Stockholms universitet

Research Report

Department of Statistics

No. 2013:5

Empirical evaluation of sparse classification boundaries and HC-feature thresholding in high-dimensional data

Annika Tillander

Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden

Empirical evaluation of sparse classification boundaries and HC-feature thresholding in high-dimensional data

Annika Tillander

October 31, 2013

Department of Statistics, Stockholm University
S-106 91 Stockholm, Sweden
E-mail: annika.tillander@stat.su.se

Abstract

The analysis of high-throughput data commonly used in modern applications poses many statistical challenges, one of which is the selection of a small subset of features that are likely to be informative for a specific project. This issue is crucial for success of supervised classification in very high-dimensional settings with sparsity patterns. In this paper, we derive an asymptotic framework that represents a sparse and weak blocks model and suggest a technique for block-wise feature selection by thresholding. Our procedure extends standard Higher Criticism (HC) thresholding to the case where the dependence structure underlying the data can be taken into account and is shown to be optimally adaptive, i.e. it performs well without knowledge of the sparsity and weakness parameters. We empirically investigate the detection boundary of our HC procedure and performance properties of some estimators of sparsity parameters. The relevance and benefits of our approach for high-dimensional classification is demonstrated using both simulation and real data.

Keywords: Higher criticism, detection boundary, high dimensionality, supervised classification, separation strength.

Introduction

The last decades′ technical advancements have created an abundance of high-dimensional data where the number of features, p, greatly exceeds the number of units, n, e.g. signal detection, image processing and RNA sequencing. In this type of high-dimensional data it has often been observed that out of the many features, in practice only a few are informative for a current project. For example, in gene expression data, only some genes demonstrate significant differences in the expression level between tumor and normal tissue, i.e. such genes that can be informative for classification are very sparse; see e.g. [19]. Further, the separation strength of those informative features turns out to be relatively low. These types of settings, usually described by the sparse and weak (SW) model, imply an especially challenging statistical problem in identification of the informative features.

For the SW model, with the proportion of informative features (*sparsity*) and the separation strength (*weakness*) heading towards zero with growing p, the possibility of detection has been extensively studied for the case of Gaussian data with independent features; see e.g. [11, 30, 4, 12]. In particular, it has been shown that the likelihood-ratio-based procedure designed for detecting informative features behaves differently depending on the sparsity and weakness parameters [11], exhibiting so-called *detection boundary* phenomena. The main results discussed in e.g. [27, 13] deal with the detection boundary for the likelihood ratio test (LRT) under different types of assumptions e.g. heterogeneous and heteroscedastic Gaussian mixtures.

When the goal is to select a subset of highly discriminative features that will be used in a classification model, a selection threshold can be identified. It is well known that, for testing hypothesis such that the separation strength of a feature is non-zero, the LRT is optimal in the sense of minimum error. Accordingly, a threshold based on the LRT procedure will be an ideal choice. However, the LRT requires knowledge of the parameters, in our case sparsity and weakness parameters. A thresholding procedure which performs asymptotically as well as the LRT but does not require parameter knowledge (i.e. optimally adaptive) is based on Higher Criticism (HC) [4, 5, 6, 27]. The concept of HC was originally introduced by [28] as an approach to multiple testing using p-value-based second level test statistics. The HC approach has been considered in the literature (not only for classification but also e.g. signal detection), however mostly under the assumption of independent Gaussian distributed features; see e.g. [17, 15].

In this study we extend the consideration to the case where dependence is allowed within groups of features (blocks) and suggest block-wise feature selection under properly adjusted sparsity and weakness assumptions (SWB). Further, we derive detection boundaries for reliable feature identification and successful classification in a high-dimensional asymptotic framework, which links p and n with sparsity and weakness parameters of SWB model. We modify the HC technique by Donoho et al. [4, 5, 6] to a more general case where the dependence structure underlying the data can be taken into account and suggest block Higher Criticism (bHC) thresholding.

The remainder of the paper is structured as follows. First, we give a background to linear classification allowing for block-wise dependence structure and introduce block-separation strength and the distributional properties of the sample based separation score. In Section 2, we consider the asymptotic sparse and weak block model (ASWB). In Section

3, we suggest a bHC thresholding procedure. In Section 4, we empirically investigate detection boundaries for ASWB. In Section 5, we evaluate performance accuracy of the estimator of the sparsity parameter. In Section 6, we compare our bHC threshholds to other commonly employed selection strategies. In Section 7, we evaluate the effect of block selection on misclassification for synthetic and real data. Finally, we present some concluding remarks in Section 8.

1 Supervised classification

In supervised classification outcome measurements, e.g. tumor tissue v.s. non-tumor tissue, exist that we wish to predict based on a set of features. In the *training* data the outcome (i.e. class variable) and the features have been observed for a set of objects. Using the training data a model or algorithm can be built that enables prediction of the outcome for new observations where only the features have been observed.

1.1 Notation and optimal classification with known parameters

Let an observation $\mathbf{x}$ represent a set of features $(x_1, \ldots, x_p)$, then n observations gives

$$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \Re^p.$$

In a supervised classification problem with $\mathcal{C}$ classes, each observation from the training data is known to belong to some class $y_j = c$ where $j = 1, \ldots, n$, $c \in \{1, \ldots, \mathcal{C}\}$ and the training data is described by

$$\mathcal{T} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}.$$

Assume that the outcome in each class is modeled by the Gaussian distribution, i.e. $\mathbf{x}_c \sim N(\boldsymbol{\mu}_c, \Sigma_c)$, where $\boldsymbol{\mu}_c$ is the class mean and Σ_c is the class-wise covariance matrix. Let $\mathcal{G}$ be a decision rule with decision regions $\Omega_c \in \Re^p$, $\Omega_c = \mathcal{G}^{-1}(c)$, which assigns an observation $\mathbf{x}$ to the class with the highest value of the linear function $D_c(\mathbf{x})$, i.e. $\mathcal{G}(\mathbf{x}) = c^*$ if $c^* = \text{argmax}_{c=1,\ldots,\mathcal{C}} D_c(\mathbf{x})$, where

$$D_c(\mathbf{x}) = \mathbf{x}'\Sigma_c^{-1}\boldsymbol{\mu}_c - \frac{1}{2}\boldsymbol{\mu}_c'\Sigma_c^{-1}\boldsymbol{\mu}_c + \log \pi_c, \tag{1.1}$$

where π_c is the prior probability of class c and $\sum_{c=1}^{\mathcal{C}} \pi_c = 1$. This classifier is analogous to the well-known *Fisher linear discriminant function* that is optimal in a sense of minimum overall misclassification probability defined as $\varepsilon = \sum_{c=1}^{\mathcal{C}} \pi_c P(\mathcal{G}(\mathbf{x}) \neq c | \mathbf{x} \in \Omega_c)$.

We will focus on two-class classification problems with equal class-wise covariance matrices (Σ). The two-class linear function can then be represented as

$$D(\mathbf{x}) = \left(\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right)' \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{1.2}$$

and an observation $\mathbf{x}$ is assigned to class 1 if $D(\mathbf{x}) \geq \log \frac{\pi_2}{\pi_1}$, or, otherwise, to class 2. If $\pi_1 = \pi_2 = 1/2$ the optimal misclassification probability can be expressed as

$$\varepsilon_{opt} = \Phi\left(-\frac{1}{2}\frac{\mathrm{E}\left[D(\mathbf{x})|\mathbf{x} \in \Omega_1\right]}{\sqrt{\mathrm{Var}\left[D(\mathbf{x})|\mathbf{x} \in \Omega_1\right]}}\right) = \Phi\left(-\frac{1}{2}\sqrt{\delta^2}\right) \tag{1.3}$$

where $\Phi(\cdot)$ is the Gaussian cumulative distribution function and $\delta^2 = \left\|\Sigma^{-1/2}\boldsymbol{\mu}\right\|^2$ is the Mahalanobis shift vector norm, where $\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ is a shift vector and $\|\cdot\|$ denotes the ℓ^2 norm.

1.2 Quantifying and estimating feature separation strength

A very important question when designing a procedure for selecting a subset of features is how to measure the separation power of a feature or a set of features in a classification framework. In this study we focus on δ^2 as such a measure. The motivation for this choice follows from (1.3) as ε_{opt} is a function of δ^2, i.e. the distance between classes. Since Φ is a monotone strictly decreasing function of $\sqrt{\delta^2}$, we can say that the separation power of a feature or a set of features is its contribution to this distance. We will need to bound the Mahalanobis shift vector norm in order to guarantee that the classifier does not degenerate, and make the assumption that there exist such constants K that $0 < K_1 \leq \delta^2 \leq K_2 < \infty$. Further we consider the separation strength under the SW model where the following hold:

- *Informative features are sparse*: The non-zero elements in the shift vector $\boldsymbol{\mu}$ are only $\tilde{p}$ out of p where $\beta = \frac{\tilde{p}}{p}$ are close to zero.
- *Informative features are weak*: The non-zero elements in the shift vector $\boldsymbol{\mu}$ have a common amplitude of μ which is small.

1.2.1 Separation strength under independence

For the classification problem when $p \gg n$ a very popular approach is to simply ignore feature covariances and standardize the features to mean 0 and variance 1, i.e. let the feature covariance matrix be the identity matrix (I_p); see e.g. [5]. Hence the separation strength will be defined for each single feature as $\delta_I^2 = \boldsymbol{\mu}' I_p \boldsymbol{\mu}$ and we define the rescaled estimate of the i features separation strength as

$$Z_i = \frac{\hat{\mu}_{1,i} - \hat{\mu}_{2,i}}{\sqrt{n}}, \tag{1.4}$$

where $\hat{\mu}_{c,i}$ is the sample class mean of the i feature for $i = 1, \ldots, p$. These Z_is can be interpreted as Z-scores of the following two sided test $H_{0,i} : Cov(y, \mathbf{x}_i) = 0$, where $\mathbf{x}_i = (x_1, \ldots, x_n)$ is the ith feature vector and with the given assumptions $Z \sim N(\theta, I_p)$ where $\theta = \sqrt{n}\mu$ and μ is the feature shift vector. According to [6] features with significantly non-zero Z_i have non-zero μ_i while other features will have Z_i values which are consistent with the null hypothesis $\mu_i = 0$. Then, selecting features with Z-score above a threshold makes sense, in Section 3 we will show how this threshold can be found in the SW model. Observe that θ provides the same information as the square root of δ^2 under the independence and unit variance.

1.2.2 Separation strength under block-diagonal dependence structure

It is obvious that the assumption of independence, while essentially simplifying the estimation problem, is not credible in real classification. Another approach to be considered is to learn the covariance structure of the data; see e.g. [9, 14, 16, 21, 18], with such

constraint that the number of features to be estimated is reduced e.g. by the assumption that the true underlying covariance structure is sparse. In this context sparse means that there are only a few features that are highly correlated with any given fixed feature. Observe that this kind of sparsity has a natural biological explanation, e.g. some groups of genes acting together in a way that are associated with the clinical outcome. This type of pattern can either come from biological expertise or simply be suggested by the underlying model.

In [9] the so called gLasso was introduced, a method to learn the sparsity patterns of the empirical feature covariances matrix ($\hat{\Sigma}$) using a regularization parameter λ. Then in [18] the gLasso-based covariance structure learning technique was suggested. It is a two-step procedure that combines gLasso with the Cuthill-McKee ordering ([3]). Its result is a block-diagonal structure approximation. The corresponding estimator of the inverse covariance matrix is given by

$$\hat{\Sigma}_{\lambda}^{-1} = \text{diag}\left[\hat{\Sigma}_{\lambda,1}^{-1}, \ldots, \hat{\Sigma}_{\lambda,b}^{-1}\right], \tag{1.5}$$

where b is the number of blocks. The existence of $\hat{\Sigma}_{\lambda,i}^{-1}$ for each i ($i = 1, \ldots, b$) is ensured by the constraint in block size $p_0 < n - 2$ imposed in [18].

Such a block diagonal segmentation of Σ^{-1} represents independence between corresponding groups of features. This in turn means that both the class means $\boldsymbol{\mu}_c$ and the observed vector $\mathbf{x}$ can be partioned into b disjoint subsets $\boldsymbol{\mu}_{c,i} = (\mu_{c,i_1}, \ldots, \mu_{c,i_{p_i}})$ and $\mathbf{x}_i = (x_{i_1}, \ldots, x_{i_{p_i}})$, ($\mathbf{x}_i \in \Re^{p_i}$) $i = 1, \ldots, b$, such that for any $i \neq j$, $\mathbf{x}_i$ and $\mathbf{x}_j$ are conditionally independent given the class variable $\mathbf{y}$. Then the two-class linear function (1.2) will have an additive structure

$$D(\mathbf{x}) = \sum_{i=1}^{b} \Big(\mathbf{x}_i - \frac{1}{2}(\boldsymbol{\mu}_{1,i} + \boldsymbol{\mu}_{2,i})\Big)' \Sigma_i^{-1} \Big(\boldsymbol{\mu}_{1,i} - \boldsymbol{\mu}_{2,i}\Big); \tag{1.6}$$

$\boldsymbol{\mu}_{c,i}$ is the class mean vector and Σ_i^{-1} is the inverse covariance matrix of the ith block. In this study we limit ourselves to equal block size, denoted p_0. Now we can define the ith block separation strength as $\delta_i^2 = \left\|\Sigma_i^{-1/2}\boldsymbol{\mu}_i\right\|^2$, i.e. as its contribution towards the total distance $\delta^2 = \sum_{i=1}^{b} \delta_i^2$. If δ_i^2 is known one can derive the feature selection or thresholding procedure by introducing an indicator function into $D(\mathbf{x})$

$$D(\mathbf{x}) = \sum_{i=1}^{b} \mathbf{1}_{\{\psi,\infty\}}(\delta_i^2) \Big(\mathbf{x}_i - \frac{1}{2}(\boldsymbol{\mu}_{1,i} + \boldsymbol{\mu}_{2,i})\Big)' \Sigma_i^{-1} \Big(\boldsymbol{\mu}_{1,i} - \boldsymbol{\mu}_{2,i}\Big), \tag{1.7}$$

where ψ is a threshold. The threshold ψ can be optimized by minimizing the misclassification probability of the $D(\mathbf{x})$. However the question remains of how to choose ψ.

Let the rescaled estimator of the ith block separation strength be defined as

$$\mathcal{S}_i^2 = \eta \hat{\boldsymbol{\mu}}_i' \hat{\Sigma}_i^{-1} \hat{\boldsymbol{\mu}}_i, \tag{1.8}$$

where $\eta = \frac{n_1 n_2}{n}$, $\hat{\boldsymbol{\mu}}_i = \hat{\boldsymbol{\mu}}_{1i} - \hat{\boldsymbol{\mu}}_{2i}$ is the shift vector of the sample class means and $\hat{\Sigma}_i$ is the maximum likelihood estimate of the covariance matrix of the ith block. In order to select

those blocks that are informative to classification, we need to specify the distribution of $\mathcal{S}_i^2$ and construct a test in a similar way as in section 1.2:

$$\mathcal{S}_i^2 = \eta \hat{\boldsymbol{\mu}}_i' \Sigma_i^{-1} \hat{\boldsymbol{\mu}}_i \frac{\hat{\boldsymbol{\mu}}_i' \hat{\Sigma}_i^{-1} \hat{\boldsymbol{\mu}}_i}{\hat{\boldsymbol{\mu}}_i' \Sigma_i^{-1} \hat{\boldsymbol{\mu}}_i}. \tag{1.9}$$

Since for the Gaussian class conditional distribution $\hat{\boldsymbol{\mu}}_i \sim N(\boldsymbol{\mu}_i, \eta^{-1}\Sigma_i)$, it follows that $\eta \hat{\boldsymbol{\mu}}_i' \Sigma_i^{-1} \hat{\boldsymbol{\mu}}_i \sim \chi^2\left(p_0, \eta\delta_i^2\right)$, where $\chi^2(p_0, \eta\delta_i^2)$ denotes the non-central χ^2 distribution with p_0 degrees of freedom and non-centrality parameter $\eta\delta_i^2$. Under the same assumptions it holds that $\hat{\Sigma}_i \sim W_{p_0}\left(\Sigma_i, n-2\right)$, where W denotes the Wishart distribution with scale matrix Σ_i and $n-2$ degrees of freedom. Hence for a given $\hat{\boldsymbol{\mu}}_i$

$$\frac{\hat{\boldsymbol{\mu}}_i' \Sigma_i^{-1} \hat{\boldsymbol{\mu}}_i}{\hat{\boldsymbol{\mu}}_i' \hat{\Sigma}_i^{-1} \hat{\boldsymbol{\mu}}_i} \sim \chi^2(n - p_0 - 1), \tag{1.10}$$

where $\chi^2(n-p_0-1)$ is the central χ^2 distribution with $n-p_0-1$ degrees of freedom. Observe that the resulting distribution in (1.10) does not involve $\hat{\boldsymbol{\mu}}_i$ which implies that the ratio in (1.9) is distributed independently on $\hat{\boldsymbol{\mu}}_i$. Now observe that (1.10) can be represented as the ratio of the non-central $\chi^2\left(p_0, \eta\delta_i^2\right)$ to an independent central $\chi^2(n-p_0-1)$, from which it follows that by using a proper standardization for both χ^2 distributions we obtain

$$\frac{n-p_0-1}{n-2} \frac{\mathcal{S}_i^2}{p_0} \sim F\left(p_0, n-p_0-1, \eta\delta_i^2\right) \tag{1.11}$$

.

Further, for a fixed p_0, $\omega_i^2 = \eta\delta_i^2$ and assuming that $0 < \omega_i^2 < \infty$ (constrained in n)

$$p_0 F\left(p_0, n-p_0-1, \eta\delta_i^2\right) \to \chi^2\left(p_0, \omega^2\right) \text{ as } n \to \infty \tag{1.12}$$

by the asymptotic properties of the non-central F distribution. In this study we will focus on the χ^2 distribution due to the advantage of fewer parameters.

