designs may automatically induce a reasonable prior. If not, then also π_0 may be estimated from training data. Consequently, in Section 7 of Storey (2003), the following algorithm for practical implementation is outlined.

Algorithm 3.1

1. *Utilizing training data* $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$*, estimate the pdfs* f_0 *and* f_1 *by* $\hat{f}_j$, $j = 0, 1$.
2. *Approximate the sets* Γ_c *for given* $c \in (0, 1)$ *by plugging* $\hat{f}_j$ *into (2) instead of* f_j, $j = 0, 1$. *The prior probability* π_0 *can either be chosen explicitly or estimated from the training data.*
3. *Estimate pFDR*(Γ_c) *and pFNR*(Γ_c) *by numerical integration in (4) with* f_j *replaced by* $\hat{f}_j$, $j = 0, 1$.

4. *Choose* $w \in (0, 1)$ *and minimize the approximation of* $(1 - w) \cdot pFDR(\Gamma_c) + w \cdot pFNR(\Gamma_c)$ *with respect to c. Denote the solution by* $\hat{c}(w)$.
5. *For the test data, form (realized) test statistics* $T_i = t_i$, $i = 1, \ldots, m$, *according to Eq. (1), again with* f_j *replaced by* $\hat{f}_j$, $j = 0, 1$.
6. *Classify* $\hat{y}_{n+i} = \mathbf{1}\{t_i \leq \hat{c}(w)\}$ *for* $i = 1, \ldots, m$.

In fact, Storey (2003) describes a slightly different approach, namely, estimating f_0 from training data drawn from the zero class and estimating the marginal $f = \pi_0 f_0 + \pi_1 f_1$ from possibly unlabeled data. This possibility makes the underlying statistical model also an appropriate framework for the statistical learning task of semi-supervised novelty detection as in Blanchard et al. (2010). Anyhow, re-writing

$$\Gamma_c = \left\{ \mathbf{x} : \frac{\pi_0 f_0(\mathbf{x})}{\pi_1 f_1(\mathbf{x})} \leq \frac{c}{1-c} \right\} = \left\{ \mathbf{x} : \lambda(\mathbf{x}) \leq \frac{\pi_1 c}{\pi_0 (1-c)} \right\}, \tag{5}$$

it becomes evident that the pdfs f_j, $j = 0, 1$, are indeed only needed in (2) via their likelihood ratio λ, defined by $\lambda(\mathbf{x}) = f_0(\mathbf{x})/f_1(\mathbf{x})$. Therefore, direct estimation of λ from the training data is a promising alternative approach, especially if labeled training data from both classes are available and the dimensionality d is larger than 1. The reason is that, according to Vapnik (1995), one should never try to solve a harder problem than the one that one is actually concerned with. Once Γ_c can approximately be determined for any fixed value c via estimation of λ, pFDR and pFNR may just be estimated by relative frequencies. Algorithmically, this alternative computational solution can be summarized as follows.

Algorithm 3.2

1. *Utilizing training data* $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, *estimate the density ratio* λ *by* $\hat{\lambda}$.
2. *Approximate the set* Γ_c *for given* $c \in (0, 1)$ *by plugging* $\hat{\lambda}$ *instead of* λ *into (5). The prior probability* π_0 *can either be chosen explicitly or estimated from the training data.*
3. *Estimate* $pFDR(\Gamma_c)$ *and* $pFNR(\Gamma_c)$ *by calculating the relative frequencies of events* $\{y_i = 0\}$ *in the training sub-dataset with* $\mathbf{x}_i \in \Gamma_c$ *and* $\{y_i = 1\}$ *in the training sub-dataset with* $\mathbf{x}_i \notin \Gamma_c$, *respectively.*
4. *Choose* $w \in (0, 1)$ *and minimize the approximation of* $(1 - w) \cdot pFDR(\Gamma_c) + w \cdot pFNR(\Gamma_c)$ *with respect to c. Denote the solution by* $\hat{c}(w)$.
5. *For the test data, form (realized) test statistics* $T_i = t_i$, $i = 1, \ldots, m$, *according to Eq. (1), again with* λ *replaced by* $\hat{\lambda}$.
6. *Classify* $\hat{y}_{n+i} = \mathbf{1}\{t_i \leq \hat{c}(w)\}$ *for* $i = 1, \ldots, m$.

Both algorithms are in concurrence and we will discuss assets and drawbacks of both. From the implementational point of view it is remarkable that in both algorithms the steps 2 to 4 can be done in parallel so that the problem essentially boils down to a single optimization with respect to $c(w)$.

For usage in non-i.i.d. situations as exemplified in Section 5, we will now formulate a relaxed framework for applying the results in Section 2 in the case that the training data set is large and f_j is systematically replaced by $\hat{f}_j$, $j = 0, 1$. Loosely speaking, we have to assume that the distributions of training and test data share important properties and that the distributions of the $\mathbf{X}_i$ can for large N approximately be characterized by stationary densities f_j, $j = 0, 1$, depending on the label $Y_i = j$. A frequently encountered modeling approach consists of weakly dependent mixtures as formalized in the following definition.

