Higher criticism for estimating proportion of non-null effect in high-dimensional multiple comparison

Annika Tillander
joint work with Tatjana Pavlenko

Karolinska Institutet, Department of Medical Epidemiology and Biostatistics

LinStat 26 August 2014

Outline

- Introduction
- Covariance structure approximation
- Separation strength
- Detection boundary
- Estimating proportion of non-null effect
- Higher criticism
- Application

Classification

We have n observations where each observation $\mathbf{x} = (x_1, \ldots, x_p)$ corresponds to p number of features. Supervised classification problem with $\mathcal{C}$ classes, class label $y_j = c$ where $j = 1, \ldots, n$, $c \in \{1, \ldots, \mathcal{C}\}$

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

Using the training data a prediction model is built which enables prediction of new observations where the outcome is unknown.

Linear Discriminant Analysis (LDA)

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.

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

π_c is the prior probability of class c and $\sum_{c=1}^{\mathcal{C}} \pi_c = 1$.

$$c^* = \text{argmax}_{c=1,\dots,\mathcal{C}} D_c(\mathbf{x})$$

Two classes

Assign $\mathbf{x}$ to class 1 if

$$\log \frac{\pi_1}{\pi_2} + \left(\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right)' \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geq 0$$

High dimensionality

It is the relation between the number of observation (n) and the number of features (p) that decides the dimensionality of the data.
In the case with more features than available observations, the problem is said to be ”high-dimensional” ($p > n$).

Standard asymptotic

p is fixed and $n \to \infty$

Asymptotic when p is not fixed

$p = p_n$ grows with n, $p_n \gg n$ for $n \to \infty$

gLasso

Using the correlation matrix $\mathcal{K}^{-1}$ instead of covariance allows for faster convergence in high-dimensional setting Rothman et al. (2008).

Let Γ denote the diagonal matrix of true standard deviations

$$\Sigma^{-1} = \Gamma^{-1}\mathcal{K}\Gamma^{-1}$$

Sparse inverse covariance estimation with the graphical lasso Friedman et al. (2009)

$$\hat{\mathcal{K}}_\lambda = \arg\min_{\mathcal{K}\succ 0}\left\{\operatorname{trace}\left(\mathcal{K}\hat{\mathcal{K}}^{-1}\right) - \log\det\mathcal{K} + \lambda\left\|\mathcal{K}^{-1}\right\|_1\right\}$$

Cuthill-McKee ordering

Reducing the bandwidth of sparse symmetric matrices Cuthill & McKee (1969)

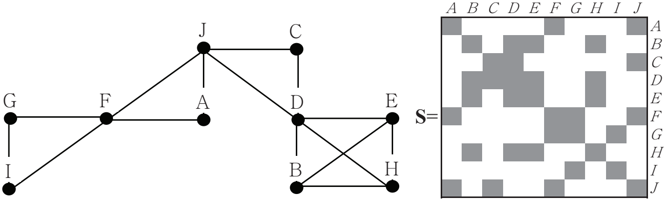
Let $\mathbf{S}$ be a $p \times p$ symmetric matrix where i denote rows and j denote columns.

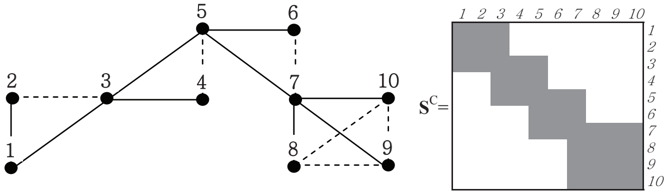
The bandwidth of $\mathbf{S}$ is the maximum value of $|i - j|$ for the non-zero elements

Determine a permutation matrix $\mathcal{P}$ such that non-zero elements will cluster about the main diagonal

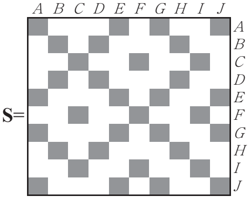
$\mathbf{S}^C = \mathcal{P}\mathbf{S}\mathcal{P}^T$

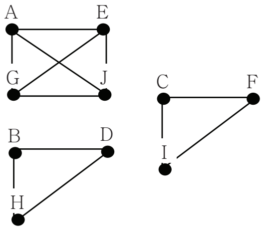
$\mathbf{S}^C = \mathcal{P}\mathbf{S}\mathcal{P}^T$