Now the $\mathcal{S}_i^2$ can be interpreted as the separation strength of the following one sided test $H_{0,i} : \boldsymbol{\mu}_i' \Sigma_i^{-1} \boldsymbol{\mu}_i = 0$ for $i = 1, \ldots, b$ with the given assumption $\mathcal{S}_i^2 \sim \chi^2(p_0, \omega^2)$ where $\omega^2 = \eta\delta^2$. Then blocks with non-zero δ_i^2 typically have significantly non-zero $\mathcal{S}_i^2$, while most other blocks will have $\mathcal{S}_i^2$ values largely consistent with the null hypothesis $\delta_i^2 = 0$. In Section 3 we suggest how a threshold can be found for the SWB that enables us to select blocks with significantly non-zero $\mathcal{S}_i^2$.

2 A model with sparse and weak features

In order to describe the sparse and weak classification model we adopt an asymptotic framework where the number of features is the driving variable, whereas the sparsity and weakness are parameterized as functions of p. First, we present the general case as can be seen in e.g. [27, 6], then extend to the block model. Observe that the weakness condition here will be related to the boundary condition of δ^2. In a high-dimensional setting we assume that number of objects n and number of features p will tend to infinity in a linked fashion, where n will remain small in comparison to p.

Asymptotic sparse and weak model (ASW)

A1: Asymptotics relating p to n. In the sequence of classification problems, both the number of features p and the number of objects n diverges to ∞ [6].

A2: Rate of growth for the number of features. In the sequence $n = \mathcal{O}(\log p^{\kappa})$ for some constant $\kappa > 0$ where $\mathcal{O}$ denotes order, i.e. $p \gg n$.

A3: As sparsity increases, β tends to 0: $\beta = p^{-\gamma}$, where $\gamma \in (0, 1)$.

A4: The separation strength decreases. The feature separation strength is parameterized as $\theta = \sqrt{2r \log p}$, where $r \in (0, 1)$.

Asymptotic sparse and weak block model (ASWB)

In the asymptotic framework for SWB it is more convenient to set $n_1 = n_2$ and the block-size $p_i = p_0$ for all $i = 1, \ldots, b$, where the latter is based on a class of asymptotically equivalent block-structure approximations where for each block i we assume that $p_i/n \to 0$ as $n \to \infty$. Then the asymptotic sparse and weak block model is represented by the following set of assumptions:

B1: Asymptotics relating b to n. In the sequence of classification problems with fixed p_0, both b and n diverges to ∞ [18].

B2: Rate of growth for the number of blocks. In the sequence $n = \mathcal{O}(\log b^{\kappa})$ for some constant $\kappa > 0$, i.e. the number of blocks can grow faster than n [18].

B3: The sparsity increases. As p tends to ∞ the fraction of informative blocks β behaves as $\beta = b^{-\gamma}$ where $\gamma \in (0, 1)$.

B4: The separation strength decreases. The block separation strength is parameterized as $\omega^2 = 2r \log b$, where $r \in (0, 1)$.

In the asymptotic setting the models under consideration can commonly be expressed as ASW$(r, \gamma; \kappa)$. The driving parameters are r and γ, since they relate the rareness and usefulness of the features/blocks in the observed vector, while parameter κ is of less importance and can be seen as incidental. Therefore we will simply write ASW(r, γ) and ASWB(r, γ) respectively. To illustrate how the sparse and weak setting differs from the more traditional setting with substantial number of informative features (*dense*) that can be highly discriminatively (*strong*), see Figure 1.

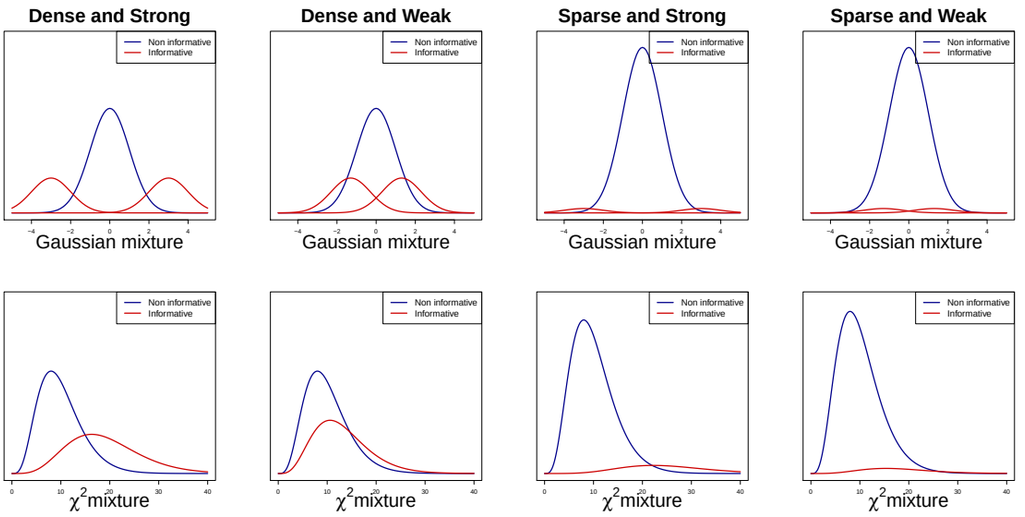

Figure 1: Density of estimated separation strength for a mixture of informative and non-informative features.

3 HC feature thresholding for sparse and weak classification problem

The problem of selecting informative features can now, for a fixed parameter γ, be formulated as follows: Given p independent features with the separation scores $Z_1, Z_2, \ldots, Z_p$, we suppose that Z_i has the probability β to be informative, i.e. having the shift parameter $\theta > 0$, where θ is calibrated as in A4. The feature selection is recast as a hypotheses testing problem:

$$\begin{aligned} H_0: &\quad Z_i \sim N(0,1) \text{ i.i.d } 1 \le i \le p \\ &\quad \text{versus } p\text{-dependent alternatives} \\ H_1: &\quad Z_i \sim (1-\beta)N(0,1) + \beta N(\theta,1) \text{ i.i.d } 1 \le i \le p \end{aligned} \tag{3.1}$$

where $\beta = p^{-\gamma}$ as defined in A3, and H_0 means that no features are informative whereas $H_1^{(p)}$ means that features are generated from a mixture of β informative and $1-\beta$ non-informative features.

To test the hypothesis we turn to the Higher Criticism (HC) approach which has been developed for optimal detection of informative features within the sparse and weak independent feature model by [4, 5, 6]. It is a three-step testing procedure:

I: The procedure starts with obtaining a p-value (π) for each observed separation score by $\pi_i = P_{H_0}\{|Z_i| \ge |z_i|\} = 2(1 - \Phi(|z_i|))$. Then the p-values are ranked in increasing order: $\pi_{(1)} \le \pi_{(2)} \le \ldots \le \pi_{(b)}$.

II: The HC objective function is defined as

$$\mathrm{HC}_{i,\pi_{(i)}} = \sqrt{p}\frac{i/p - \pi_{(i)}}{\sqrt{i/p(1-i/p)}}, \tag{3.2}$$

where $i = 1, \ldots, p$. For a fixed $\alpha_0 \in (0, 1)$, the HC test statistic is

$$\mathrm{HC}^* = \max_{1 \le i \le (\alpha_0 \times p)} \mathrm{HC}_{i,\pi_{(i)}}. \tag{3.3}$$

In rare and weak situations HC seems insensitive to the selection of α_0 and a common choice is $\alpha_0 = 0.1$ see [6]. The HC objective function ($\mathrm{HC}_{i,\pi_{(i)}}$) and the absolute separation scores $|Z|_{(i)}$ are ordered correspondingly to the p-values.

III: The last step is the thresholding. Let HC^* be achieved at index i^*; the HC thresholding (HCT) is then the value $t^{*HC} = |Z|_{(i^*)}$. Features with Z-scores exceeding t^{*HC} in magnitude are then selected as informative.

We adopt this method for block selection and adjust for ASWB(r, γ). Now, again, for a fixed parameter γ we formulate the problem of block-wise feature selecting as follows: Given b independent blocks with observed block separation strength $\mathcal{S}_1^2, \mathcal{S}_2^2, \ldots, \mathcal{S}_b^2$, we suppose that the $\mathcal{S}_i^2$ has the probability β to be informative, i.e. having the non-centrality parameter $\omega^2 > 0$ where ω^2 is calibrated as in B4. We recast the block threshold selection into the hypotheses testing problem:

$$\begin{aligned} H_0: \quad & \mathcal{S}_i^2 \sim \chi^2_{p_0}(\cdot; 0) \text{ i.i.d, } 1 \le i \le b \\ & \text{versus } b\text{-dependent alternatives} \\ H^{1:} \quad & \mathcal{S}_i^2 \sim (1-\beta)\,\chi^2_{p_0}(\cdot; 0) + \beta \chi^2_{p_0}\left(\cdot; \omega^2\right) \text{ i.i.d, } 1 \le i \le b, \end{aligned} \tag{3.4}$$

where $\beta = b^{-\gamma}$ as in B3. In words, the goal of testing the hypothesis problem is as follows: we model non-informative blocks as samples from $\chi^2_{p_0}(\cdot; 0)$ and informative blocks as samples from $\chi^2_{p_0}(\cdot; \omega^2)$. Then β can be interpreted as a proportion of informative blocks with ω^2 representing the block separation strength. Using this interpretation the problem (3.4) can be equivalently stated as testing the hypothesis $\beta = 0$.

For b observed separation strength the p-value for each block is obtained as $\pi_i = P_{H_0}\{\mathcal{S}_i^2 \ge s_i^2\} = \bar{\chi}^2_{p_0}(s_i^2; 0)$, where $\bar{\chi}^2_{p_0}(\cdot; 0) = 1 - \chi^2_{p_0}(\cdot; 0)$ is the survival function of $\chi^2_{p_0}(\cdot; 0)$. After the p-values have been ranked, the block higher criticism (bHC) objective function is calculated as

$$\mathrm{bHC}_{i,\pi_{(i)}} = \sqrt{b}\frac{i/b - \pi_{(i)}}{\sqrt{i/b(1 - i/b)}}, \tag{3.5}$$

where $i = 1, \ldots, b$ and for fixed $\alpha_0 \in (0, 1)$ the bHC test statistic is

$$\mathrm{bHC}^* = \max_{1 \le i \le (\alpha_0 \times b)} \mathrm{bHC}_{i,\pi_{(i)}}, \tag{3.6}$$

In words, $\mathrm{bHC}_{i,\pi_{(i)}}$ (3.5) resembles the standard discrepancy between the expected and the observed behavior of a bulk of $\pi_{(i)}$ obtained from (3.4) and is used here for assessing the significance of the whole bulk of p-values. When this discrepancy is large, we reject H_0 since the whole set of p-values is not consistent with the null hypothesis. Now we consider the feature thresholding using the bHC testing. Given the collection of p-values from the test problem (3.4) suppose that the bHC in (3.6) is achieved at index i^*. Then the block higher criticism threshold, bHC threshold, is the value of $\tau^* = \mathcal{S}_{i^*}^2$ and the blocks whose separation strength $\mathcal{S}_i^2$ exceeds τ^* in magnitude are selected as informative for the classification. Figure 2 illustrates the thresholding technique. A and B present a probability plot(PP plot) of the ordered separation scores and p-values plotted against

$\frac{i}{p}/\frac{i}{b}$ respectively, and C illustrates the threshold value that corresponds to i^* yielding the largest discrepancy in (3.5).

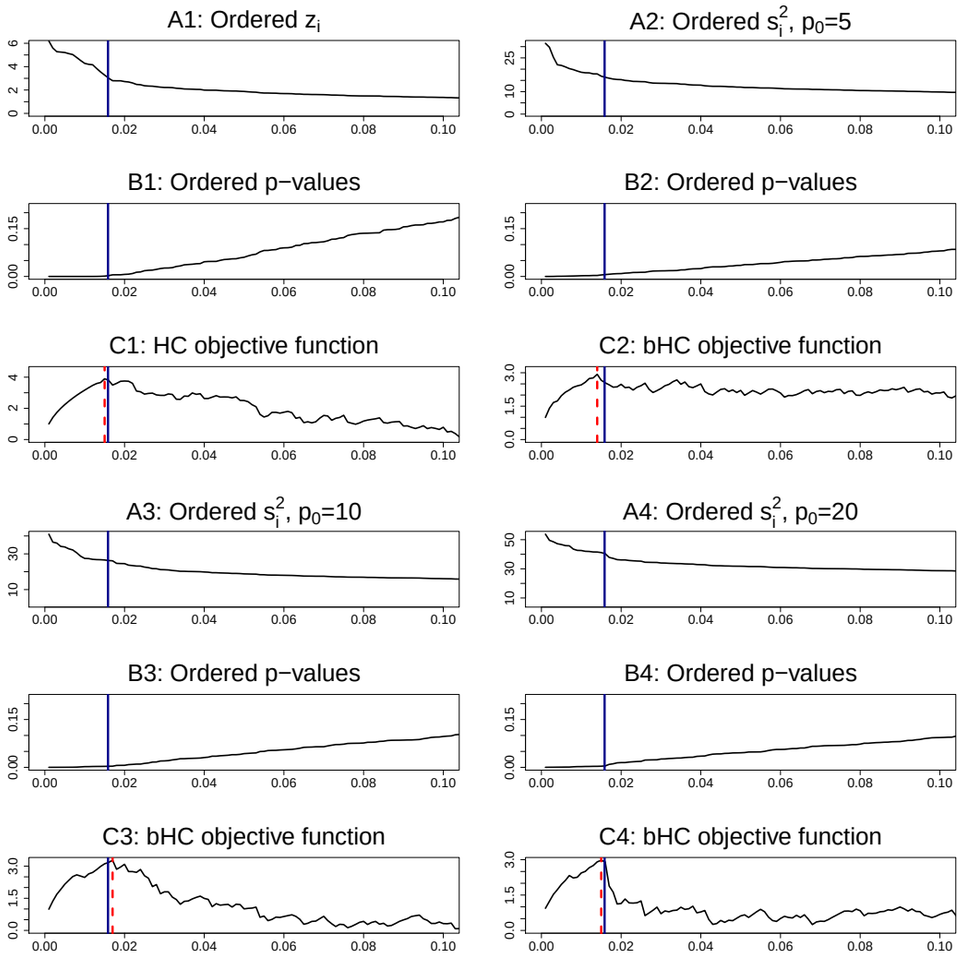

Figure 2: Illustration of HC and bHC thresholding on the x-axes i/p for 1 and i/b for 2:4. A) The ordered separation scores $|Z|$ and $\mathcal{S}^2$. B) The corresponding ordered p-values. C) The HC objective function. The solid blue vertical line shows true β and the dotted red line the HC threshold.

4 Detection boundaries

Even though it is clearly seen that HC thresholding is a very beneficial and easily implemented procedure, it has certain boundaries. The problem of detecting informative features for the ASW has been extensively studied within statistical literature; see e.g. [11, 6, 30, 12]. Asymptotically successful identification of informative features are possible for many choices of (γ, r) whereas for another large family of choices it is impossible. The concept $\gamma - r$ plane have been introduced to bring understanding to this fact where the $\gamma - r$ plane is the two-dimensional domain of $\gamma < 1$ and $0 < r$. The domain is divided

by boundaries into regions with different possibilities of detection. First, we introduce the detection boundaries that have been presented for ASW and then we empirically investigate the detection boundary for block-diagonal dependence structure.

4.1 Under independence

For the $\gamma - r$ plane the focus is to characterize the so-called detection boundaries which are curves that partition the plane into regions of different possibilities of separating the informative features from non-informative features. When applying the likelihood ratio test (LRT) for testing, (3.1), it has been shown that there is an abrupt change in the behavior on the test depending on the choice of parameters $\gamma - r$, namely, the sum of type I and type II error tends to 1 in one part of $\gamma - r$ plane and 0 in another part [11]. These two areas can be described by ρ

If $r > \rho^*(\gamma)$, H_0 and H_1 separate asymptotically.

If $r < \rho^*(\gamma)$, H_0 and H_1 merge asymptotically.

The curve that partitions these two areas is known as the *detection boundary* function [11].

$$\rho_1^*(\gamma) = \begin{cases} \gamma - 1/2 & 1/2 < \gamma \leq 3/4 \\ \left(1 - \sqrt{1-\gamma}\right)^2 & 3/4 < \gamma \leq 1 \end{cases}$$

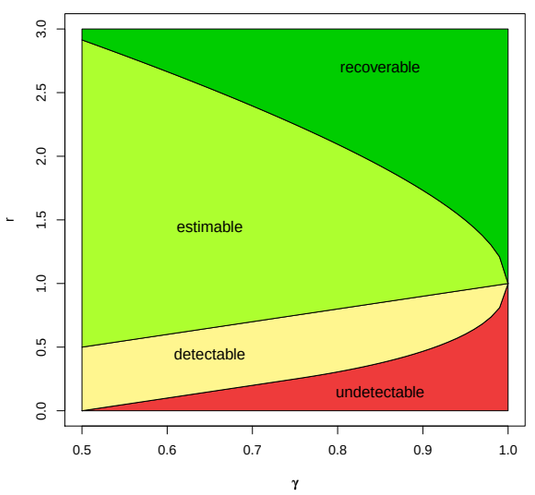

Figure 3: Detection boundary for signal identification within $\gamma - r$ plane for Gaussian case [15].