Definition 3.3 (Weakly dependent mixture model). Under the general setup of binary classification as presented in Section 2, the multivariate distribution of the data tuples $\{(\mathbf{X}_i, Y_i)\}_{i=1}^N$ on $(\mathbb{R}^d \times \{0,1\})^N$ is called a weakly dependent mixture model if the conditions (6)–(8) are fulfilled.

$$\text{Defining } \gamma_N = n/N, \text{ it holds } 0 < \liminf_{N\to\infty} \gamma_N \leq \limsup_{N\to\infty} \gamma_N < 1. \tag{6}$$

Independently of the values in the sequence $\{\gamma_N\}_N$, it holds

$$N^{-1}\sum_{i=1}^{N}(1-Y_i) \underset{a.s.}{\to} \pi_0 = 1-\pi_1 \in (0,1). \tag{7}$$

There exist continuous cdfs $F_0 \neq F_1$ on $\mathbb{R}^d$, not depending on the values in the sequence $\{\gamma_N\}_N$, with the property that for all $\mathbf{x} \in \mathbb{R}^d$ it holds

$$N_0^{-1}\sum_{i=1}^{N}(1-Y_i)\,\mathbf{1}\{\mathbf{X}_i \leq \mathbf{x}\} \underset{a.s.}{\to} F_0(\mathbf{x}) \text{ and } N_1^{-1}\sum_{i=1}^{N} Y_i\,\mathbf{1}\{\mathbf{X}_i \leq \mathbf{x}\} \underset{a.s.}{\to} F_1(\mathbf{x}), \tag{8}$$

with self-explaining counts N_0 and N_1.

The distributional assumptions in Definition 3.3 imply that the Y_i asymptotically follow a Bernoulli(π_1)-distribution and the $\mathbf{X}_i$ are asymptotically distributed as a mixture $(1-Y_i)F_0 + Y_iF_1$. Under these assumptions, at least for large training data sets, the asymptotic pdfs f_0 and f_1 can consistently be estimated from the training data where labeling is known and separate estimation of the densities corresponding to $y_i = 0$ and $y_i = 1$, respectively, is possible. Condition (7) together with (6) ensures that, if training and test phase are systematically connected, one can use the relative frequency of $\{y_i = 0\}$, $i = 1, \ldots, n$, in the training phase in order to obtain an objective value for π_0. If, however, the test phase is de-coupled from training with respect to the relative occurrence of labels, one can also use noninformative prior probabilities $\pi_0 = \pi_1 = 1/2$ (or any other Bernoulli prior). For instance, the experimental design may change from the training to the test phase as it is occasionally the case in applications from the field of brain-computer interfacing, see Section 5. Furthermore, notice that Definition 3.3 is only concerned with the dependency structure between distinct random vectors, while the "inner-vector" dependency structure, i.e., the dependence between the components of a particular $\mathbf{X}_i$, is not affected. The latter dependency will implicitly be addressed in the density (ratio) estimation.

Remark 3.4

1. The concept of weakly dependent mixtures or weak dependency in general was introduced in the multiple testing context by Storey et al. (2004). From Vapnik–Chervonenkis theory, it is known that sequences of (conditionally to the labels) iid. random vectors fulfill (8) with an exponential rate of convergence, see, e.g., Theorems 12.11 and 12.12 in DasGupta (2008). However, it may be assumed that the class of weakly dependent models is considerably broader. Exemplary models fulfilling the weak dependency assumption in the case of $d = 1$ are models with block dependencies (cf. Chapter 4 in Gontscharuk, 2010), moving average processes (see Clarke and Hall, 2009), or α-mixing processes with α not "too large" (see Farcomeni, 2007).
2. For estimation of f_0 and f_1, parametric and nonparametric multivariate density estimation methods may be employed. A comparison of different nonparametric methods can be found in Hwang et al. (1994). It has to be kept in mind that the resulting estimated densities $\hat{f}_j$, $j = 0, 1$ should not be too involved, because one has to perform (numerical) integration of the $\hat{f}_j$'s over Γ_c for evaluating pFDR and pFNR. Therefore, if no parametric family of densities can be assumed, we propose (at least for d not too large) to employ a fixed-width kernel density estimator with a Gaussian kernel and empirically sphered data. Excellent textbook references for density estimation are Silverman (1986) and Härdle et al. (2004).