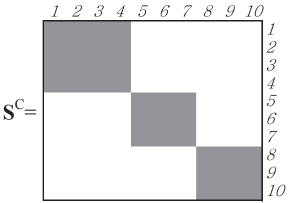




True block-structure







Algorithm

Combining gLasso and Cuthill-McKee ordering

Bootstrap sample, calculate $\hat{\mathcal{K}}_j^{-1}$

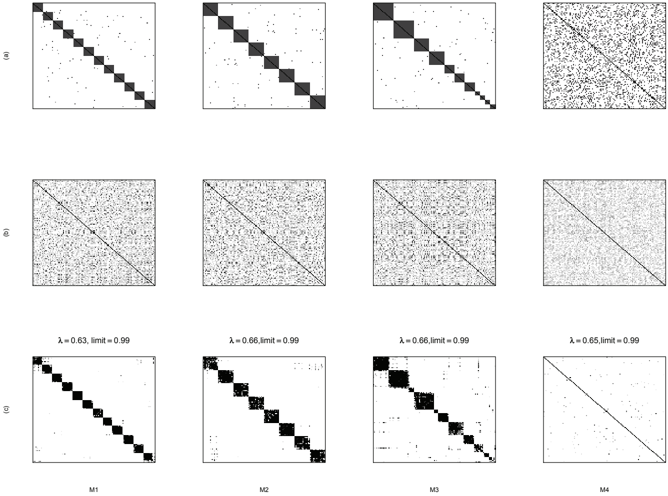
Estimate $\hat{\mathcal{K}}_j\left[\lambda_i\right]$ with gLasso

$\mathbf{S}_{ij} = \mathbf{1}_{\hat{\mathcal{K}}_j[\lambda_i]>0}$

$\tilde{\mathbf{S}}_{ik} = \mathbf{1}_{\frac{\sum_{j=1}^{r} \mathbf{s}_{ij}}{r} > q_k}$

Find permutation matrix $\mathcal{P}_{ik}$ for skeleton $\tilde{\mathbf{S}}_{ik}$ with Cuthill-McKee ordering algorithm

Identifying block-structure



Additive classifier

Block diagonal segmentation

$$\Sigma^{-1} = \text{diag}\left[\Sigma_1^{-1}, \ldots, \Sigma_b^{-1}\right]$$

where b is the number of blocks. 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 $\mathbf{x}_i = (x_{i_1}, \ldots, x_{i_{p_i}})$, p_i is the block size, $i = 1, \ldots, b$, such that for any $i \neq j$, $\mathbf{x}_i$ and $\mathbf{x}_j$ are conditionally independent given the class variable y.

Two-class linear function with additive structure

$$D(\mathbf{x}) = \sum_{i=1}^{b} \left(\mathbf{x}_i - \tfrac{1}{2}(\boldsymbol{\mu}_{1,i} + \boldsymbol{\mu}_{2,i})\right)' \Sigma_i^{-1}(\boldsymbol{\mu}_{1,i} - \boldsymbol{\mu}_{2,i})$$

Mahalanobis distance

Let $\pi_1 = \pi_2 = 1/2$ then the optimal misclassification probability can be expressed as

$$\varepsilon = \Phi\left(-\frac{1}{2}\sqrt{\delta^2}\right)$$

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.

Separation strength

The ith block separation strength

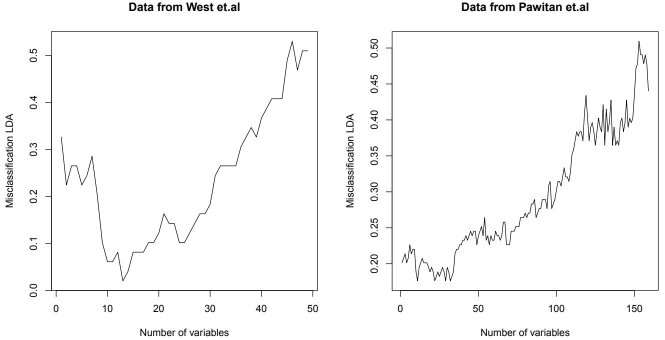
$$\delta_i^2 = \left\| \Sigma_i^{-1/2} \boldsymbol{\mu}_i \right\|^2$$

Rescaled estimate of the ith separation strength

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

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. $\mathcal{S}^2 \sim \chi^2(p_0, \omega^2)$ where p_0 degrees of freedom and $\omega^2 = \eta\delta^2$ the non-centrality parameter.