Further boundaries have been identified and the $\gamma - r$ plane can be divided into four different regions corresponding to the detection possibilities; see Figure 3 [15]. The second curve, when $1/2 < \gamma \leq 1$, is the *identification boundary* $\rho_2^*(\gamma) = \gamma$ reported by [4]. Above this boundary, information identification by thresholding is possible (estimable); directly

below it is only possible to detect information but not to estimate. The third curve is the *recovery boundary* $\rho_3^*(\gamma) = \left(1 + \sqrt{1-\gamma}\right)^2$ reported by [30, 12], where in the region above the curve almost all information can be completely identified.

4.1.1 Detection boundaries and HC

The LRT gives the detection boundaries assuming exact knowledge of sparsity and weakness parameters and, since it requires the knowledge of parameter values, it is less practical for real data where the parameters are often unknown. There are methods for estimating the parameters (see e.g. [17]) though HC does not require knowledge of γ and r, which simplifies the calculations. Also, it has been shown that HC allows the same performance accuracy as the LRT. In fact HC replicates the behaviors of the LRT; in the so-called detectable areas the informative features are reliable identified by HC thresholding. This property of HC, as earlier stated, is known as optimal adaptivity; see [4, 27].

4.2 Under block-diagonal structure

The main goal of our paper is how to select a subset of features that will improve classification accuracy, and we are interested in relating the ASWB to the classification task. Much of the work with hypothesis testing (3.1) in ASW is tightly connected to signal identification, i.e. detection of informative features as the main goal (see e.g. [27, 15]). Though it is fairly easy to see the connection between hypothesis testing and the classification task, this as ω^2 is calibrated according to B4 and is related to ε (1.3) through $\delta_i^2 = \frac{\omega_b^2(r;\gamma)}{c_n}$. Then, the combination of the sparsity and weakness parameters directly affects classification possibilities. As a basic example we let $\delta^2 = b^{-\gamma} b 2r \log(b)$ since $(1 - b^{-\gamma})\, b$ blocks do not contribute with any information to the total distance. As a constraint we bind the misclassification from below to a minimum of $\varepsilon > 1e^{-5}$ and calculate (ε), see Figure 4. The color corresponds to the misclassification probability and shows the area where it is possible to discriminate between two classes. This area is coherent to the area of possible detection of informative features.

As $\mathcal{S}_i^2 \sim \chi^2_{p_0}(\cdot, \omega^2)$ we are interested in detection boundaries for the χ^2-distribution. It has been shown that the optimal detection boundary for the χ^2-distribution with two degrees of freedom is asymptotically similar to the Gaussian case [13]. We empirically study how the detection problem for the χ^2-distribution behaves for growing numbers of degrees of freedom within the $r - \gamma$ plane. We use the LRT to identify the detection boundaries since the detectable region can be defined as where the LRT has the sum of type I and II error probabilities that tends to 0 as the number of features goes to infinity [27]. Hence we plot the sum of types I and II error for the LRT with all parameters known within the $r - \gamma$-plane.

We start by evaluating the possibility of detecting the presence of informative features, considering the stated hypotheses (3.4) and using the likelihood ratio (LR) where $LR_i = \frac{f_{H_1}(s_i^2)}{f_{H_0}(s_i^2)}$ where the likelihood corresponds to the density function of s_i^2, the observed separation strength for the ith block. Respectively the hypothesis gives

$$LR_i = \frac{(1-\beta)\, \chi^2_{p_0}\left(s_i^2; 0\right) + \beta \chi^2_{p_0}\left(s_i^2; \omega^2\right)}{\chi^2_{p_0}\left(s_i^2; 0\right)}. \tag{4.1}$$

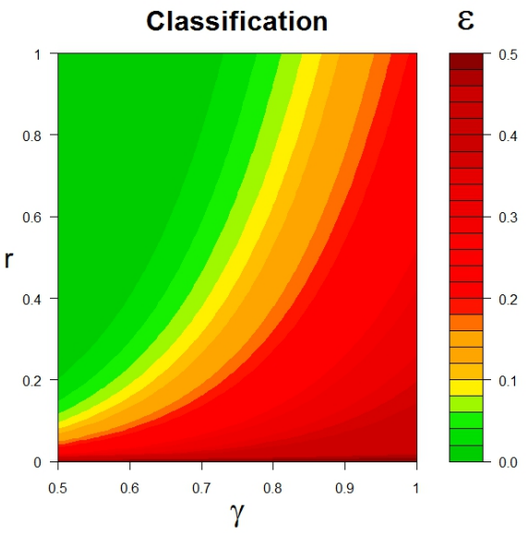

Figure 4: The relation between $r - \gamma$ and the misclassification probability (ε) for discriminating between to classes.

For b blocks $LR_B = LR_B\left(s_1^2, s_2^2, \ldots, s_b^2; \gamma, r\right)$, then we consider the LRT which reject H_0 if and only if

$$\log(LR_B) > 0. \tag{4.2}$$

We consider two different scenarios to demonstrate the behavior of the detection regions $\mathcal{S}_i^2$. The fact that blocks were identified by the covariance structure does not guarantee that the features merged have the same separation strength. As stated before, blocks can also be built on expert knowledge in the specific area, resulting in highly informative but independent features constructing blocks. In order to represent a variety of situations for the distribution of $\mathcal{S}_i^2$ over the of blocks we regard these scenarios that represent the two extremes:

Scenario I

Scenario I is the restrictive approach that regards a block as one unit minimizes separation strength in relation to block structure. A whole block does not provide more information than one single feature. The separation strength only depends on the weakness parameter r and number of blocks; block size has no effect. In this setting we generate a separation score from the χ^2-distribution.

Scenario II

Scenario II is the liberal approach that regards the information of each feature within the block and maximizes the separation strength in relation to block structure. For the informative block the separation strength increases for each added feature and grows with degrees of freedom. In this setting we generate the score g from the Gaussian distribution and then calculate the separation score as $s_i = \mathbf{g}'_{p_0} I_{p_0} \mathbf{g}_{p_0}$ where $\mathbf{g}_{p_0}$ is a vector of Gaussian

scores with the length of the block size and I_{p_0} is the identity matrix with the same size as the block.

To evaluate where it is possible to detect the presence of informative features within the $\gamma - r$ plane, we start by conducting a Monte Carlo simulation for **Scenario I**, with 160 true null hypotheses as $s_0 \sim \chi^2(p_0, 0)$ under H_0 (3.4) for $b = 100000$, $\gamma = (0.5 : 1)$ with step 0.01, $r = (0 : 1)$ with step 0.02 and $p_0 = [5, 10, 20]$ and 160 true alternative hypotheses as $s_1 \sim (1 - \beta)\chi^2(p_0, 0) + \beta\chi^2(p_0, \omega^2)$ under H_1 (3.4) for $b = 100000$, $\gamma = (0.5 : 1)$ with step 0.01, and $r = (0 : 1)$ with step 0.02 and $p_0 = [5, 10, 20]$. For each hypothesis the LR_B is estimated, H_0 is rejected if $LR_B > 0$ and the errors are calculated. A 95% confidence interval for the sum of errors over the 320 hypotheses is calculated and the upper bound displayed as a contour plot for each degree of freedom in the $r - \gamma$ plane; see the top row of Figure 5.

Next, we conduct a Monte Carlo simulation for **Scenario II**, again with 160 true null hypotheses but now with $g_0 \sim N(0, 1)$ under H_0 for $p = p_0 10000$, $\gamma = [0.5 : 1]$ with step 0.01 and $r = [0 : 1]$ with step 0.02 and (3.1) and 160 true alternative hypotheses as $g_1 \sim (1 - p^{-\gamma})N(0, 1) + p^{-\gamma}N(\sqrt{(}2r \log(p)), 1)$ under H_1 (3.1) for $p = p_0 10000$, $\gamma = [0.5 : 1]$ with step 0.01 and $r = [0 : 1]$ with step 0.02. Then we calculate the separation strength for the following different blocks size: $p_0 = [5, 10, 20]$. For each hypothesis, the LR_B is estimated, H_0 is rejected if $LR_B > 0$ and the errors are calculated. A 95% confidence interval for the sum of errors over the 320 hypotheses is calculated and the upper bound displayed as contour plot for each degree of freedom in the $r - \gamma$ plane; see bottom row of Figure 5.

As can be seen in Figure 5, for the restrictive approach of Scenario I the detectable area is smaller than for the Gaussian distribution, and it also decreases with a growing number of degrees of freedom. This indicates that if blocks do not contribute with more information than a separate single feature the set of informative features needs to be less sparse and weak to be detectable. In the case of the liberal approach the detection pattern is similar to that of the Gaussian distribution and, as expected, the detection area is somewhat larger for larger blocks. This indicate that if the blocks are constructed from informative features the detectable region is rather intact despite the change in distributional properties.

The previous gives where it is possible to detect presence of informative features. The next step is to evaluate where it is possible to identify the informative features, i.e. we wish to test the hypotheses.

$$H_{0,i} : \mathcal{S}_i \sim \chi^2_{p_0}(\cdot; 0) \text{ versus } H_{1,i} : \mathcal{S}_i \sim \chi^2_{p_0}(\cdot; \omega^2) \tag{4.3}$$

This gives the following:

$$LR_i = \frac{\chi^2_{p_0}\left(s_i^2; \omega^2\right)}{\chi^2_{p_0}\left(s_i^2; 0\right)}. \tag{4.4}$$

Ratios are then ranked in increasing order, $LR_1 < \ldots < LR_b$. We use three different approaches for rejecting the null hypothesis. For the first LRT we consider the empirical distribution of the LR and the top β-quantile $\left(\hat{\ell}\right)$ is used as the reference value for rejecting H_0. Since we know the true proportion of informative blocks and when the two hypotheses are well-separated, the top β-quantile corresponds to a true alternative

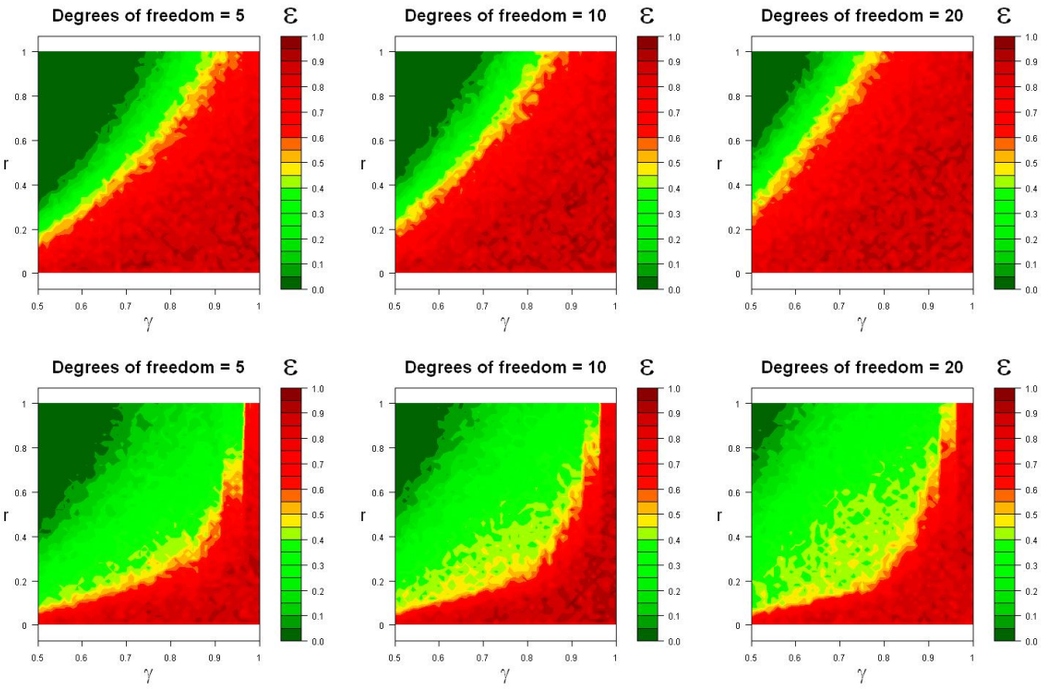

Figure 5: The boundaries for detecting the presence of informative features. The first row corresponds to data from Scenario I and the second row to data from Scenario II.

hypothesis. For the second LRT we widen the cutoff by allowing a size α type I error and consider $\alpha + \beta$ proportion highest values of LR from the empirical distribution. For the third approach, we consider the LRT which rejects H_0 if and only if $\log(\text{LR}_i) > 0$, i.e. strictly favoring the alternative hypothesis when the probability density for the non-central χ^2 exceeds the probability density for the central χ^2. An example of the difference in cutoff between the three approaches can be seen in Figure 6.

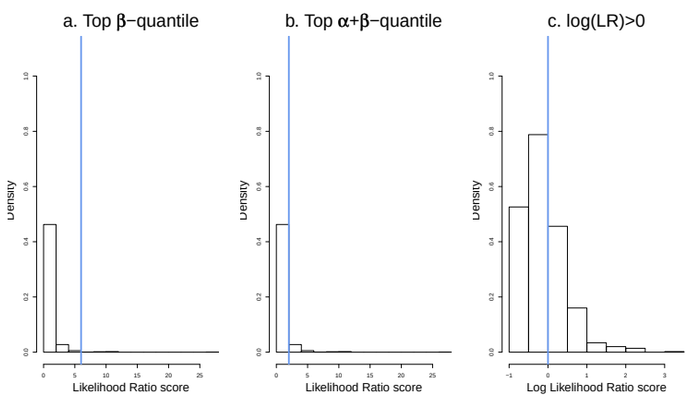

Figure 6: The empirical distribution for the likelihood ratio test, where the lines show the cutoff for rejecting the null hypothesis.

When the LR has been estimated for each block and the null hypothesis rejected according to the approaches described above, we can evaluate performance accuracy. Since the ultimate objective is to select a subset of features that will be useful for the classification task we want to identify as many truly informative as possible while keeping the noise (falsely informative features) to a minimum. Hence, we will estimate the errors separately as follows and then standardize the sum to 1. We calculate the Attained Significance Level (ASL) representing the type I error as

$$\begin{aligned} ASL^1\left(r,\gamma,LR\right) &= \frac{\#of\left\{LR > \hat{\ell}_\beta | H_0 \text{ is true } \right\}}{\left(1-b^{-\gamma}\right)b} \\ ASL^2\left(r,\gamma,LR\right) &= \frac{\#of\left\{LR > \hat{\ell}_{(\alpha+\beta)} | H_0 \text{ is true } \right\}}{\left(1-b^{-\gamma}\right)b} \\ ASL^3\left(r,\gamma,\log LR\right) &= \frac{\#of\left\{\log LR > 0 | H_0 \text{ is true } \right\}}{\left(1-b^{-\gamma}\right)b}, \end{aligned} \tag{4.5}$$

and the Empirical Power (EP), i.e. one minus the type II error, calculates as

$$\begin{aligned} EP^1\left(r,\gamma,LR\right) &= \frac{\#of\left\{LR > \hat{\ell}_\beta | H_1 \text{ is true } \right\}}{b^{-\gamma}b} \\ EP^2\left(r,\gamma,LR\right) &= \frac{\#of\left\{LR > \hat{\ell}_{(\alpha+\beta)} | H_1 \text{ is true } \right\}}{b^{-\gamma}b} \\ EP^3\left(r,\gamma,\log LR\right) &= \frac{\#of\left\{\log LR > 0 | H_1 \text{ is true } \right\}}{b^{-\gamma}b}. \end{aligned} \tag{4.6}$$

Now we want investigate when we can identify the informative features, so we conduct a Monte Carlo simulation under the alternative hypothesis in (3.4) for **Scenario I** with $\left(1-b^{-\gamma}\right)b\ s_0 \sim \chi^2(p_0,0)$ under H_0 (4.3) and $b^{-\gamma}b\ s_1 \sim \chi^2(p_0, 2r\log(b))$ under H_1 (4.3) for $b = 100000$, $\gamma = (0.5:1)$ with step 0.01, $r = (0:1)$ with step 0.02 and $p_0 = [5, 10, 20]$.

The separation scores are then run through the procedure described as Algorithm 1 to acquire the type I and II errors. This is repeated 10 times and a 95% confidence interval for the sum of errors is calculated. The upper bound of the confidence interval is graphically displayed as a contour plot for each degree of freedom in the $r-\gamma$ plane for all types of cutoffs in Figure 7. The region of possible identification corresponds to the green area, where the sum of errors is small. For the first type of cutoff (only top β-quantile) the estimable area is very small as expected. We can see that the groups are only well-separated for small block sizes ($p_0 \leq 10$) and large r and small γ. Widening the rejection area and accounting for α as well really improved the results and the pattern for the estimable area is now more coherent to the Gaussian case. Since we are in the domain of high dimensional data and in the sparse regime one rejected true null hypothesis has much less impact than a not rejecting a false null hypothesis, increasing the proportion of rejected hypotheses lowers the total error. For the third cutoff, where we strictly favor the alternative hypothesis and allow the widest rejection area, the estimable region is larger than for the Gaussian case. In this case we must be aware that more noise will be in the selected subset.

Next, we conduct a Monte Carlo simulation for **Scenario II**, with $\left(1-p^{-\gamma}\right)p\ g_0 \sim N(0,1)$ under H_0 (4.3) and $p^{-\gamma}p\ g_1 \sim N(\sqrt{(}2r\log(p)),1)$ under H_1 (4.3) for $p = 900000$,

Algorithm 1 Types I and II errors calculated with Likelihood Ratio Test

Input: s^2, a vector with observed separation strength for b blocks; $\mathbf{I} \in (0,1)$, an indicator vector with 1 for a truly informative block and 0 for a non-informative block; p_0, the block size, ω^2, the true non-centrality parameter for informative blocks and the cutoff values α and β.