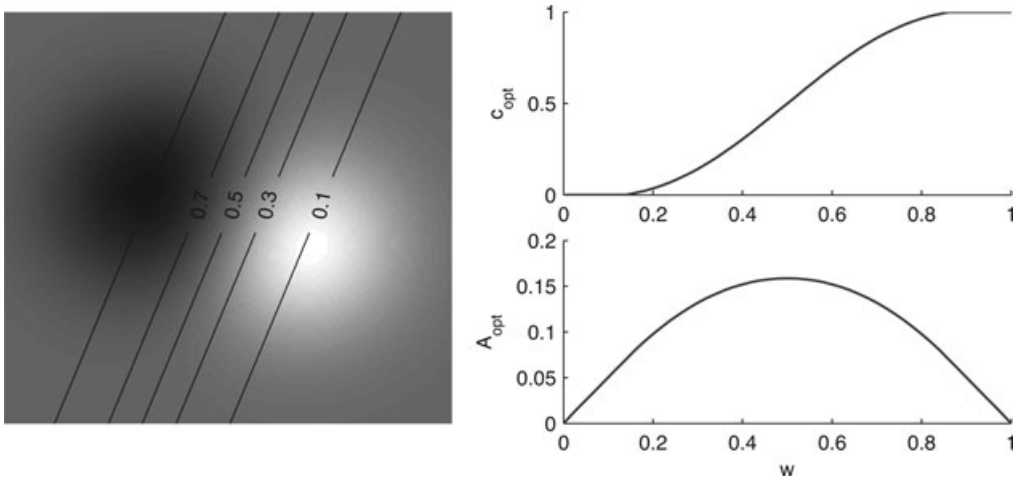

Figure 1 Class-conditional densities f_j, $j = 0, 1$ on $\mathbb{R}^2$ are given by spherical Gaussians with identical covariance matrices. The brightness in the left panel encodes the density difference $f_0 - f_1$. Different values of c correspond to linear and parallel decision thresholds (black lines). The upper right panel shows the value $c(w)$ of the optimal threshold as a function of w and the lower right panel the corresponding value of the cost function $A(w) = (1 - w)\text{pFDR} + w\text{pFNR}$.

3. In a series of papers (cf. Sugiyama et al., 2009 and references therein), a group of Japanese researchers developed methods for and discussed applications of direct estimation of density ratios avoiding plug-in of estimated densities. In Sections 2.8 and 4 of Sugiyama et al. (2009), especially the so-called `uLSIF` algorithm is propagated.

4 Computer simulations

In order to assess and compare the accuracy of the estimated values of $c(w)$ resulting from Algorithms 3.1 and 3.2 in the case of correlated data, we simulated a multivariate time series model for which a strictly stationary solution is known to exist and available in an explicit form. To this end, we considered vector-valued moving average processes driven by Gaussian noise as described, e.g., in Chapter 11 of Brockwell and Davis (1991). For ease of graphical presentation, we restrict our attention to dimensionality $d = 2$ here. The data space of some of the real data examples in Section 5 has higher dimensionality.

We start with the simple classification task where the class-conditional stationary densities are given by two spherical Gaussians with identical covariance matrices. In this case, there exists an explicit solution to the classification problem. From Eqs. (3) and (5), we know that for given weight parameter w the value $c(w)$, which defines the border of the rejection region, only depends on the prior-weighted density ratio $\pi_0 f_0(x)/(\pi_1 f_1(x))$. In our special case, the iso-lines that correspond to different values of c are straight lines, perpendicular to the difference vector between the class means (see Fig. 1, left panel).

Therefore, the optimal classifier is the well-known solution of *linear discriminant analysis* (LDA). Depending on the choice of the weight w, an optimal value of c (i.e., decision threshold) can be determined. Obviously, $c(w)$ increases strictly monotonously with w, but in general this relationship is nonlinear; here, it corresponds to a sigmoidal curve (Fig. 1, upper right panel). The lower right panel shows the graph of the cost function $A(w) = (1 - w)\text{pFDR} + w\text{pFNR}$. For $w = 0$ only misclassified

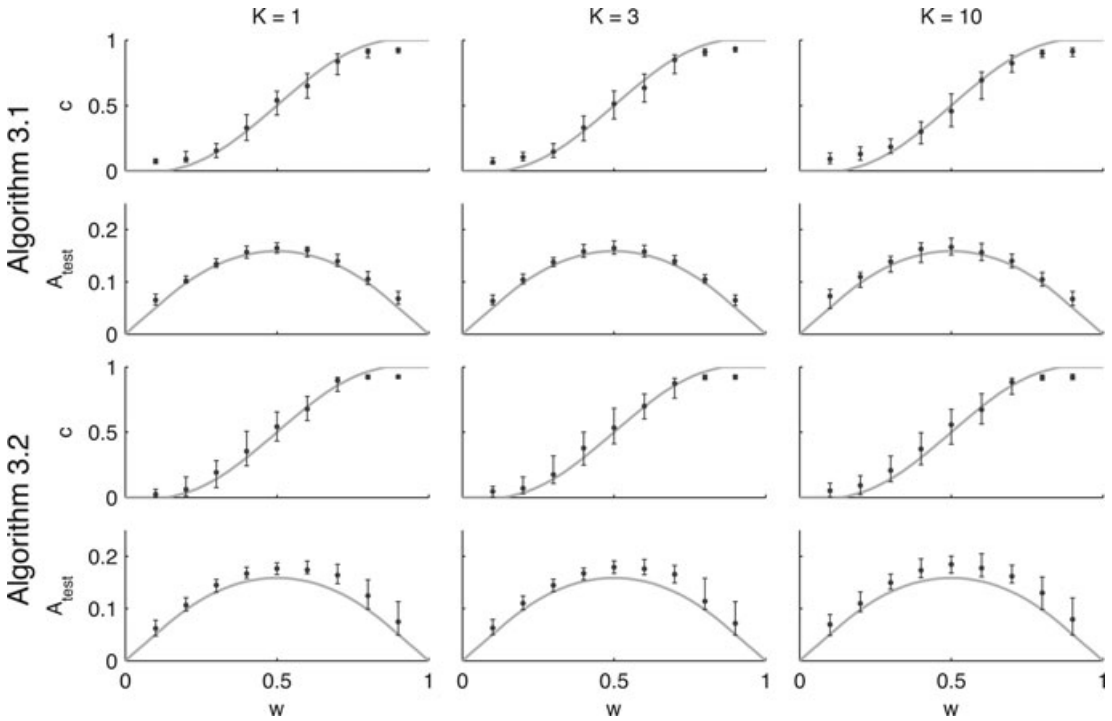

Figure 2 Simulation results for Algorithms 3.1 and 3.2, respectively, in the case of identical covariance matrices. The first and the third row show the values $\hat{c}(w)$ estimated from the training data and the second and fourth row the achieved cost function values on the test data (median and 66% range over 100 repetitions). The gray lines correspond to the respective theoretical optimum also shown in Fig. 1. Every column corresponds to a different range K of autocorrelation.

samples from one class contribute to the cost function, so that its value can trivially be brought to zero by always deciding for this class. The same holds true for $w = 1$. Since the classes overlap (the normal distribution is supported on its entire probability space), $A(w)$ takes a maximum in between.