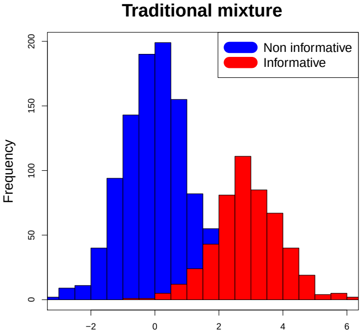
Misclassifiation

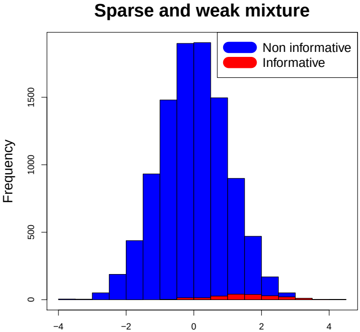


[1]

[1] Data from West et al. (2001) and Pawitan et al. (2005)

Sparse and weak setting





Detection boundary

β: sparsity parameter, proportion of non-null effect
ω^2: weakness parameter, effect strength

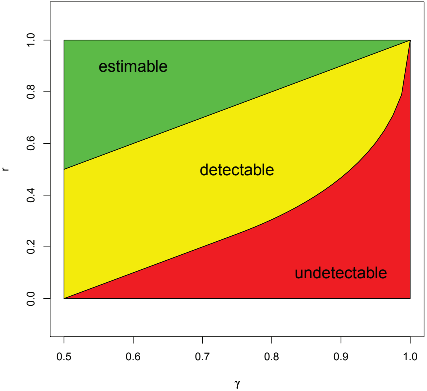
Asymptotic sparse and weak model

$\beta = b^{-\gamma}$ where $\gamma \in (0, 1)$
$\omega^2 = 2r \log b$ where $r \in (0, 1)$

For $p_0 = 1$ and $p = b$

$$H_0: \quad \mathcal{S}_i \sim N(0,1) \text{ i.i.d } 1 \leq i \leq p$$
$$H_1: \quad \mathcal{S}_i \sim (1-\beta)N(0,1) + \beta N(\sqrt{\omega^2}, 1) \text{ i.i.d } 1 \leq i \leq p$$

Ingster (1999), Donoho & Jin (2008), Klaus & Strimmer (2013)



Detection boundary for blocks

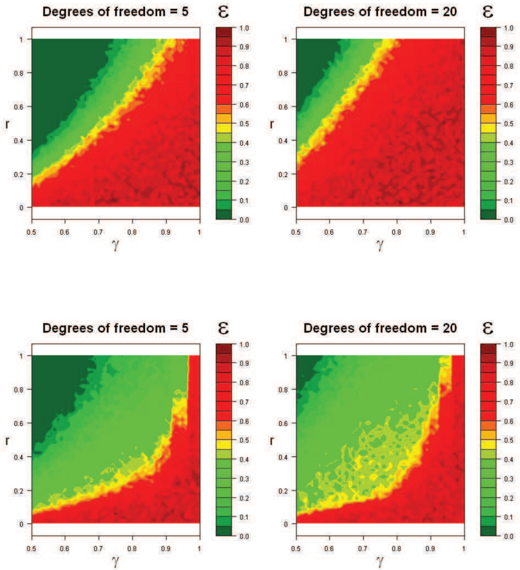
$$H_0: \quad \mathcal{S}_i^2 \sim \chi^2_{p_0}(\cdot;0) \text{ i.i.d, } 1 \le i \le b$$
$$H_1: \quad \mathcal{S}_i^2 \sim (1-\beta)\chi^2_{p_0}(\cdot;0) + \beta\chi^2_{p_0}\left(\cdot;\omega^2\right) \text{ i.i.d, } 1 \le i \le b$$

$$LR_i = \frac{f_{H_1}(s_i^2)}{f_{H_0}(s_i^2)} = \frac{(1-\beta)\chi^2_{p_0}(s_i^2;0) + \beta\chi^2_{p_0}(s_i^2;\omega^2)}{\chi^2_{p_0}(s_i^2;0)}$$
$$LR_b = LR_b(s_1^2, s_2^2, \ldots, s_b^2; \beta, \omega^2)$$

Reject H_0 iff

$$\log(LR_b) > 0$$

Detection boundary for blocks



Estimating proportion of non-null effect

Let π_i denote the significance level ("*p*-value")

Under the global null hypothesis

$$\pi_i \sim U(0,1)$$

where U denotes the uniform distribution.