Output: Types I and II errors for the three cutoffs.

for $i = 1$ to b **do**

  $\lambda_i = \frac{\chi^2_{p_0}(s_i^2;\omega^2)}{\chi^2_{p_0}(s_i^2;0)}$

end for

$\lambda_{(b)} \geq \ldots \geq \lambda_{(1)}$ order the Likelihood ratio in decreasing order

$\hat{\ell}_\beta = \lambda_{(i)}[\beta b]$ select a cutoff value that corresponds β

$\hat{\ell}_{\alpha+\beta} = \lambda_{(i)}[(\alpha+\beta)b]$ select a cutoff value that corresponds to α and β

for $i = 1$ to b **do**

  $R_i^\beta = 1$ IF $\lambda_{(i)} \geq \hat{\ell}_\beta$ ELSE $R_i^\beta = 0$

  $R_i^{\alpha\beta} = 1$ IF $\lambda_{(i)} \geq \hat{\ell}_{\alpha+\beta}$ ELSE $R_i^{\alpha\beta} = 0$

  $R_i^{\log} = 1$ IF $\log\lambda_{(i)} > 0$ ELSE $R_i^{\log} = 0$

end for

for $i = 1$ to b **do**

  $\epsilon_i^{I_\beta} = 1$ IF $R_i^\beta = 1 | \mathbf{I}_i = 0$ and $\epsilon_i^{II_\beta} = 1$ IF $R_i^\beta = 0 | \mathbf{I}_i = 1$

  $\epsilon_i^{I_{\alpha\beta}} = 1$ IF $R_i^{\alpha\beta} = 1 | \mathbf{I}_i = 0$ and $\epsilon_i^{II_{\alpha\beta}} = 1$ IF $R_i^{\alpha\beta} = 0 | \mathbf{I}_i = 1$

  $\epsilon_i^{I_{\log}} = 1$ IF $R_i^{\log} = 1 | \mathbf{I}_i = 0$ and $\epsilon_i^{II_{\log}} = 1$ IF $R_i^{\log} = 0 | \mathbf{I}_i = 1$

end for

type $\mathrm{I}_\beta = \frac{\sum_{i=1}^{b} \epsilon_i^{I_\beta}}{b-\beta b}$ and type $\mathrm{II}_\beta = \frac{\sum_{i=1}^{b} \epsilon_i^{II_\beta}}{\beta b}$

type $\mathrm{I}_{\alpha\beta} = \frac{\sum_{i=1}^{b} \epsilon_i^{I_{\alpha\beta}}}{b-\beta b}$ and type $\mathrm{II}_{\alpha\beta} = \frac{\sum_{i=1}^{b} \epsilon_i^{II_{\alpha\beta}}}{\beta b}$

type $\mathrm{I}_{\log} = \frac{\sum_{i=1}^{b} \epsilon_i^{I_{\log}}}{b-\beta b}$ and type $\mathrm{II}_{\log} = \frac{\sum_{i=1}^{b} \epsilon_i^{II_{\log}}}{\beta b}$

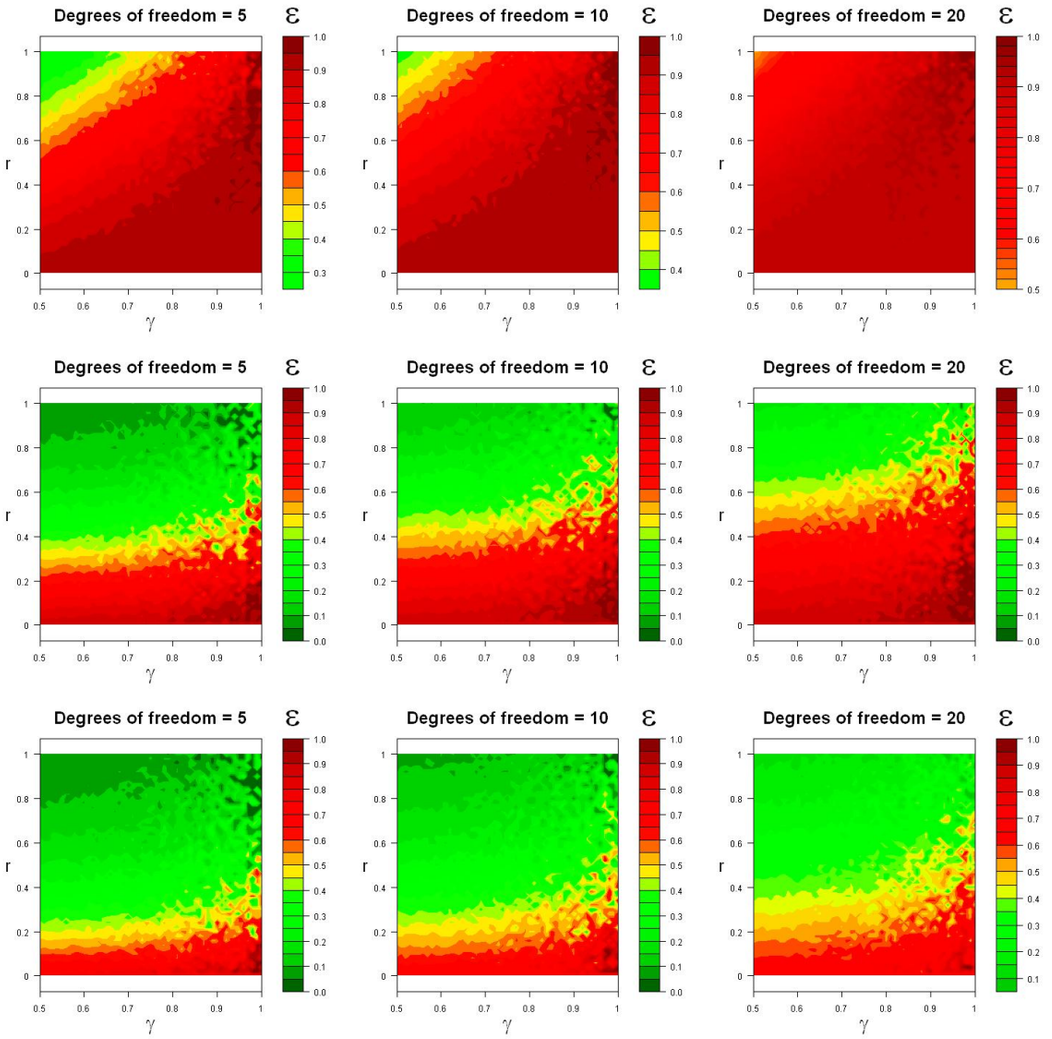

Figure 7: Empirical detection boundary in sparse setting for the χ^2-distribution using the empirical distribution of the likelihood ratio. Top row: the top β-quantile is used as the reference value for rejecting H_0. Mid row: the top $\alpha + \beta$-quantile is used as the reference value. Bottom row: the empirical distribution of log likelihood ratio, strictly favoring the alternative hypothesis.

$\gamma = [0.5 : 1]$ with step 0.01 and $r = [0 : 1]$ with step 0.02. Then we calculate the separation strength for the block sizes $p_0 = [5, 10, 20]$ and run through the procedure described as Algorithm 1 to acquire the type I and II errors. This is repeated 10 times and the confidence interval for the sum of errors is calculated. The upper bound for the confidence interval is displayed in Figure 8. In this setting the the pattern is reversed and, as expected, shows lower missed identification for larger blocks. Though in this scenario we observe an abrupt change in identification, the informative blocks become too rare for detection for lower values of γ since the number of informative features are too few. Given the liberal approach, sets of informative features are easily identified as long as $\gamma < 0.8$.

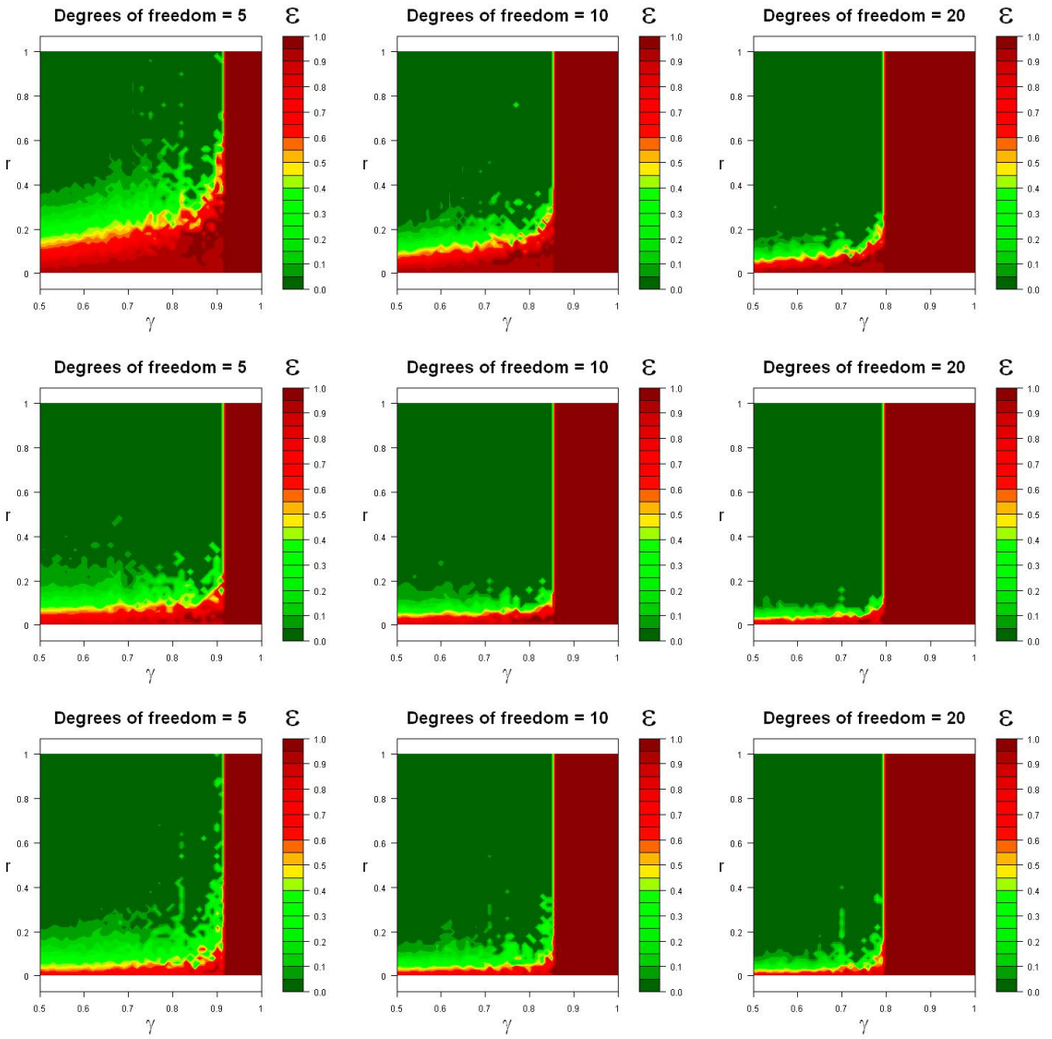

Figure 8: Empirical detection boundary in a sparse setting for the χ^2-distribution using the empirical distribution of the likelihood ratio for the second simulation approach. Top row: the top β-quantile is used as the reference value for rejecting H_0. Mid row: the top $\alpha + \beta$-quantile is used as the reference value. Bottom row: the empirical distribution of log likelihood ratio when strictly favoring the alternative hypotheses.

5 Estimating the proportion informative blocks

For classification, a relevant question is if there are any informative blocks within a data set, and by estimating β we get a quick answer. Though $\beta \neq 0$ indicates the presence of separation strength, it does not give the answer to which blocks are actually informative. This intermediate step is important for feature selection in classification. We explore in this section how estimation of β can affect further thresholding.

We start with the simple Bonferroni correction as a baseline and compare it with two other methods. The analyses are based on the distribution of the p-values of the

hypothesis tests, that if a null hypothesis is true $\pi_i \sim U(0,1)$, where U denotes the uniform distribution. In this context our β denotes the proportion of false null hypotheses and can be calculated as

$$\beta = \frac{\sum_{i=1}^{b} \mathbf{1}\{\omega_i \neq 0\}}{b}. \tag{5.1}$$

A lower bound for $\hat{\beta}$ with the property $P\left(\hat{\beta} \leq \beta\right) \geq 1-\alpha$ can be constructed for a specified confidence level $1-\alpha$. Then the proportion of false null hypotheses is at least $\hat{\beta}$ and the global null hypothesis, there are no false null hypotheses ($\beta = 0$), can be tested at level α by rejecting when $\beta > 0$.

A standard estimate for proportion of null values $\varpi_0 = 1-\beta$ based on p-values starts with ranking the p-values in increasing order $\pi_{(1)} \leq \pi_{(2)} \leq \ldots \leq \pi_{(b)}$; then

$$\hat{\varpi}_0(t) = \frac{\sum_{1=i}^{b} \mathbf{1}\left\{\pi_{(i)} > t\right\}}{(1-t)\,b}. \tag{5.2}$$

The quantity $\hat{\beta} = 1 - \hat{\varpi}_0$ is then an estimate for a lower bound on the proportion of false null hypotheses [23]. A disadvantage for this estimate is that it is based on the choicee of t; see e.g. [20].

The third method for estimating the proportion of false null hypotheses for a given $t \in (0,1)$ is suggested by [17]

$$\hat{\beta} = \max_{t \in (0,1)} \frac{F_b(t) - t - \mathrm{B}_{b,\alpha}\Delta(t)}{1-t} \tag{5.3}$$

where $F_b(t) = \frac{\sum_{i=1}^{b} \mathbf{1}\{\pi_i \leq t\}}{b}$ is the empirical distribution of p-values, $\Delta(t) = \sqrt{t(1-t)}$ the standard deviation-proportional bounding function and $\mathrm{B}_{b,\alpha}$ the bounding sequence for $\Delta(t)$ at level α given by

$$\mathrm{B}_{b,\alpha} = \frac{G^{-1}(1-\alpha) + m_b}{n_b} \tag{5.4}$$

where G is the Gumbel distribution, $m_b = 2\log_2 b + \frac{1}{2}\log_3 b - \frac{1}{2}\log 4\pi$[1] and $n_b = \sqrt{2b\log_2 b}$.

To compare the three methods numerically we employ the two earlier described scenarios. First we consider **Scenario I** and generate a Monte Carlo simulation with $(1-b^{-\gamma})\,b$ $s_0 \sim \chi^2(p_0, 0)$ under H_0 and $b^{-\gamma}b$ $s_1 \sim \chi^2(p_0, 2r\log(b))$ under H_1 for $b = 5000$, $\gamma = (0.5:1)$ with step 0.01, $r = (0:1)$ with step 0.02 and $p_0 = [5, 10, 20, 30]$. Then the proportion false null hypotheses is estimated, $\hat{\beta}$, with respective method for significance level and threshold value set as $\alpha = t = 0.05$. For the method suggested by [17] we use the `R` package HOWMANY and the total number of estimated false null hypotheses for all b blocks. The ratio between the estimate and the true proportion is calculated $\left(\frac{\hat{\beta}}{\beta}\right)$. To standardize the maximum ratio to one when the estimated proportion is larger than the true proportion, the ratio calculation is reversed and calculated as $\frac{\beta}{\hat{\beta}}$. This is then repeated 10 times and the average calculated and presented as a contour plot within the $r-\gamma$ plane, where green

[1] $\log_2 = \log\log$ and $\log_3 = \log\log\log$

indicates high concordance and red discrepancy; see Figure 9. As can easily be seen, all three methods have severe difficulties with getting accurate estimates of β in SWB. Even in the detectable area given by the LRT the methods fail at estimating the true proportion of false null hypotheses. It is only the method suggested by [23] that comes somewhat close to good estimates of β in the very top of the detectable area.

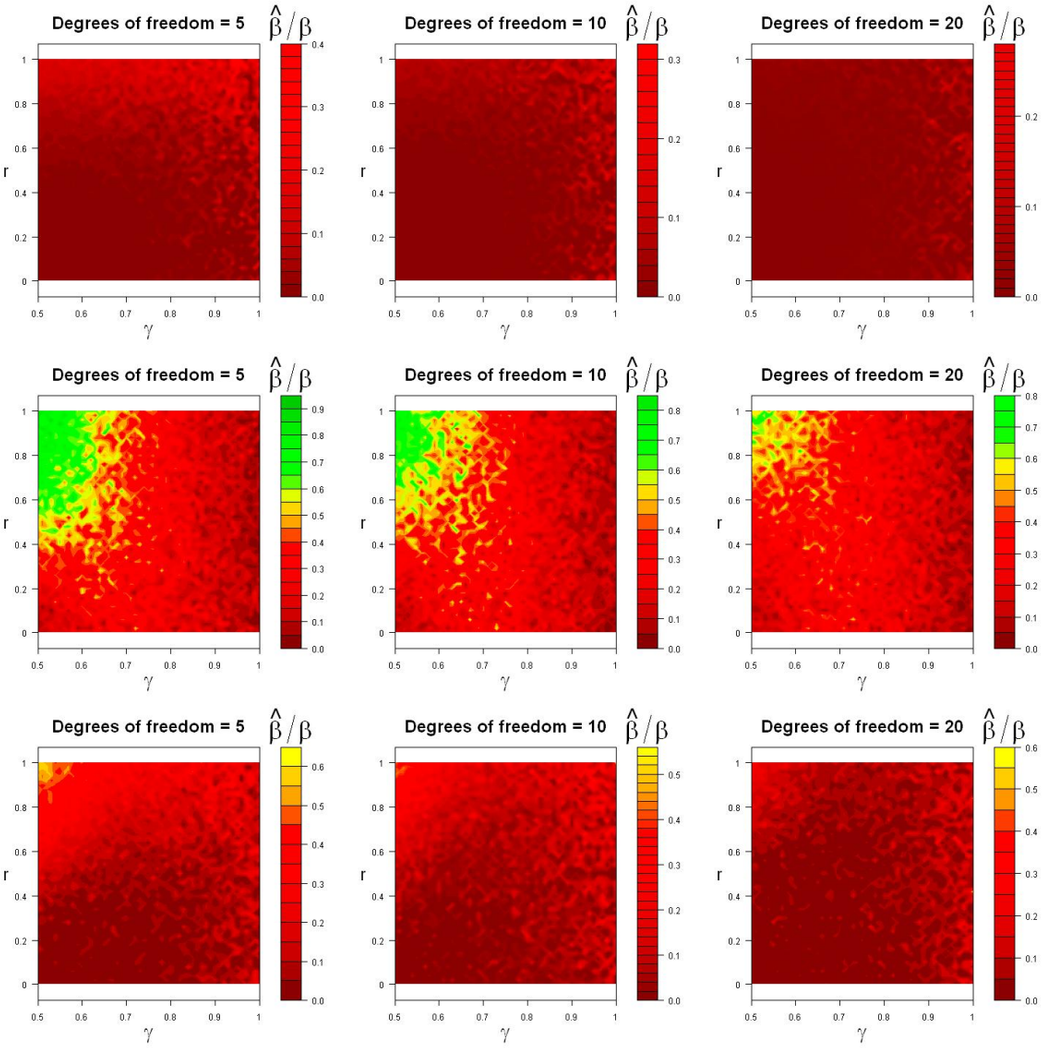

Figure 9: Contour plot of the ratio between the estimated proportion of false null hypotheses and the true proportion (β) in the $r - \gamma$ plane, a ratio of one indicating full concordance and the lower the value, the higher the discrepancy between the estimated proportion and the true value. Top row: $\hat{\beta}$ from the Bonferroni correction. Mid row: $\hat{\beta}$ from method by [23]. Bottom row: $\hat{\beta}$ from method by [17].