To test how well the proposed algorithms, which do not impose any structural assumption on f_j, $j = 0, 1$, can reproduce these theoretically optimal functions in presence of (nontrivial) autocorrelations, we applied them to finite samples of data drawn from these distributions. The number of training data points was chosen as 200 for each class, and the classification accuracy was assessed on a test set of same size, drawn independently from the training data. Each classification task was repeated 100 times on a different sample and for different choices of w. The upper panel in Fig. 2 displays the results obtained with Algorithm 3.1: In the first row we draw the empirical thresholds $\hat{c}(w)$ estimated from the training set for varying values of w and in the second row the corresponding values of the weighted sum of pFDR and pFNR, achieved on the test data set. For comparison, the continuous gray lines in Fig. 2 correspond to the theoretically optimum values.

While the first column represents results obtained with iid. data $\mathbf{x} = (\mathbf{x}_t : t = 1, \ldots, T = 200)$, the data points $\tilde{\mathbf{x}} = (\tilde{\mathbf{x}}_t : t = 1, \ldots, T)$ used in the second and third column have nonvanishing autocorrelations at lags up to $K - 1$ (with $K = \{1, 3, 10\}$, respectively). This autocorrelation structure has been introduced by moving-average filters, i.e., by constructing

$$\tilde{\mathbf{x}}_t = \sum_{k=1}^{K} b_k \, \mathbf{x}_{t-k+1} \qquad \text{with normalization} \qquad \sum_{k=1}^{K} b_k^2 = 1, \quad K \le t \le T.$$

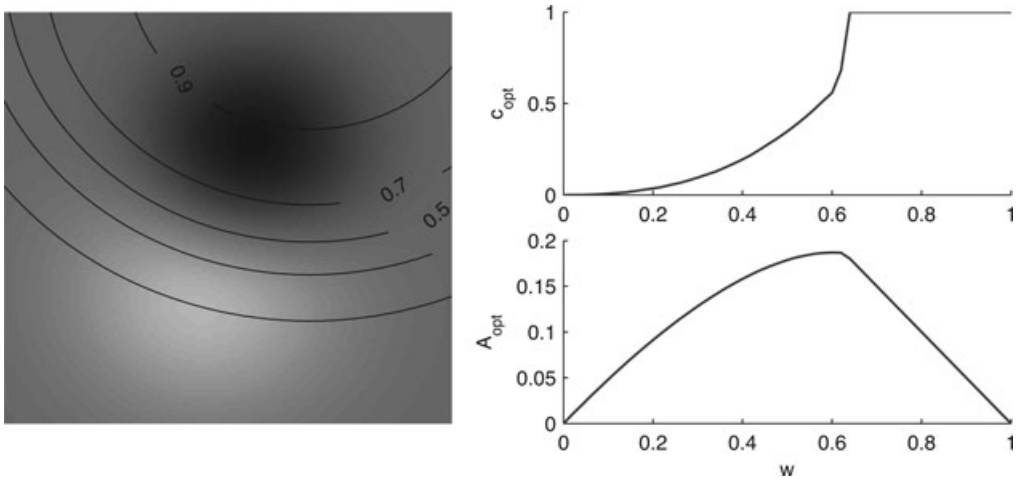

Figure 3 Class-conditional densities f_j, $j = 0, 1$ on $\mathbb{R}^2$ are given by spherical Gaussians with different covariance matrices. The brightness in the left panel encodes the density difference $f_0 - f_1$. Different values of c correspond to quadratic decision thresholds (black lines). The upper right panel shows the value $c(w)$ of the optimal threshold as a function of w and the lower right panel the corresponding value of the cost function $A(w) = (1 - w)\text{pFDR} + w\text{pFNR}$.

The filter coefficients b_k have been randomly drawn from a standard normal distribution and normalized (ensuring that $\tilde{\mathbf{X}}_t$ had the same stationary distribution as $\mathbf{X}_t$ for all time points t). The figure shows that Algorithm 3.1 reproduces the S-shaped relationship between w and the corresponding threshold $c(w)$ well. However, for w far from 1/2, there is a small but systematic bias toward less extreme values of c (compared to the theoretical optimum). This is due to the fact that on a finite sample the classification result remains constant if we increase (decrease) c above (below) a certain threshold due to discreteness of the obtainable empirical classification accuracies. Consequently, this effect has small impact on the actual cost function value evaluated on the test set. Changing the length of autocorrelation has virtually no effect on the performance of Algorithm 3.1; only the error bars grow slightly larger, because we effectively decreased the number of independent data points.