β denote the proportion false null hypotheses

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

Estimating proportion of non-null effect

Rank the p-values in increasing order $\pi_{(1)} \leq \pi_{(2)} \leq \ldots \leq \pi_{(b)}$

Storey & Tibshirani (2003)

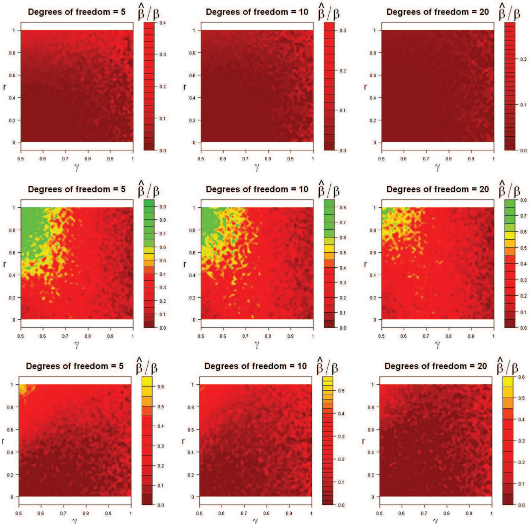
$$\hat{\beta} = 1 - \frac{\sum_{1=i}^{b} \mathbf{1}\left\{\pi_{(i)} > t\right\}}{(1-t)b}$$

Meinshausen & Rice (2006)

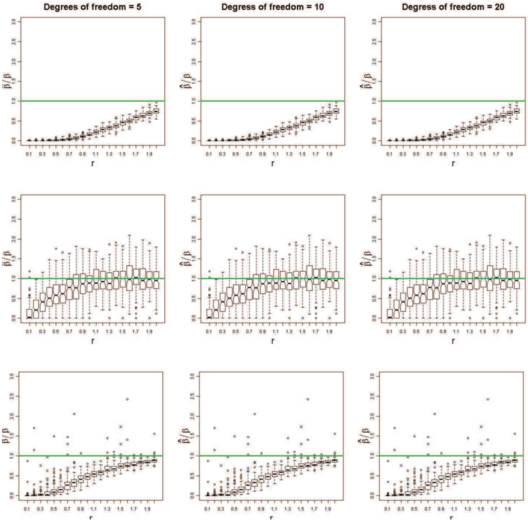
$$\hat{\beta} = \max_{t \in (0,1)} \frac{F_b(t) - t - \mathrm{B}_{b,\alpha}\Delta(t)}{1-t}$$

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 $\mathrm{B}_{b,\alpha} = \frac{G^{-1}(1-\alpha)+m_b}{n_b}$ where G is the Gumbel distribution, $m_b = 2\log_2 b + \frac{1}{2}\log_3 b - \frac{1}{2}\log 4\pi$

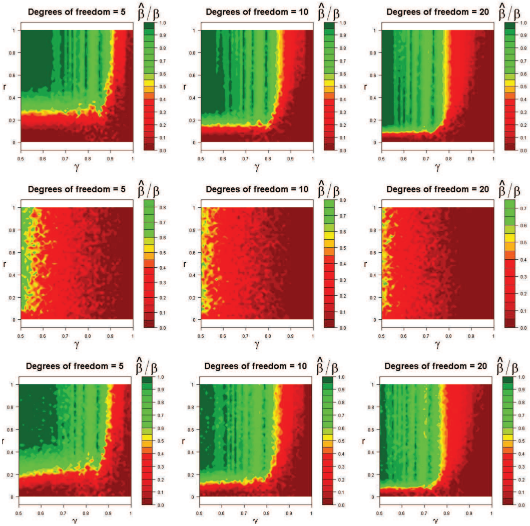
Estimating proportion of non-null effect



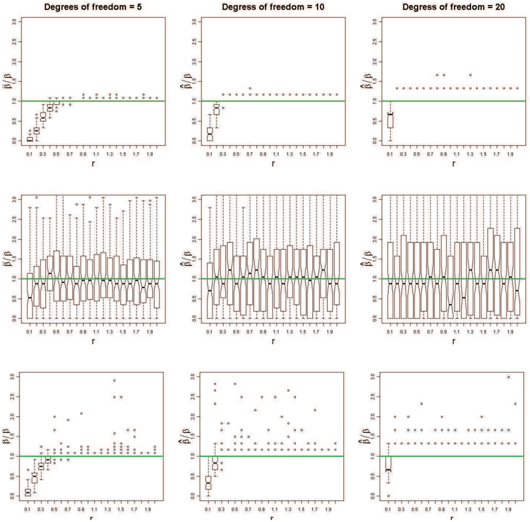
Estimating proportion of non-null effect



Estimating proportion of non-null effect



Estimating proportion of non-null effect



Higher criticism

Donoho & Jin (2004, 2008, 2009)

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

$i = 1, \ldots, p$ and for fixed $\alpha_0 \in (0,1)$ the HC test statistic is

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

Block higher criticism

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

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

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

False Discovery Rates

Benjamini & Hochberg (1995)

$$\mathrm{Fdr}(\pi_i) = P\left(\text{Non informative} \middle| \Pi \leq \pi_i\right) = \frac{(1-\beta)\pi_i}{F(\pi_i)} \leq \frac{p}{i}\pi_{(i)}$$