To investigate how much separation strength is required to get reliable estimates of the proportion false null hypotheses we conduct a second Monte Carlo simulation, still as a mixture of $(1 - b^{-\gamma})\, b$ $s_0 \sim \chi^2(p_0, 0)$ and $b^{-\gamma} b$ $s_1 \sim \chi^2(p_0, 2r \log(b))$ for $b = 5000$ but now for a moderately sparse setting with γ fixed to 0.55 and increasing the separation

strength, $r = [0.1 : 2]$ with step 0.1. We estimate the ratio $\frac{\hat{\beta}}{\beta}$ and repeat 100 times. As can be seen in Figure 10, the Bonferroni correction is too restrictive and consistently underestimates the proportion of false null hypotheses, the method suggested by Storey & Tibshirani overestimates the proportion when the ratio > 1 and underestimates when the ratio < 1, leading to the median of the ratios being close to one, whereas the Meinshausen & Rice method is more in line with the restrictiveness of the Bonferroni correction but comes closer to the true proportion, especially for more degrees of freedom. For all of the methods the information strength needs to be substantial, e.g. $r > 1$, to come close to the true proportion of false null hypotheses, and the higher the degrees of freedom, the less well all of the methods perform.

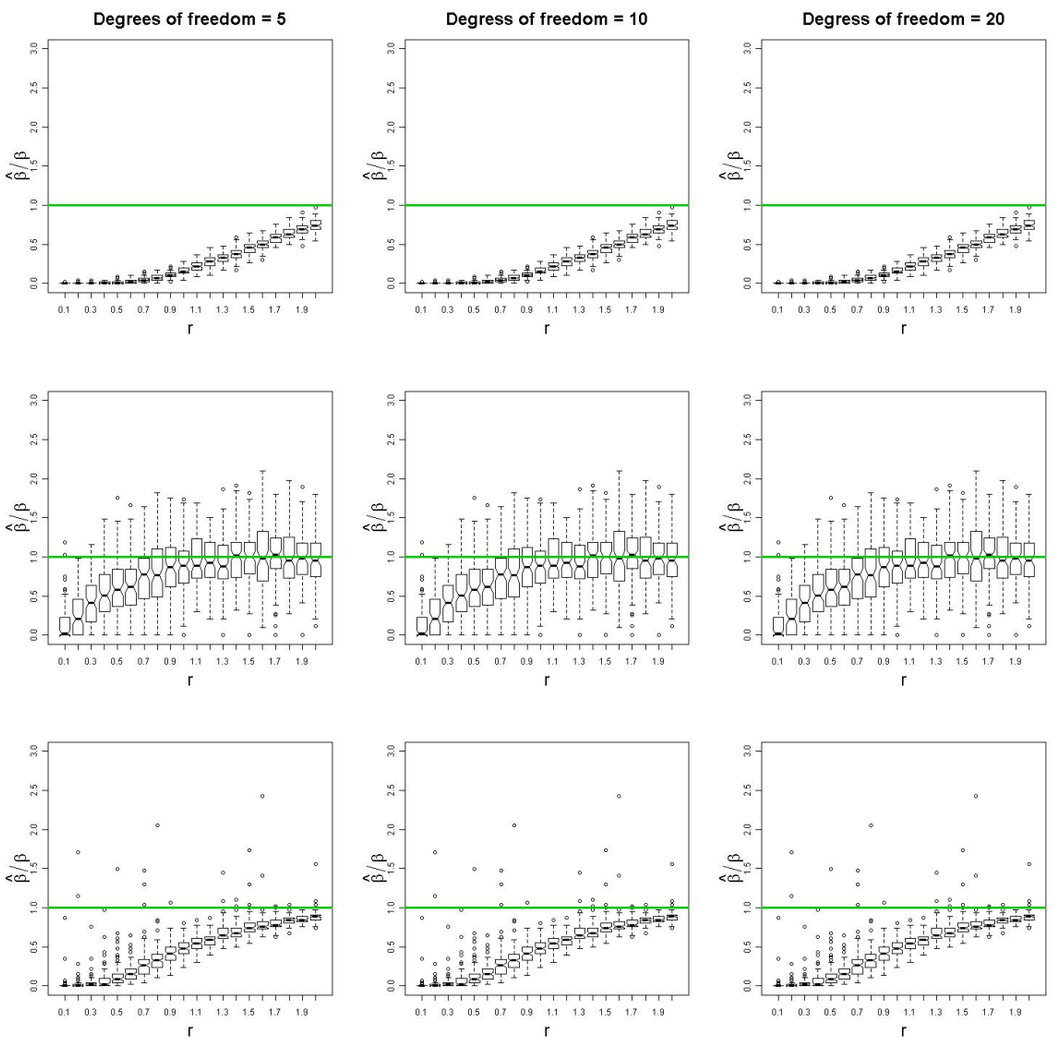

Figure 10: The ratio between estimated proportion of false null hypotheses and the true proportion as a function of information strength r for a given sparsity of ($\gamma = 0.55$), for different block sizes. Results are shown for estimate from the Bonferroni correction (top row), the Storey & Tibshirani method (middle row) and the Meinshausen & Rice method (bottom row).

We instead consider **Scenario II**, where larger blocks indicate higher separation strength, and repeat the Monte Carlo simulation starting with Gaussian distribution and then constructing blocks. All parameters are the same as in the first setup of the second scenario except for number of features, which is changed to $p = 60000$. As for the first scenario, the proportion of false null hypotheses is estimated with the above described methods and the ratio $\left(\frac{\hat{\beta}}{\beta}\right)$ calculated. Also here this is repeated 10 times and the average calculated and presented as contour plot within the $r - \gamma$ plane where green indicates high concordance and red discrepancy; see Figure 11. Now the area of high concordance corresponds to the detectable area for $\hat{\beta}$ estimated with both the Bonferroni correction and the Meinshausen & Rice method, whereas the estimate with Storey & Tibshirani method fails.

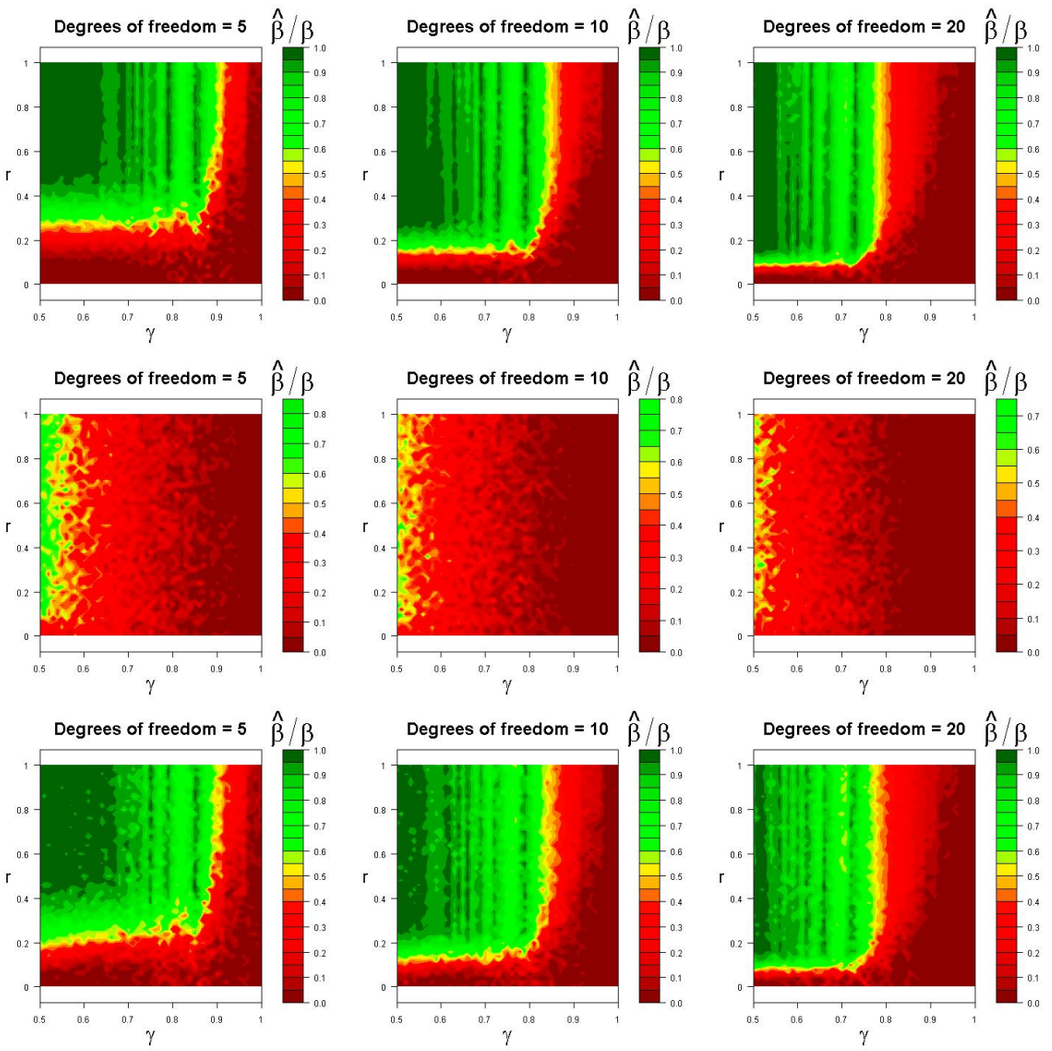

Figure 11: Contour plot of the ratio between the estimated proportion of false null hypotheses and the true proportion (β) in the $r - \gamma$ plane, a ratio of one indicating full concordance and the lower value, the higher the discrepancy between the estimated proportion and the true value. Top row: $\hat{\beta}$ from Bonferroni correction. Mid row: $\hat{\beta}$ from method by [23]. Bottom row: $\hat{\beta}$ from method by [17].

We replicate the second Monte Carlo simulation from Scenario I now for **Scenario II** with the same set of parameters except for keeping γ fixed at 0.55 and letting separation strength increase, $r = (0.1 : 2)$ with step 0.1. Using the same calculation as before and repeated 100 times, the result is shown in Figure 12. Also here can we see a tendency to correct estimate of the proportion of false null hypotheses for very low separation strength in this setting using the Bonferroni correction or the Meinshausen & Rice method. Now the tendency is overestimation of the proportion rather than underestimation, and the basic the Bonferroni correction is the method that gives the best results. When $\hat{\beta}$ is calculated with Storey & Tibshirani′s method, even if the median corresponds to a ratio of one, the estimates are too widely spread and do not give consistent results.

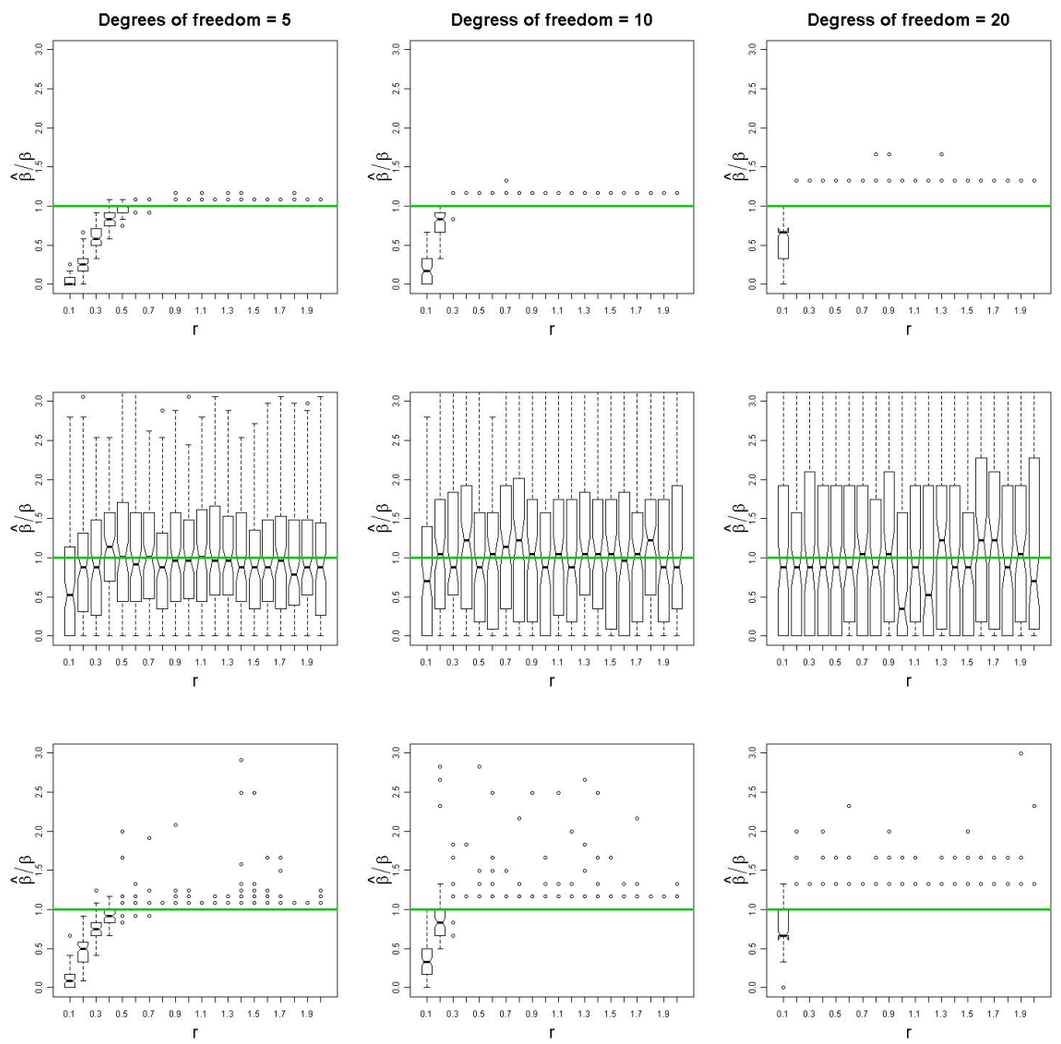

Figure 12: The ratio between estimated proportion of false null hypotheses and the true proportion as a function of information strength r for a given sparsity of ($\gamma = 0.55$) for different block sizes. Results are shown for estimates from the Bonferroni correction (top row), the Storey & Tibshirani method (middle row) and the Meinshausen & Rice method (bottom row).

Meinshausen & Rice (2006) showed, that for their method, the ratio between $\hat{\beta}$ and the true proportion tended to one for a growing number of features in all areas of the SW model where LRT succeeds. This could not be seen in the restrictive approach for the SWB model. We can conclude that in the liberal approach it is possible to get good estimates of the proportion of false null hypotheses, but for the sparse and weak block model we cannot take for granted strong $\mathcal{S}_i^2$. This strongly motivates us to turn to bHC since the main advantage of HC is the optimal adaptivity to unknown parameters; see e.g.[27].

6 Block Selection by Thresholding

Here we present some alternative commonly employed selection strategies, to which we will compare the performance of our bHC.