The lower panel in Fig. 2 shows the same plots for Algorithm 3.2 (using the `uLSIF` implementation). This algorithm also captures the true structure, leading to a good approximation. However, now the symmetry between the two classes is broken. This can be seen most prominently in the bottom row displaying the cost function $A(w)$. For high values of w, the results are slightly less accurate than for small values and also the error bars are larger. This is due to the fact that by directly estimating the density ratio, we do not treat the two densities symmetrically. Obviously, fitting f_0/f_1 in a least-squares sense is not the same as fitting f_1/f_0 and taking the reciprocal value. While this issue might be fixed by designing a symmetrical density ratio estimator, this is beyond the scope of the present paper. Here, we only point out that even though it might be faster and sometimes easier to directly estimate the density ratio, this approach can introduce apparent asymmetries between the classes, even if the true problem is symmetric.

In order to cover a case where the underlying classes are truly asymmetric in their stationary distributions, we simulated data such that both class-specific stationary densities are still given by spherical Gaussians, but this time with different covariance matrices, cf. Fig. 3. Here, the optimum threshold lines for different values of c (Fig. 3, left panel) are given by cone sections, which are the solutions of *quadratic discriminant analysis* (QDA). Depending on the choice of w, the optimal threshold $c(w)$ can be calculated (upper right panel). We see that, after a certain weight value, $c(w)$ saturates at 1, meaning that all samples will be classified into the same class, regardless of their

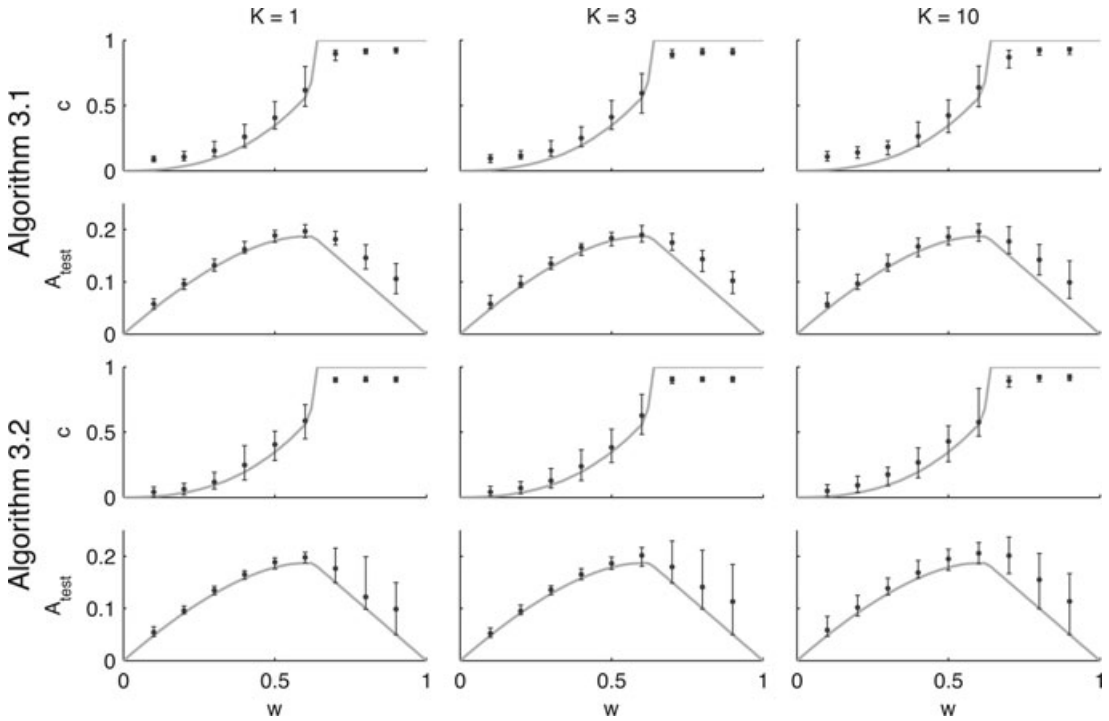

Figure 4 Simulation results for Algorithms 3.1 and 3.2, respectively, in the case of class-specific, different covariance matrices. The first and the third row show the values $\hat{c}(w)$ estimated from the training data and the second and fourth row the achieved cost function values on the test data (median and 66% range over 100 repetitions). The gray lines correspond to the respective theoretical optimum also shown in Fig. 3. Every column corresponds to a different range K of autocorrelation.

position in the data space. Consequently, the cost function declines linearly from this point on (lower right panel), since pFNR $= 0$ and pFDR $= const$.

Figure 4 summarizes the simulation results for Algorithms 3.1 and 3.2, respectively, in this asymmetric case. As in the former setting, the autocorrelations of the data points have only minor effects, leading only to slightly larger error bars. Moreover, the estimated values $\hat{c}(w)$ do not take extreme values (close to 0 or 1), leading to relatively larger fluctuations of $A(w)$ especially in the plateau phase where the true optimal value $c(w)$ is equal to 1 (i.e., for $w > 0.6$).

5 Application: Classification of single-trial EEG data

In this section, we use the term "trial" as in the discussion below (1), meaning that every test data point to be classified results in one classification trial, leading to m trials altogether. The classification of *single-trial* electroencephalogram (EEG) data poses a considerable challenge for data analysis algorithms due to the high trial-to-trial variability and the resulting low signal-to-noise ratio. Mastering of this challenge has gained more and more interest in the research field of brain-computer interfaces (BCIs) (see Wolpaw and Wolpaw, 2012; Dornhege et al., 2007 for an overview and introduction). A BCI is a system that converts brain activity in real-time into related control signals. This can allow, for instance, paralyzed patients who are deprived of other means of communication to interface with the

Table 1 Results of classification with Algorithms 3.1 and 3.2 for BCI data sets from four participants of the screening study described by Dickhaus et al. (2009).