6.1 False discovery rate

False Discovery Rate (Fdr) was introduced as a useful approach to simultaneous testing by [2]. It is a tail-areas based procedure that aims to control the number of false discovery rates, i.e. the expected ratio of the number of false informatives among all rejected features. The Fdr is distribution-based and on the p-value scale it is defined as

$$\mathrm{Fdr}(\pi_i) = P\left(\text{Non informative}|\Pi \leq \pi_i\right) = \frac{(1-\beta)\pi_i}{F(\pi_i)}, \tag{6.1}$$

where $F(\pi_i)$ is the mixed distribution of the p-values and $1-\beta$ is the proportion of non-informative features. When estimating the tail-area based Fdr, the p-values are ranked in increasing order($\pi_{(1)} \leq \pi_{(2)} \leq \ldots \leq \pi_{(p)}$) and the empirical estimate of Fdr is the rule of Benjamini and Hochberg (1995):

$$\widehat{\mathrm{Fdr}}(\pi_{(i)}) = \frac{(1-\hat{\beta})\pi_{(i)}}{\hat{F}(\pi_{(i)})} \leq \frac{p}{i}\pi_{(i)}. \tag{6.2}$$

There is also another false discovery rate criterion that is based on densities, the Local False Discovery Rate (Lfdr), suggested by [8, 7]. This is an empirical Bayes version of Benjamin and Hochberg′s (1995) rule and focus on densities rather than tail areas. It starts with a simple Bayes model: assume that the p test score-values fall into two classes, *Informative* or *Non-informative*, depending on whether or not the test score is generated according to the null hypothesis. The two classes have the prior probabilities $(1-\beta)$ and β and the test score (we will use p-values) has density either $f_{H_0}(\pi_i)$ or $f_{H_1}(\pi_i)$ depending on its class:

$$\begin{aligned} (1-\beta) &= P\left\{\text{Non-informative}\right\},\ f_{H_0}(\pi_i) \text{ density if Non-informative (Null)} \\ \beta &= P\left\{\text{Informative}\right\},\ f_{H_1}(\pi_i) \text{ density if Informative (Non-null).} \end{aligned} \tag{6.3}$$

These are then combined to give the mixed density

$$f(\pi_i) = (1-\beta)f_{H_0}(\pi_i) + \beta f_{H_1}(\pi_i). \tag{6.4}$$

Then the Lfdr is defined according to Bayes' theorem as the *a posteriori* probability of being in the *Non-informative* class given π_i:

$$\text{Lfdr}(\pi_i) = P\left\{\text{Non-informative}|\pi_i\right\} = \frac{(1-\beta)f_{H_0}(\pi_i)}{f(\pi_i)}. \tag{6.5}$$

The two Fdr methods are closely related but the key difference is that the Lfdr, being density based, implicitly assumes that the number of hypotheses is large ($p \to \infty$). There is a direct relationship between Fdr and Lfdr:

$$\text{Fdr}(\Pi) = E\left\{\text{Lfdr}(\pi)|\pi \in \Pi\right\}. \tag{6.6}$$

Both the local and the tail-area based Fdrs are designed to work with p-values as input test statistics, similar to HC. We will identify the thresholds for both false discovery rate methods based on the p-values using the R package `fdrtool` [24]. This package uses a modified Grenander approach for density estimation; for details see [25].

In simulations a fourth method is used, Oracle local false discovery rate (Ofdr), where we use test score instead of the p-values and assume that the parameters for the density are known. This method is also based on a multiple-testing procedure, though it will not rely on p-values, as individual p-values are appropriate for testing single hypothesis but they fail as building blocks in multiple testing according to [26]. Instead it is based on the LR and defined as $\text{Ofdr}(x) = \frac{(1-\beta)f_{H_0}(x)}{f_{H_1}(x)}$. Since β is a global parameter the Ofdr implies that the relative importance of the observations can be ranked according to the LRs. The procedure to identify informative blocks consists of three steps: first, equipping the general adaptive testing with consistent estimates (β, ω^2) (in our oracle setting they are known) and calculating the Ofdr for observed block separation strength as

$$\text{Ofdr}(s_i^2) = \frac{(1-\beta)\,\chi^2_{p_0}\left(s_i^2; 0\right)}{(1-\beta)\,\chi^2_{p_0}\left(s_i^2; 0\right) + \beta\chi^2_{p_0}\left(s_i^2; \omega^2\right)}. \tag{6.7}$$

The second step is then to rank the Ofdr in an increasing order: $\text{Ofdr}(s_i^2)_{(1)} \leq \text{Ofdr}(s_i^2)_{(2)} \leq \ldots \leq \text{Ofdr}(s_i^2)_{(b)}$. The third step is to reject all $H_0^{(i)}$, $i = 1, \ldots, k$ where

$$k = \max_{1 \leq i \leq k}\left\{i; \frac{1}{i}\sum_{j=1}^{i}\text{Ofdr}_{(j)} \leq \alpha\right\}. \tag{6.8}$$

Synthetic multivariate block data

The main goal with selecting a subset of informative blocks/features is to then use them for classification between objects. Here we will look at multivariate Gaussian data with two classes and estimate the separation score. We generate data blockwise for a given block dependence structure (Σ_{p_0}) and sample from different distributions depending on class and number of informative features. First, we fix the parameters $(p, p_0, \beta, \varepsilon, \Sigma_{p_0}, n_1, n_2)$, and then the number of blocks $b = \frac{p}{p_0}$, number of informative features $\tilde{p} = \beta p$ /blocks $\tilde{b} = \frac{\beta p}{p_0}$ and the sparsity measure $\gamma = -\frac{\log\beta}{p/p_0}$. Since $\sum_{i=1}^{\tilde{b}}\delta_i^2 = -2\Phi^{-1}(\varepsilon)$, given that we set $\mu_1 = 0$ for the first class, we get $\mu_2 = \sqrt{\frac{\delta_i^2}{\mathbf{1}'_{p_0}\Sigma_{p_0}^{-1}\mathbf{1}_{p_0}}}$ for the second class. Now, for

the null hypothesis class, we draw p/p_0 samples from $N(\mu_1, \Sigma_{p_0})$; for the alternative hypothesis, we first draw $p(1-\beta)/p_0$ samples from $N(\mu_1, \Sigma_{p_0})$ and then draw $\tilde{p}/p_0$ samples from $N(\mu_2, \Sigma_{p_0})$. Even though the separation score is now calculated from data that originally comes from Gaussian distribution (as in Scenario II), this is more in line with the restricted setup in Scenario I since the feature contrast decreases with the number of degrees of freedom.

We start with comparing the performance of the thresholding methods for single features versus blocks. The parameters are set as $\beta = 0.01$, $\epsilon = 0.03$, $p = 10000$, and $n1 = n2 = 50$. For the covariance the true block size is 10 and variance for each feature is 1 and between features 0.6. For the block size (p_0) we start with 1, i.e. a single feature, and then continue with 5, 10 and 15 [2]. For the single feature we calculate the Students t-test as an approximation of Z and for the blocks s_i^2 and the p-values were estimated from the central t-distribution with $n-2$ degrees of freedom and the central χ^2-distribution with degrees of freedom corresponding to the block size; see the first two rows of Figure 13. Then the four thresholding methods are implemented and the number of false null hypotheses estimated and shown as vertical lines in Figure 13. As can be seen, all four methods give good estimates in all of the settings. The bHC and the different fdr methods work as well for the χ^2-distributed score as for single features.

[2] For these settings the parameters correspond to $\gamma = (0.5, 0.6, 0.67, 0.71)$ and $r = (0.22, 0.29, 0.32, 0.33)$.

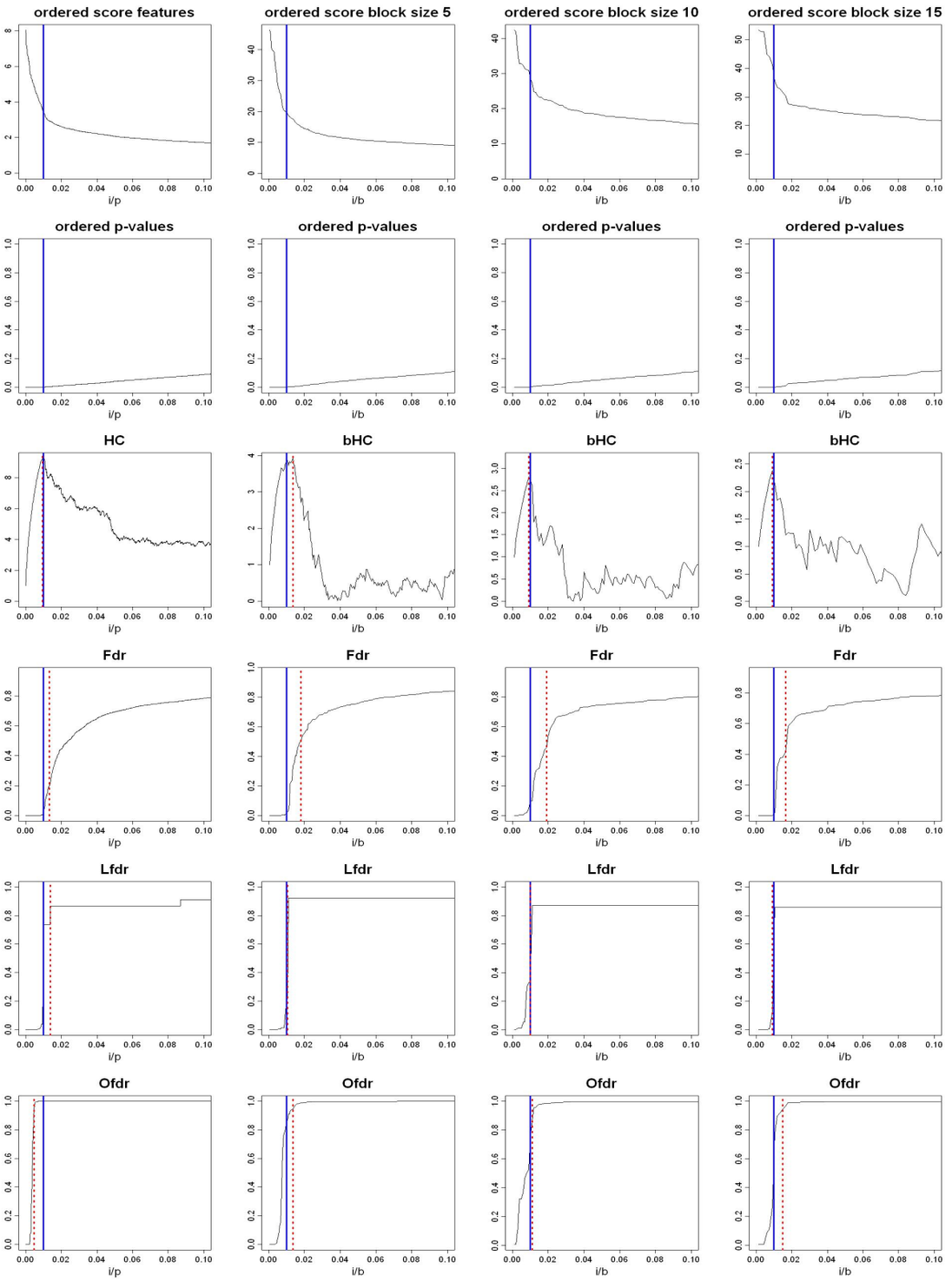

Figure 13: Results for the test score, ordered p-values and the different thresholding methods for features versus growing number of blocks. The solid blue line represents the true proportion of false null hypotheses (β) and the dotted red line the estimated ($\hat{\beta}$).

7 Block selection and the effect on misclassification

Now we know it is possible to select informative features with different thresholding methods and we can turn to the main goal, classification between subjects. Here we evaluate how successful in detecting useful features the methods are by using the selected subset in classification. We compare misclassification error regarding classes between methods as well as the misclassification that is obtained if the full data set is used. We assign the subjects in the test data to either class using the additive classifier and then the misclassification error is calculated as

$$mc = \frac{\#\{\hat{y}_j = 1 | y_j = 2\} + \#\{\hat{y}_j = 2 | y_j = 1\}}{n}, \tag{7.1}$$

where $\hat{y}_j$ is the predicted class and y_j is the true class for the jth observation. Here we use 10-fold cross-validation to divide the data into training and test data.

7.1 Synthetic data

To generate class data we use the synthetic multivariate block data described above and employ Monte Carlo simulation with the parameters set as follows: $\beta = [0.01, 0.02, 0.03]$, $\varepsilon = [0.05, 0.1, 0.15]$, $n_1 = n_2 = 50$, $\sigma^2 = 1$ and $\rho = 0.8$ for $p = 2500$ with $p_0 = 5$, $p = 5000$ with $p_0 = 10$ and $p = 10000$ with $p_0 = 20$. For each of the nine data sets we estimate thresholds to select a subset of informative blocks for the classification using the four different thresholding methods. In order to make the classification run smoothly, a condition is added to set the number of informative blocks to one if no suitable threshold been identified.

As an illustrative example we start by estimating the misclassification for accumulation of blocks, i.e. estimating mc using only the block with the highest separation score in the classifier, then using the blocks with two the highest separation scores, and so on until the mc estimate is based on using all blocks in the additive classifier. For each set of data we selected a subset of blocks based on the threshold choice from the four thresholding methods; the vertical lines in Figures 14 to 16 correspond to the number of blocks that are included in the subset for each of the methods.

For the small block size ($p_0 = 5$) and when the classes are well-separated, the numbers of selected blocks are quite similar. Then as the classification difficulty increases, the bHC thresholding tends to include more blocks in the subset than actually correspond to the true number of informative blocks, but still gives lower misclassification.

When the block size is moderate ($p_0 = 10$) this is even more noticeable as for all nine settings bHC results in the highest number of chosen blocks but also the lowest misclassification, whereas the number of blocks selected by the other methods are more similar and in line with β.

For the final block size ($p_0 = 20$) the subset chosen by Fdr and Lfdr includes few blocks for all settings; for $\varepsilon \geq 0.1$ it is likely that no threshold was identified. In the case of well-separated classes bHC selects the largest number of blocks and receives the lowest misclassification, whereas ε increase the Ofdr, in some cases, has better success.

Next we repeat the Monte Carlo simulation for the given parameters 100 times with the true misclassification rate extended to $\varepsilon = [0.005, 0.01, 0.05, 0.1, 0.15]$ for $p = 5000$ with $p_0 = 5$, $p = 10000$ with $p_0 = 10$ and $p = 20000$ with $p_0 = 20$. We compare the four

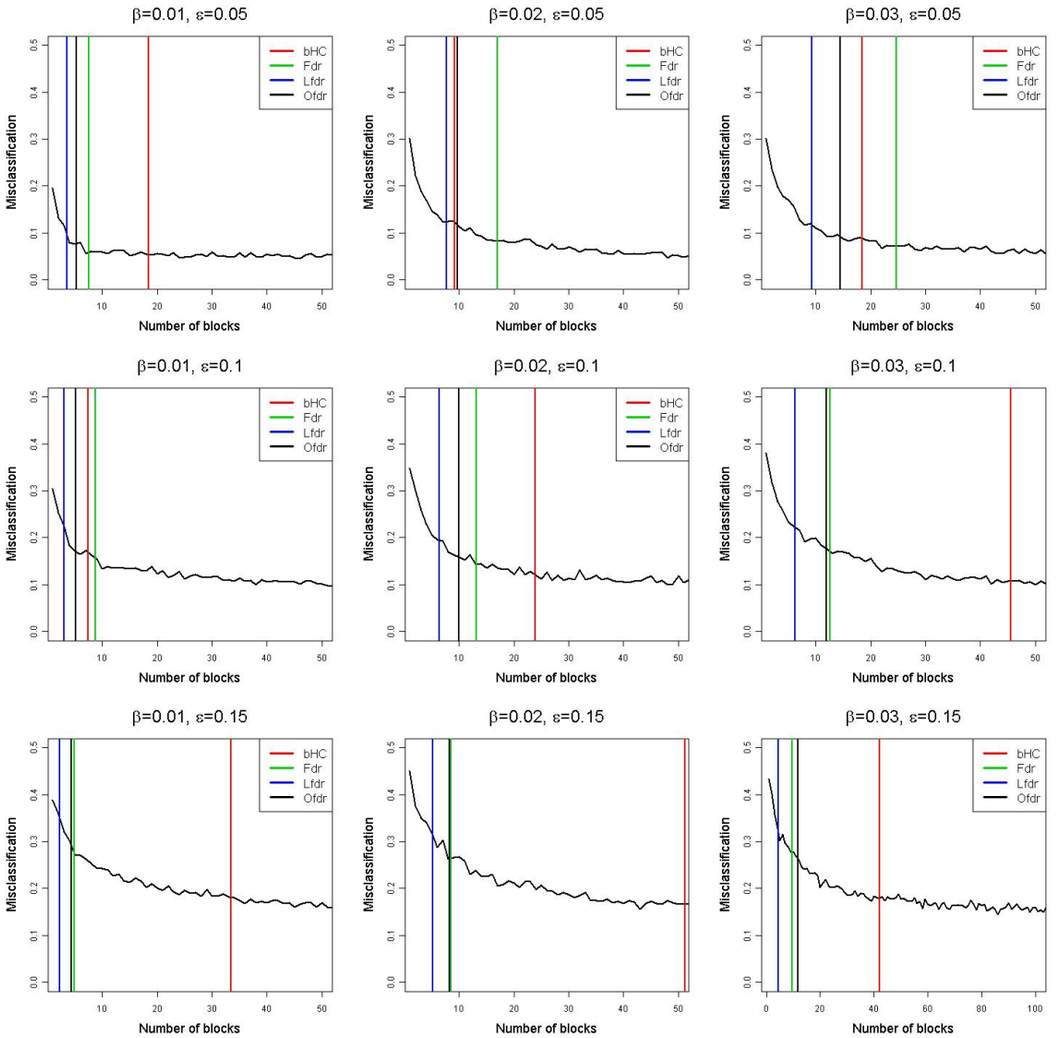

Figure 14: The misclassification rate for accumulation of blocks and the cutoffs for the four thresholding methods (black line Ofdr, red line HC, blue line Fdr and green line Lfdr) for a block size of 5 to 500 blocks.

methods of misclassification as well as the estimate of number of informative blocks and the proportion of falsely informative blocks defined as

$$fpr = \frac{\#\{H_0 \text{ rejected} | H_0 \text{ true}\}}{(1-\beta)b} \tag{7.2}$$

and the results are shown in Tables 1 to 3.

For the small block size ($p_0 = 5$) when the classes are well-separated ($\varepsilon < 0.05$) all methods come close to the true number of informative blocks and have low misclassification. bHC and the Ofdr have the best performance. Even though bHC somewhat overestimates the number of informative blocks for the more difficult classification setting resulting in the highest fpr, it still gives very low misclassification. Both Fdr and Lfdr

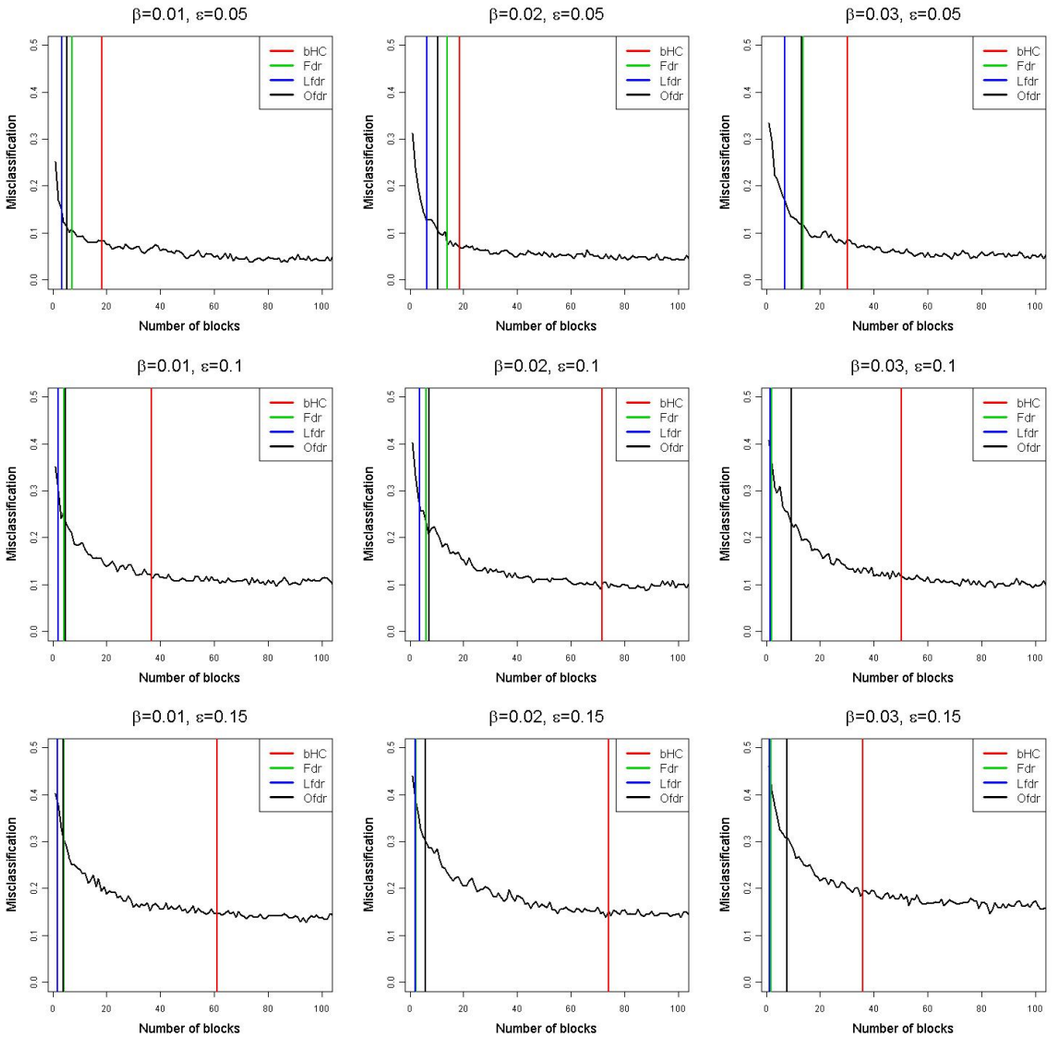

Figure 15: The misclassification rate for accumulation of blocks and the cutoffs for the four thresholding methods (black line Ofdr, red line HC, blue line Fdr and green line Lfdr) for a block size of 10 to 500 blocks.

tend more to underestimate the number of informative blocks, which results in very low fpr and also generates more misclassification.

When the block doubles in size, $p_0 = 10$, bHC still gives consistently low misclassification even though the tendency is still to overestimate the number of informative blocks. The Ofdr also still performs well even though it is not as good as bHC and more restrictive in the estimate of number of informative blocks. For well-separated classes the Fdr generates low misclassification and comes close to the correct number of informative blocks. When the classes come closer together the selected number of informative blocks becomes too small for good classification. The tendency for the Lfdr is the same as for Fdr but selects even fewer blocks.

For the last block size ($p_0 = 20$), bHC shows a reversed pattern by overestimating

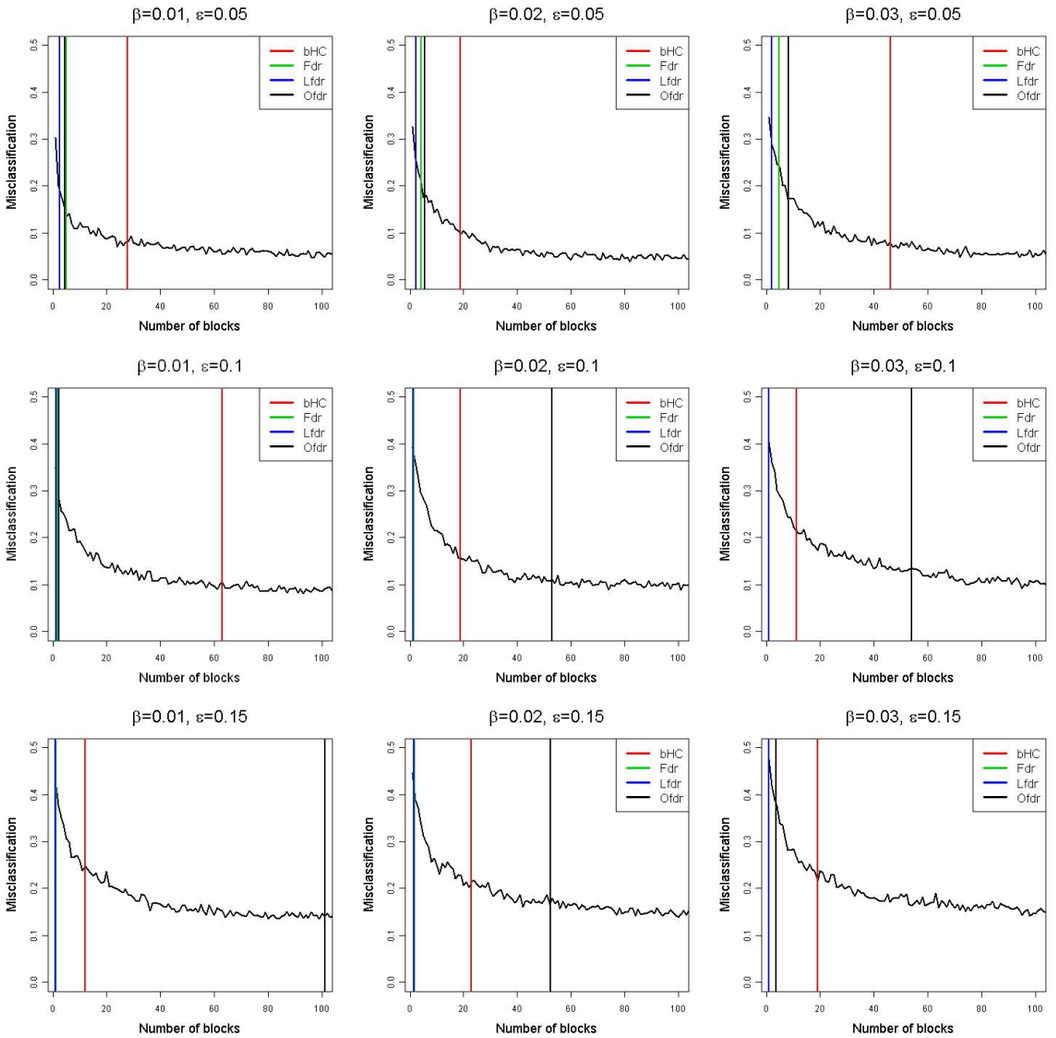

Figure 16: The misclassification rate for accumulation of blocks and the cutoffs for the four thresholding methods (black line Ofdr, red line HC, blue line Fdr and green line Lfdr) for a block size of 20 to 500 blocks.

the number of informative blocks for the well-separated classes and underestimating for when the classes are closer together, but the misclassification rate is still surprisingly low and accurate. Now the Ofdr performs notably less well in comparison to bHC; it comes close only for the well-separated classes. Both the Fdr and Lfdr struggle with identifying informative blocks, which results in low classification accuracy.

		bHC						Fdr						Lfdr						Ofdr					
		$\tilde{b}$		fpr		mc		$\tilde{b}$		fpr		mc		$\tilde{b}$		fpr		mc		$\tilde{b}$		fpr		mc	
β	ϵ	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd
0.010	0.005	11	2.6	0.00	0.00	0.01	0.01	17	5.7	0.01	0.00	0.02	0.03	8	3.0	0.00	0.00	0.03	0.05	10	0.6	0.00	0.00	0.01	0.01
0.020	0.005	22	19.7	0.00	0.02	0.01	0.01	36	9.5	0.02	0.01	0.01	0.03	16	4.3	0.00	0.00	0.02	0.03	21	1.8	0.00	0.00	0.01	0.01
0.030	0.005	32	21.0	0.01	0.02	0.01	0.01	53	10.6	0.03	0.01	0.01	0.01	24	5.5	0.00	0.00	0.02	0.01	30	3.2	0.01	0.00	0.01	0.01
0.010	0.010	12	13.1	0.00	0.01	0.01	0.01	17	6.2	0.01	0.01	0.02	0.04	8	2.6	0.00	0.00	0.03	0.04	10	0.6	0.00	0.00	0.01	0.01
0.020	0.010	20	5.4	0.00	0.00	0.02	0.02	34	9.0	0.02	0.01	0.01	0.01	15	4.9	0.00	0.00	0.04	0.04	21	2.1	0.00	0.00	0.02	0.01
0.030	0.010	37	33.1	0.02	0.03	0.02	0.02	47	13.6	0.02	0.01	0.01	0.01	21	6.5	0.00	0.00	0.03	0.02	29	3.3	0.01	0.00	0.02	0.01
0.010	0.050	22	48.3	0.01	0.05	0.06	0.04	15	7.5	0.01	0.01	0.07	0.07	7	4.0	0.00	0.00	0.12	0.08	10	2.0	0.00	0.00	0.07	0.03
0.020	0.050	57	79.7	0.05	0.08	0.06	0.05	19	9.5	0.01	0.01	0.08	0.05	9	5.3	0.00	0.00	0.14	0.07	16	3.3	0.01	0.00	0.08	0.03
0.030	0.050	96	90.8	0.08	0.09	0.07	0.05	19	10.9	0.01	0.01	0.09	0.07	9	6.1	0.00	0.00	0.15	0.08	24	3.9	0.01	0.00	0.06	0.03
0.010	0.100	65	89.9	0.06	0.09	0.12	0.07	9	6.1	0.00	0.00	0.15	0.09	4	3.1	0.00	0.00	0.21	0.08	8	2.3	0.00	0.00	0.14	0.05
0.020	0.100	118	104.8	0.11	0.10	0.10	0.06	6	6.2	0.00	0.00	0.23	0.10	3	4.0	0.00	0.00	0.27	0.08	13	3.6	0.01	0.00	0.12	0.05
0.030	0.100	140	108.8	0.13	0.11	0.11	0.09	4	4.6	0.00	0.00	0.26	0.10	2	2.8	0.00	0.00	0.30	0.08	20	4.0	0.01	0.00	0.11	0.03
0.010	0.150	94	97.2	0.09	0.10	0.18	0.10	4	4.4	0.00	0.00	0.25	0.10	2	1.9	0.00	0.00	0.30	0.07	6	2.5	0.00	0.00	0.20	0.06
0.020	0.150	103	111.6	0.10	0.11	0.16	0.12	2	2.7	0.00	0.00	0.31	0.08	1	1.4	0.00	0.00	0.33	0.06	11	3.5	0.01	0.00	0.16	0.05
0.030	0.150	108	109.6	0.10	0.11	0.17	0.12	2	1.8	0.00	0.00	0.32	0.07	1	1.0	0.00	0.00	0.34	0.05	17	4.1	0.01	0.00	0.15	0.04

Table 1: Number of blocks selected as informative $\left(\tilde{b}\right)$, proportion of falsely selected blocks (fpr) and the misclassification rate (mc) averaged over 100 runs, presented as mean (m) and standard deviation (sd) for block size $p_0 = 5$.

		bHC						Fdr						Lfdr						Ofdr					
		$\tilde{b}$		fpr		mc		$\tilde{b}$		fpr		mc		$\tilde{b}$		fpr		mc		$\tilde{b}$		fpr		mc	
β	ϵ	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd
0.01	0.005	13	28.0	0.01	0.03	0.01	0.01	14	6.9	0.01	0.00	0.04	0.06	7	3.5	0.0	0.0	0.06	0.07	10	0.7	0.00	0.0	0.01	0.01
0.02	0.005	23	39.6	0.01	0.04	0.02	0.02	29	9.4	0.01	0.01	0.02	0.04	14	5.2	0.0	0.0	0.04	0.05	20	2.6	0.00	0.0	0.02	0.01
0.03	0.005	33	36.9	0.01	0.04	0.02	0.02	36	11.1	0.01	0.01	0.02	0.03	17	6.4	0.0	0.0	0.04	0.03	28	3.8	0.01	0.0	0.02	0.02
0.01	0.010	14	29.2	0.01	0.03	0.02	0.02	12	7.5	0.00	0.00	0.07	0.09	6	3.7	0.0	0.0	0.09	0.09	10	0.8	0.00	0.0	0.02	0.02
0.02	0.010	21	19.9	0.01	0.02	0.03	0.02	23	8.2	0.01	0.01	0.03	0.05	11	5.1	0.0	0.0	0.07	0.06	18	2.7	0.00	0.0	0.03	0.02
0.03	0.010	65	79.9	0.05	0.08	0.03	0.03	28	10.9	0.01	0.01	0.04	0.03	13	8.0	0.0	0.0	0.08	0.05	25	3.6	0.01	0.0	0.03	0.02
0.01	0.050	57	90.1	0.05	0.09	0.07	0.05	9	5.9	0.00	0.00	0.12	0.08	4	3.2	0.0	0.0	0.17	0.08	8	1.8	0.00	0.0	0.10	0.03
0.02	0.050	93	104.0	0.08	0.10	0.06	0.08	6	5.6	0.00	0.00	0.20	0.11	3	2.9	0.0	0.0	0.26	0.09	13	3.1	0.01	0.0	0.10	0.04
0.03	0.050	143	114.7	0.13	0.11	0.05	0.08	3	3.4	0.00	0.00	0.27	0.09	2	1.7	0.0	0.0	0.31	0.07	18	3.8	0.01	0.0	0.09	0.04
0.01	0.100	84	94.0	0.08	0.09	0.11	0.07	3	3.4	0.00	0.00	0.24	0.09	2	1.8	0.0	0.0	0.27	0.07	6	2.3	0.00	0.0	0.17	0.06
0.02	0.100	107	109.4	0.10	0.11	0.10	0.11	2	1.9	0.00	0.00	0.32	0.07	1	1.1	0.0	0.0	0.32	0.05	10	2.6	0.01	0.0	0.14	0.04
0.03	0.100	96	113.9	0.09	0.11	0.10	0.12	2	1.7	0.00	0.00	0.32	0.07	1	0.9	0.0	0.0	0.33	0.05	15	3.5	0.01	0.0	0.11	0.04
0.01	0.150	79	94.9	0.07	0.09	0.17	0.10	2	1.9	0.00	0.00	0.32	0.07	1	1.3	0.0	0.0	0.32	0.06	14	9.6	0.01	0.1	0.24	0.07
0.02	0.150	53	83.0	0.05	0.08	0.16	0.13	1	1.5	0.00	0.00	0.34	0.06	1	0.9	0.0	0.0	0.34	0.05	7	2.6	0.01	0.0	0.19	0.06
0.03	0.150	55	88.4	0.05	0.09	0.16	0.13	1	0.4	0.00	0.00	0.34	0.04	1	0.1	0.0	0.0	0.35	0.03	11	3.3	0.01	0.0	0.14	0.05

Table 2: Number of blocks selected as informative $\left(\tilde{b}\right)$, proportion of falsely selected blocks (fpr) and the misclassification rate (mc) averaged over 100 runs, presented as mean (m) and standard deviation (sd) for block size $p_0 = 10$.