		Algorithm 3.1				**Algorithm** 3.2			
Data set	w	$\hat{c}(w)$	pFDR	pFNR	Error [%]	$\hat{c}(w)$	pFDR	pFNR	Error [%]
1	0.1	0.067	0	0.189	11.67	0.450	0.018	0.204	13.33
	0.3	0.255	0.023	0.130	8.33	0.718	0.024	0.169	11.00
	0.5	0.465	0.043	0.111	8.00	0.718	0.024	0.169	11.00
2	0.1	0.079	0.069	0.324	25.00	0.520	0.045	0.444	40.00
	0.3	0.280	0.076	0.223	16.50	0.866	0.184	0.257	22.50
	0.5	0.497	0.108	0.159	13.50	0.866	0.184	0.257	22.50
3	0.1	0.013	0	0.477	45.67	0.253	0	0.436	38.67
	0.3	0.157	0	0.400	33.33	0.253	0	0.436	38.67
	0.5	0.562	0.144	0.193	17.00	0.668	0.052	0.289	21.33
4	0.1	0.076	0.007	0.080	4.67	0.671	0.143	0.086	11.67
	0.3	0.252	0.014	0.063	4.00	0.671	0.143	0.086	11.67
	0.5	0.460	0.027	0.052	4.00	0.437	0.148	0.087	12.00

external world via BCI-controlled rehabilitative tools. Recently, also nonmedical applications of BCI technologies are being explored (see Blankertz et al., 2010b).

In order to demonstrate the applicability of the algorithms proposed in Section 3 in this context, we refer to a prominent classification problem in the BCI field: the discrimination of *motor intentions* referring to the *left* or *right* hand. Consequently, a BCI could allow a paralyzed patient to continuously transmit a binary control signal via her/his intended movements and thereby control a computer application or an assistive device (see Williamson et al., 2009; Wolpaw and Wolpaw, 2012).

Here, for exemplary purposes, we take BCI data sets from four participants of the screening study described by Dickhaus et al. (2009). In each data set, there are 150 data points (75 for each of the classes *left* and *right* hand motor intention) to train the classification model and 300 data points (again evenly distributed over both classes) for testing. In order to derive feature vectors from each motor intention, we used state-of-the-art methods from BCI research (see Blankertz et al., 2008). Details about the experimental design can be found in Blankertz et al. (2010a).

The obtained features are then processed by Algorithms 3.1 and 3.2 as explained above ($\pi_0 = 1/2$ was implied by the experimental design). In Table 1, we report the classification results for feedback data sets (i.e., out-of-sample classification accuracies) from four experimental subjects and for three different choices of the weight parameter w, namely 0.1, 0.3, and 0.5. In addition to the classification error, the corresponding values of $\hat{c}(w)$ and empirical versions of pFDR and pFNR are listed.

In line with the theoretical properties of classification routines based on the cost function $A(w)$, we observe that $\hat{c}(w)$ is isotone in w and that the empirical values of pFDR and pFNR are monotonously increasing and decreasing in w, respectively. If only the plain classification accuracy is targeted, the best choice for w is obviously $w = 1/2$, since the problem is totally symmetric (equal sample sizes by experimental design and, consequently, $\pi_0 = 1/2$). However, if the BCI shall for instance be used to operate a device with different functions depending on the chosen classes, it may be appropriate to weight classification errors in the two directions differently, because erroneous execution of one command may have a more severe negative impact on the whole system than the other one. This is the case, e.g., in the control of a wheelchair (where false positive motion commands can possibly be dangerous, see Tsui et al., 2011), or in the automatic recognition of BCI errors in mental typewriters

(where the erroneous "correction" of correctly spelled characters is very annoying for the user, see Schmidt et al., 2012).

Comparing Algorithms 3.1 and 3.2, it becomes apparent that, quite consistently, Algorithm 3.1 performs slightly better over the four investigated data sets. This is an unexpected finding. On the whole, however, it is remarkable that with both algorithms the so-called "level criterion" of 70%, which is agreed on in the BCI community to provide a lower bound on the classification accuracy necessary to operate a BCI successfully in practice, can be achieved. As a quasi-standard for classifying logarithmic bandpower features (as used in this study), typically LDA is applied. This choice is due to the experience that such a feature representation allows linear separation of the class-specific data for many BCI users, see Blankertz et al. (2011, 2008). Indeed, in terms of raw classification accuracy, LDA outperforms Algorithm 3.1 on the datasets in Table 1. However, notice that our approach avoids relying on such prior knowledge, because it is fully nonparametric. Therefore, it may also be applied to data from new experimental paradigms for which no laboratory experience is existing yet.

6 Concluding remarks

We have theoretically derived a binary classification rule optimizing the cost function $A(w)$ which is based on multiple testing error rates. The discussion around Eq. (4) showed that, in contrast to the Bayes risk $R_\pi(\Gamma)$ which is a local quantity, $A(w)$ addresses the problem that $m > 1$ individual classification problems have to be solved rather than a single one.