		bHC						Fdr						Lfdr						Ofdr					
		$\tilde{b}$		fpr		mc		$\tilde{b}$		fpr		mc		$\tilde{b}$		fpr		mc		$\tilde{b}$		fpr		mc	
β	ϵ	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd	m	sd
0.010	0.005	36	41.1	0.03	0.04	0.02	0.02	11	5.9	0.00	0.00	0.06	0.08	6	3.0	0.00	0.00	0.07	0.07	9	1.0	0.00	0.00	0.03	0.02
0.020	0.005	42	64.5	0.03	0.06	0.04	0.03	12	7.6	0.00	0.00	0.09	0.09	6	3.9	0.00	0.00	0.13	0.08	14	2.7	0.00	0.00	0.04	0.02
0.030	0.005	94	106.6	0.08	0.10	0.04	0.05	7	7.7	0.00	0.00	0.17	0.11	4	5.3	0.00	0.00	0.22	0.10	19	3.8	0.01	0.00	0.05	0.03
0.010	0.010	39	42.1	0.03	0.04	0.03	0.03	9	6.1	0.00	0.00	0.10	0.10	5	3.0	0.00	0.00	0.12	0.09	9	1.3	0.00	0.00	0.04	0.02
0.020	0.010	62	68.8	0.05	0.07	0.04	0.05	7	7.2	0.00	0.00	0.15	0.11	3	3.2	0.00	0.00	0.20	0.09	12	2.9	0.00	0.00	0.07	0.03
0.030	0.010	99	103.7	0.09	0.10	0.05	0.08	3	4.0	0.00	0.00	0.26	0.10	2	2.9	0.00	0.00	0.28	0.08	16	3.7	0.01	0.00	0.07	0.04
0.010	0.050	53	55.0	0.05	0.05	0.06	0.07	2	2.1	0.00	0.00	0.26	0.08	1	0.7	0.00	0.00	0.28	0.05	24	140.1	0.02	0.14	0.18	0.08
0.020	0.050	30	58.2	0.03	0.06	0.14	0.11	1	0.4	0.00	0.00	0.33	0.05	1	0.1	0.00	0.00	0.33	0.04	16	99.4	0.01	0.10	0.18	0.07
0.030	0.050	21	24.5	0.02	0.02	0.16	0.13	1	0.4	0.00	0.00	0.34	0.04	1	0.2	0.00	0.00	0.34	0.04	10	3.5	0.01	0.00	0.14	0.05
0.010	0.100	21	21.2	0.02	0.02	0.14	0.12	1	0.6	0.00	0.00	0.32	0.05	1	0.4	0.00	0.00	0.32	0.05	162	367.6	0.16	0.37	0.31	0.12
0.020	0.100	14	20.5	0.01	0.02	0.20	0.13	1	0.2	0.00	0.00	0.34	0.03	1	0.0	0.00	0.00	0.34	0.04	53	218.3	0.05	0.22	0.27	0.10
0.030	0.100	19	34.4	0.02	0.03	0.17	0.12	1	0.3	0.00	0.00	0.33	0.05	1	0.1	0.00	0.00	0.34	0.04	6	2.6	0.01	0.00	0.19	0.06
0.010	0.150	23	32.1	0.02	0.03	0.18	0.13	1	0.2	0.00	0.00	0.34	0.04	1	0.0	0.00	0.00	0.33	0.04	401	491.6	0.40	0.49	0.39	0.13
0.020	0.150	12	16.1	0.01	0.02	0.22	0.12	1	0.0	0.00	0.00	0.34	0.04	1	0.0	0.00	0.00	0.34	0.04	152	358.1	0.15	0.36	0.32	0.10
0.030	0.150	17	31.1	0.02	0.03	0.18	0.12	1	0.0	0.00	0.00	0.34	0.03	1	0.0	0.00	0.00	0.34	0.03	34	170.8	0.03	0.17	0.25	0.08

Table 3: Number of blocks selected as informative $\left(\tilde{b}\right)$, proportion of falsely selected blocks (fpr) and the misclassification rate (mc) averaged over 100 runs, presented as mean (m) and standard deviation (sd) for block size $p_0 = 20$.

7.2 Real cancer data

In this section we apply the three different thresholding methods bHC, Fdr and Lfdr to real data. We consider classification of four real gene expression data sets from cancer tumor samples.

Colon cancer data: The first set of data consists of filtered and processed gene expression information from colon cancer data of which $p = 2000$ genes were chosen with largest minimal intensity over $n = 62$ tissue samples. As a class variable we use non-tumor vs. tumor tissue with $n_0 = 22$ and $n_1 = 40$ observations, respectively. Additional information about tissue samples and the preprocessing procedure can be found in [1].

Breast cancer data I: For the second set of data, the tumor samples were selected from the population-based Stockholm-Gotland breast cancer registrary on the basis of several criteria; for more details on the data see [19]. This data matrix monitors $p = 6573$ genes in $n = 159$ breast tumor samples and as a class variable we use 5-year breast-cancer death (true = death within 5 years).

Prostate cancer data: The third data is available through the Broad institute and consists of gene expression patterns from 52 tumorous and 50 normal prostate specimens for approximately 12,600 genes; for more details see [22].

Breast cancer data II: For the last data the tumor samples were selected from the Duke Breast Cancer SPORE tissue bank on the basis of several criteria; for more details on the data see [29]. This data matrix monitors $p = 7129$ genes in $n = 49$ breast tumor samples and as class variable we use whether the tumors are positive or negative estrogen receptors.

The data was scaled and centered. Next the data was divided into blocks. We test eight different block sizes ($p_0 = [2, 5, 10, 15, 20, 30, 40, 50]$). Then we estimate the separation strength and order the blocks according to strength from the highest to the lowest. Using the thresholding methods we estimate number of informative blocks to include in the additive classifier. We divide the data into training and test data using 10-fold cross validation and calculate the misclassification according to the above; see Table 4. There are great differences in the number of selected blocks between bHC and the two false-discovery rate methods. The bHC consistently shows lower misclassification than both Fdr and Lfdr as well as compared to including the full data in the classifier, as can be seen in Table 4.

Next we choose features with the highest variance to be able to search for true block structures within the data. We calculated the variance for all feature variables, then the features were ranked according to the variance and we selected a subsample of the top 50% for the two breast cancer data and the top 25% for the prostate cancer data. For colon cancer the whole data set was included. Then we estimate the block-diagonal structure of $\hat{\Sigma}$ through gLasso with bootstrap and Cuthill-McKee ordering; for more information see [18]. All data sets were then arranged according to the estimated block structure. Then we followed the above procedure to estimate the number of informative blocks and misclassification; see Table 5. This approach is beneficial for the *Breast cancer data II*, where consistently low misclassification is showed over the different block sizes. For the *Prostate cancer data* there were better results estimating the threshold from the full data set, indicating that choosing the features with the highest variance was not optimal for this data. The *Breast cancer data II* shows similiar results for the two approaches. It is still bHC that performs with the lowest misclassification rates, but the two Fdr methods

come closer in results; see Table 5.

It is clearly seen that misclassification varies between block sizes and in this study we have limited ourselves to equal block sizes. Allowing for data-adjusted blocks could improve the classification even further. In comparison to published results in [5, 15] our bHC results in lower misclassification for some block sizes.

Block	No.selected blocks			Misclassification rate			
size	bHC	Fdr	Lfdr	bHC	Fdr	Lfdr	All
Breast cancer data I							
1	*657*	*999*	*583*	*0.24*	*0.23*	*0.24*	-
2	328	1461	804	0.23	0.23	**0.22**	0.28
5	131	1219	929	**0.22**	0.27	0.25	0.26
10	65	657	601	**0.20**	0.26	0.25	0.28
15	43	438	393	**0.21**	0.28	0.26	0.28
20	32	328	308	**0.19**	0.28	0.28	0.26
30	21	219	209	**0.14**	0.26	0.25	0.30
40	16	164	156	**0.17**	0.26	0.25	0.26
50	13	131	121	**0.15**	0.25	0.26	0.25
Prostate cancer data							
1	*126*	*808*	*442*	*0.10*	*0.20*	*0.13*	-
2	630	5547	4082	**0.14**	0.38	0.35	0.36
5	252	2520	2427	**0.09**	0.26	0.28	0.25
10	126	1260	1254	**0.08**	0.19	0.19	0.17
15	84	840	838	**0.07**	0.14	0.13	0.15
20	63	630	620	**0.06**	0.13	0.12	0.13
30	42	420	411	**0.05**	0.11	0.13	0.10
40	31	315	309	**0.05**	0.11	0.09	0.12
50	25	252	247	**0.04**	0.12	0.09	0.13
Breast cancer data II							
1	*712*	*165*	*61*	*0.02*	*0.08*	*0.04*	-
2	712	1254	696	0.06	0.06	**0.04**	0.10
5	285	1313	759	**0.04**	0.08	0.06	0.14
10	142	712	705	**0.04**	0.12	0.10	0.08
15	95	475	443	**0.06**	0.10	0.14	0.16
20	71	356	351	**0.02**	0.16	0.10	0.10
30	1	237	235	0.10	**0.08**	**0.08**	0.10
40	1	178	176				
50							

Table 4: Number of blocks selected as informative and misclassification using whole data sets.

Block size	No.inf.blocks bHC	Fdr	Lfdr	Misclassification rate bHC	Fdr	Lfdr	All
Colon cancer data							
1	*200*	*388*	*240*	*0.13*	*0.13*	*0.13*	-
2	198	994	664	**0.15**	0.26	0.26	0.27
5	79	89	87	**0.08**	0.19	0.18	0.15
10	39	198	196	0.13	**0.11**	0.16	0.18
15	26	132	130	**0.15**	0.16	0.18	0.19
20	19	99	1	**0.11**	0.13	0.21	**0.11**
30	13	66	64	**0.11**	0.18	0.15	0.18
40	9	49	1	**0.10**	0.18	0.19	0.18
50	1	39	37	0.32	0.27	0.19	**0.18**
Breast cancer data I							
1	*328*	*550*	*315*	*0.22*	*0.21*	*0.23*	-
2	164	614	342	0.22	0.22	**0.21**	0.28
5	65	565	384	**0.21**	0.26	0.23	0.26
10	32	328	286	**0.21**	0.30	0.28	0.27
15	21	219	188	**0.20**	0.27	0.28	0.26
20	16	164	152	**0.16**	0.28	0.27	0.28
30	10	109	105	**0.18**	0.24	0.28	0.26
40	8	82	77	**0.18**	0.25	0.27	0.26
50	6	65	60	**0.20**	0.27	0.24	0.26
Prostate cancer data							
1	*315*	*89*	*43*	*0.36*	*0.34*	*0.28*	-
2	157	150	54	0.27	0.30	**0.23**	0.38
5	63	260	99	**0.16**	0.25	0.20	0.31
10	31	218	101	**0.09**	0.19	0.14	0.25
15	21	188	127	**0.06**	0.21	0.16	0.23
20	15	155	134	**0.11**	0.18	0.16	0.19
30	10	105	97	0.11	0.14	**0.10**	0.15
40	7	78	72	**0.09**	0.10	0.12	0.13
50	6	63	60	0.15	**0.14**	**0.14**	0.16
Breast cancer data II							
1	*356*	*230*	*133*	*0.22*	*0.20*	*0.22*	-
2	307	217	97	**0.02**	**0.02**	**0.02**	0.27
5	142	461	289	**0.02**	0.06	**0.02**	0.14
10	71	356	271	**0.02**	0.12	0.06	0.14
15	47	237	225	**0.04**	0.12	0.16	0.12
20	35	178	174	**0.02**	0.12	0.12	0.18
30	1	118	115	0.12	0.10	**0.08**	0.10
40	1	89	87	0.18	0.16	**0.12**	0.16
50							

Table 5: Number of selected blocks and misclassification using a subset of the real data where true block structure has been identified.

8 Concluding remarks

The results obtained in this paper constitute a part of study on high-dimensional classification, where sparsity patterns in feature relevance are combined with feature weakness motivating the suggested SWB model. Further, the model was validated using both simulated and real datasets. We could see that that the detectable region for SWB decreases with growing degrees of freedom, hence, in the case of weak separation strength, the block size should be kept moderate.

The empirical analysis of the suggested feature selection procedure shows that bHC thresholding has near optimal behavior for a variety of block sizes. This behavior was studied using the *ideal* threshold interpretation, meaning that with knowledge of n, p, β and ω the LR-based procedure choose the very best threshold within the detectable area. Since the bHC procedure "mimics" behavior of LR while not requiring knowledge of parameters, we conclude that bHC is optimally adaptive to unknown sparsity and separation strength. These results are in line with e.g. [27] and [4], where optimal adaptivity of the HC procedure was considered in the context of signal detection.

The bHC principle has also been motivated by our results of estimation of the proportion of informative blocks, i.e proportion of non nulls in [17]. We embed the technique by [23, 17] into high-dimensional sparse classification and show that the problem of estimation becomes more subtle, in particular due to problems with reliable estimates of $\hat{\beta}$.

HC was previously applied for feature selection in high-dimensional classification; see [5, 10]. However, a crucial difference is in selecting blocks of features (not individual features) at the learning stage; our procedure selects blocks when designing a sample-based linear classifier and the actual allocation decision is made by the classifier when presented with a new observed feature vector.

An essential part of the current work is applications, especially using bHC in classification of cancer data. A number of studies in the problem report promise asymptotic results, with a warning that it could be difficult to demonstrate the observed phenomena in the finite sample case; see e.g. [10] Applications of HC to realistic data are even more restrictive due to the assumptions of independence. We have investigated our bHC thresholding in classification of tumor samples for four real data sets with varying sparsity and amplitude of separation strength, and showed that for a number of cases our techniques outperform existing feature selection procedures, giving lower misclassification rates. It is also important to note that our bHC technique requires an extra stage of structure learning. That the bHC classifier matches or outperforms HC based on the independence assumption clearly indicates the relevance of the SWB model and the benefits of properly taking into account the dependence structure underlying the data. In this study we limited ourselves to equal block sizes, but it can easily be extended to allow for different block sizes, which can further improve the results.

References

[1] Uri Alon, Naama Barkai, Daniel A. Notterman, Kurt Gish, Sunny Ybarra, Deborah Mack, and Arnold J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS*, 96(12):6745–6750, June 1999.

[2] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):pp. 289–300, 1995.

[3] Elizabeth Cuthill and J. McKee. Reducing the bandwidth of sparse symmetric matrices. In *Proceedings of the 1969 24th national conference*, ACM '69, pages 157–172, New York, NY, USA, 1969. ACM.

[4] David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist*, pages 962–994, 2004.

[5] David Donoho and Jiashun Jin. Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences*, 105(39):14790–14795, 2008.

[6] David Donoho and Jiashun Jin. Feature selection by higher criticism thresholding achieves the optimal phase diagram. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4449–4470, 2009.

[7] Bradley Efron. Local false discovery rates. Technical report, Department of Statistics, Stanford University, 2004.

[8] Bradley Efron, John D. Storey, and Robert Tibshirani. Microarrays, empirical bayes methods, and false discovery rates. *Genet. Epidemiol*, 23:70–86, 2001.

[9] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[10] Peter Hall and Jiashun Jin. Properties of higher criticism under strong dependence. *The Annals of Statistics*, 36(1):381–402, 2008.

[11] Yu. I. Ingster. Minimax detection of a signal for l_n^p balls. *Mathematical Methods of Statistics*, 7:401–428, 1999.

[12] Pengsheng Ji and Jiashun Jin. Ups delivers optimal phase diagram in high-dimensional variable selection. *Annals of Statistics*, 40(1):3107–3123, 2012.

[13] Jiashun Jin. *Detecting and Estimating Sparse Mixtures*. PhD thesis, Stanford University, 2003.

[14] Markus Kalisch and Peter Buhlmann. Robustification of the pc-algorithm for directed acyclic graphs. *Journal of Computational and Graphical Statistics*, 17(4):773–789, 2008.

[15] Bernd Klaus and Korbinian Strimmer. Signal identification for rare and weak features: higher criticism or false discovery rates? *Biostatistics*, 14:129, 2013.

[16] Nicolai Meinshausen and Peter Buhlmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

[17] Nicolai Meinshausen and John Rice. Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Annals of Statistics*, 34(1):373–393, 2006.

[18] Tatjana Pavlenko, Anders Bjorkstrom, and Annika Tillander. Covariance structure approximation via glasso in high-dimensional supervised classification. *Journal of Applied Statistics*, 39(8):1643–1666, 2012.

[19] Yudi Pawitan, Judith Bjohle, Lukas Amler, AnnaLena Borg, Suzanne Egyhazi, Per Hall, Xia Han, Lars Holmberg, Fei Huang, Sigrid Klaar, Edison T. Liu, Lance Miller, Hans Nordgren, Alexander Ploner, Kerstin Sandelin, Peter M. Shaw, Johanna Smeds, Lambert Skoog, Sara Wedren, and Jonas Bergh. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Research*, 7(6):953–964, 2005.

[20] Yudi Pawitan, Karuturi R. Krishna Murthy, Stefan Michiels, and Alexander Ploner. Bias in the estimation of false discovery rate in microarray studies. *Bioinformatics*, 21(20):3865–3872, 2005.

[21] Philipp Rutimann and Peter Buhlmann. High dimensional sparse covariance estimation via directed acyclic graphs. *Electron. J. Statist.*, 3:1133–1160, 2009.

[22] Dinesh Singh, Phillip G. Febbo, Kenneth Ross, Donald G. Jackson, Judith Manola, Christine Ladd, Pablo Tamayo, Andrew A. Renshaw, Anthony V. D'Amico, and Jerome P. Richie. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203–209, March 2002.

[23] John D. Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[24] Korbinian Strimmer. fdrtool: a versatile r package for estimating local and tail area-based false discovery rates. *Bioinformatics*, 24(1):1461–1462, 2008.

[25] Korbinian Strimmer. A unified approach to false discovery rate estimation. *BMC Bioinformatics*, 9(1):303, 2008.

[26] Wenguang Sun and T. Tony Cai. Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102(479):901–912, 2007.

[27] T. Tony Cai, X. Jessie Jeng, and Jiashun Jin. Optimal detection of heterogeneous and heteroscedastic mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):629–662, 2011.

[28] John W. Tukey. T13 n: The higher criticism. Course Notes, Statistics 411, Princeton University, 1976.

[29] Mike West, Carrie Blanchette, Holly Dressman, Erich Huang, Seiichi Ishida, Rainer Spang, Harry Zuzan, John A. Olson, Jeffrey R. Marks, and Joseph R. Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *PNAS*, 98(20):11462–11467, September 2001.

[30] Jichun Xie, T. Tony Cai, and Hongzhe Li. Sample size and power analysis for sparse signal recovery in genome-wide association studies. *Biometrika*, 98(2):273–290, 2011.