Furthermore, we discussed two algorithms for implementing the classifier into computer software. Algorithm 3.1 seems to be more time-consuming, but has shown better performance on real data and some more convenient properties regarding symmetric treatment of the classes for problems that are symmetric in nature in Section 4. Consequently, from our investigations, Algorithm 3.1 seems preferable. `MATLAB` programs with which the computer simulations presented in Section 4 can be reproduced are available from the authors upon request. EEG raw data, as considered in Section 5, require a signal processing that can most conveniently be done in `MATLAB`, which offers its own toolbox for this task. This is why we also programmed all subsequent data analysis routines in `MATLAB`, so that software platforms do not need to be switched during the classification workflow and the programs can easily be integrated into laboratory usage.

As a direction for further research, it may be interesting to explore how a fully Bayesian optimization approach for $A(w)$ would compare to our empirical Bayesian approach and to discuss reasonable priors for the stationary densities. Moreover, it has to be mentioned that the assumption of stationary (limiting) distributions for both classes is crucial in our present framework. This assumption is not always justified in practice and possible extensions of pFDR-pFNR-based classification to cases where nonstationarities have to be assumed would be desirable. Another interesting idea would be to study a generalized class of multiple testing based cost functions of the form

$$(1-w)g_1(\mathbb{P}^{(V,R)}) + wg_2(\mathbb{P}^{(T,W)}),$$

where g_1 and g_2 are given functionals. The cost function $A(w)$ is a special member of this class, but for instance the weighted average of the generalized family-wise error rate and its type II analogue may also be a suitable criterion for classification.

Finally, we would like to express the opinion that a closer cooperation between the multiple testing community and the field of statistical machine learning, where classification is an important research topic, will be profitable for both sides. The research presented in this work was only realizable by joining the respective expertise. Unfortunately, statistical methodology shows up to now little overlap in both fields. On the other hand, there are even problems where the two inferential problems multiple testing and classification occur at the same time. Consider biomarker studies aiming at classifying

subjects into disease groups on the basis of their biomarker profiles. In a first stage, a subset of relevant markers has to be selected from the very large set of all available biomarkers (a multiple testing problem). Then, in a second stage, classification of subjects is performed on the basis of feature vectors built from the selected biomarkers. Such a two-stage design is chosen by Freidlin et al. (2010), for example. They are considered with the specific problem of identifying a subgroup of cancer patients which is responsive to treatment on the basis of gene expression levels. Freidlin et al. (2010) employ a resampling scheme for the entire two-stage procedure in order to assess statistical significance of treatment effects in the identified subgroup. This can be regarded as a proxy for classification accuracy in this particular context. How misclassification errors resulting from the two-stage classification procedure can be controlled in a general way is an interesting and highly relevant question that requires cooperative research.

Acknowledgments We thank the Editor Lutz Edler, the responsible editor of the Special Issue and a referee for their constructive comments. Special thanks are due to the organizers of MCP 2011 for the successful conference. F.C.M. gratefully acknowledges support by the German Ministry of Education and Research (BMBF) through the "Adaptive BCI" Project, FKZ 01GQ1115.

Conflict of interest
The authors have declared no conflict of interest.

References

Abramovich, F., Benjamini, Y., Donoho, D. L. and Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Annals of Statistics* **34**, 584–653, doi:10.1214/009053606000000074.

Benjamini, Y. (2010). Simultaneous and selective inference: current successes and future challenges. *Biometrical Journal* **52**, 708–721.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **57**, 289–300.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd edn.). Springer Series in Statistics. Springer-Verlag, New York.

Blanchard, G., Lee, G. and Scott, C. (2010). Semi-supervised novelty detection. *Journal of Machine Learning Research* **11**, 2973–3009.

Blanchard, G. and Roquain, E. (2008). Two simple sufficient conditions for FDR control. *Electronic Journal of Statistics* **2**, 963–992.

Blanchard, G. and Roquain, E. (2009). Adaptive false discovery rate control under independence and dependence. *Journal of Machine Learning Research* **10**, 2837–2871.

Blankertz, B., Lemm, S., Treder, M. S., Haufe, S. and Müller, K. R. (2011). Single-trial analysis and classification of ERP components—a tutorial. *NeuroImage* **56**, 814–825, doi:10.1016/j.neuroimage.2010.06.048.

Blankertz, B., Sannelli, C., Halder, S., Hammer, E. M., Kübler, A., Müller, K. R., Curio, G. and Dickhaus, T. (2010a). Neurophysiological predictor of SMR-based BCI performance. *NeuroImage* **51**, 1303–1309, doi:10.1016/j.neuroimage.2010.03.022.

Blankertz, B., Tangermann, M., Vidaurre, C., Fazli, S., Sannelli, C., Haufe, S., Maeder, C., Ramsey, L. E., Sturm, I., Curio, G. and Müller, K. R. (2010b). The Berlin Brain-Computer Interface: non-medical uses of BCI technology. *Frontiers in Neuroscience* **4**, 198, doi:10.3389/fnins.2010.00198. Open Access.

Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M. and Müller, K. R. (2008). Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Processing Magazine* **25**, 41–56, doi:10.1109/MSP.2008.4408441.

Bogdan, M., Chakrabarti, A., Frommlet, F. and Ghosh, J. K. (2011). Asymptotic Bayes-optimality under sparsity of some multiple testing procedures. *Annals of Statistics* **39**, 1551–1579, doi:10.1214/10-AOS869.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods* (2nd edn.). Springer Series in Statistics. Springer-Verlag, Berlin

Clarke, S. and Hall, P. (2009). Robustness of multiple testing procedures against dependence. *Annals of Statistics* **37**, 332–358, doi:10.1214/07-AOS557.

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. Springer Texts in Statistics. Springer, New York, NY

Dickhaus, T., Sannelli, C., Müller, K. R., Curio, G. and Blankertz, B. (2009). Predicting BCI performance to study BCI illiteracy. *BMC Neuroscience* **10** (Suppl 1), P84, doi:10.1186/1471-2202-10-S1-P84.

Do, K. A., Müller, P. and Tang, F. (2005). A Bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society, Series B, Applied Statistics* **54**, 627–644, doi:10.1111/j.1467-9876.2005.05593.x.

Donoho, D. and Jin, J. (2008). Higher criticism thresholding: optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences USA* **105**, 14790–14795.

Dornhege, G., del R. Millán, J., Hinterberger, T., McFarland, D. and Müller, K. R. (Eds.) (2007). *Toward Brain-Computer Interfacing*, MIT Press, Cambridge, MA.

Efron, B. (2003). Robbins, empirical Bayes and microarrays. *Annals of Statistics* **31**, 366–378, doi:10.1214/aos/1051027871.

Efron, B. (2008). Microarrays, empirical bayes and the two-groups model. *Statistical Science* **23**, 1–22.

Farcomeni, A. (2007). Some results on the control of the false discovery rate under dependence. *Scandinavian Journal of Statistics* **34**, 275–297, doi:10.1111/j.1467-9469.2006.00530.x.

Finner, H., Dickhaus, T. and Roters, M. (2009). On the false discovery rate and an asymptotically optimal rejection curve. *Annals of Statistics* **37**, 596–618, doi:10.1214/07-AOS569.

Freidlin, B., Jiang, W. and Simon, R. (2010). The cross-validated adaptive signature design. *Clinical Cancer Research* **16**, 691–698.

Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *Annals of Statistics* **32**, 1035–1061, doi:10.1214/009053604000000283.

Gontscharuk, V. (2010). *Asymptotic and Exact Results on FWER and FDR in Multiple Hypotheses Testing*. Ph.D. thesis, Heinrich-Heine Universität, Düsseldorf.

Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004). *Nonparametric and Semiparametric Models*. Springer Series in Statistics. Springer, Berlin.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction* (2nd ed.). Springer Series in Statistics. Springer, New York, NY. doi:10.1007.b94608.

Hwang, J. N., Lay, S. R. and Lippman, A. (1994). Nonparametric multivariate density estimation: a comparative study. *IEEE Transactions on Signal Processing* **42**, 2795–2810.

Lipkovich, I., Dmitrienko, A., Denne, J. and Enas, G. (2011) Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine* **30**, 2601–2621.

Müller, P., Parmigiani, G. and Rice, K. (2007). FDR and Bayesian multiple comparisons rules. In: J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West (eds.): Bayesian Statistics 8. *Proceedings of ISBA 8th World Meeting on Bayesian Statistics*. Oxford University Press, Oxford, pp. 349–370.

Müller, P., Parmigiani, G., Robert, C. and Rousseau, J. (2004). Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American Statistical Association* **99**, 990–1001, doi:10.1198/016214504000001646.

Schmidt, N. M., Blankertz, B. and Treder, M. S. (2012). Online detection of error-related potentials boosts the performance of mental typewriters. *BMC Neuroscience* **13**, 19, doi:10.1186/1471-2202-13-19.

Scott, C., Bellala, G. and Willett, R. (2009). The false discovery rate for statistical pattern recognition. *Electronic Journal of Statistics* **3**, 651–677.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability. Chapman and Hall, London, New York.

Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of Statistics* **31**, 2013–2035, doi:10.1214/aos/1074290335.

Storey, J. D., Taylor, J. E. and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of The Royal Statistical Society Series B (Statistical Methodology)* **66**, 187–205, doi:10.1111/j.1467-9868.2004.00439.x.

Sugiyama, M., Kanamori, T., Suzuki, T., Hido, S., Sese, J., Takeuchi, I. and Wang, L. (2009). A density-ratio framework for statistical data processing. *IPSJ Transactions on Computer Vision and Applications* **1**, 183–208.

Tsui, C. S., Gan, J. Q. and Hu, H. (2011). A self-paced motor imagery based brain-computer interface for robotic wheelchair control. *Clinical EEG and Neuroscience* **42**, 225–229.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, Berlin.

Vapnik, V. N. (1998). *Statistical Learning Theory, Adaptive and Learning Systems for Signal Processing, Communications, and Control*. Wiley, Chichester.

Williamson, J., Murray-Smith, R., Blankertz, B., Krauledat, M. and Müller, K. R. (2009). Designing for uncertain, asymmetric control: interaction design for brain-computer interfaces. *International Journal of Human-Computer Studies* **67**, 827–841, doi:10.1016/j.ijhcs.2009.05.009.

Wolpaw, J. R. and Wolpaw, E. W. (Eds) (2012). *Brain-Computer Interfaces: Principles and Practice*. Oxford University Press, New York. ISBN-13: 978-0195388855.