Efron et al. (2001), Efron (2004)

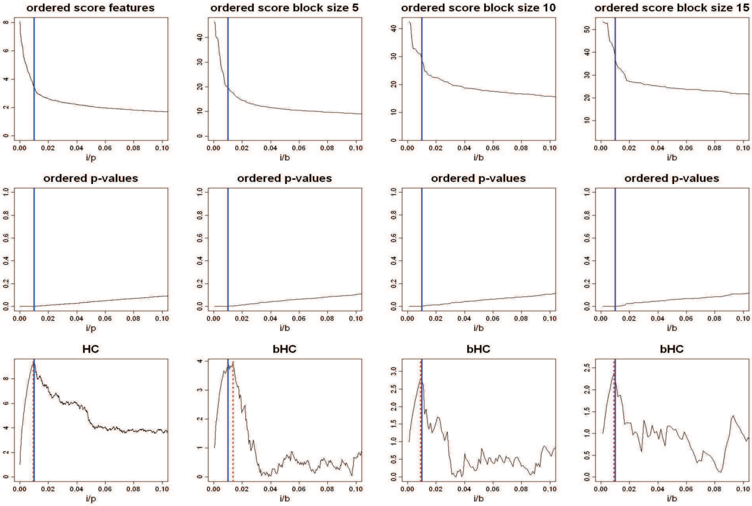
$$\mathrm{Lfdr}(\pi_i) = P\left\{\text{Non-informative} \middle| \pi_i\right\} = \frac{(1-\beta)f_{H_0}(\pi_i)}{f(\pi_i)}$$

Sun & Cai (2007)

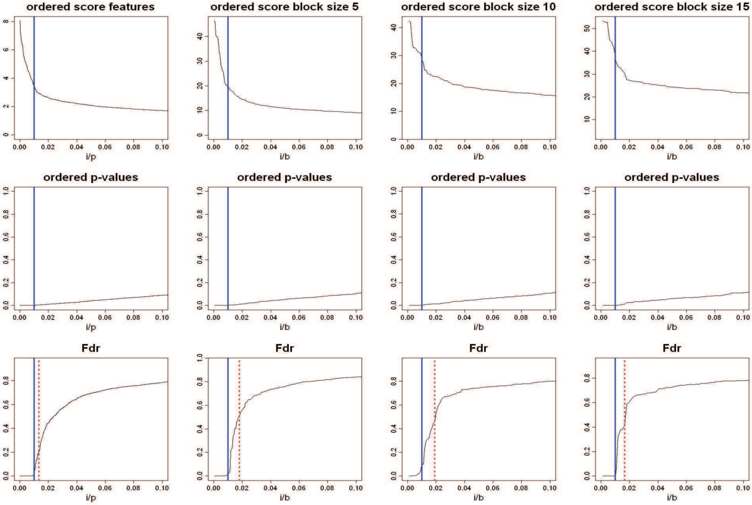
$$\mathrm{Ofdr}(s_i^2) = \frac{(1-\beta)\chi^2_{p_0}\left(s_i^2;0\right)}{(1-\beta)\chi^2_{p_0}\left(s_i^2;0\right)+\beta\chi^2_{p_0}\left(s_i^2;\omega^2\right)}$$

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

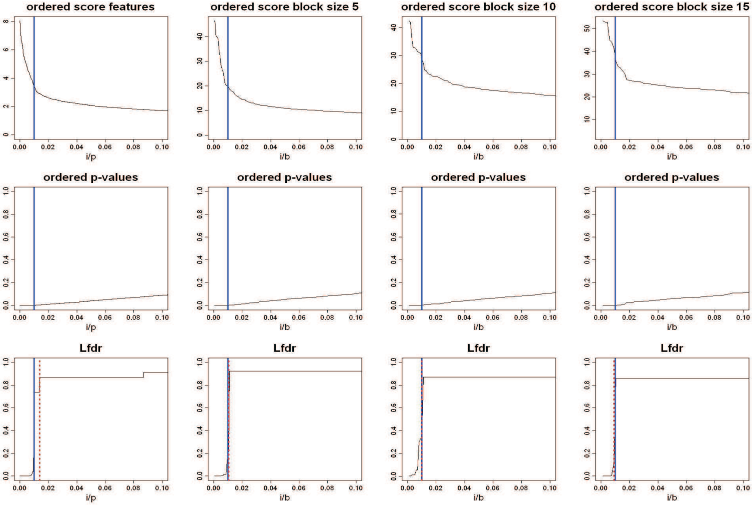
Simulated data



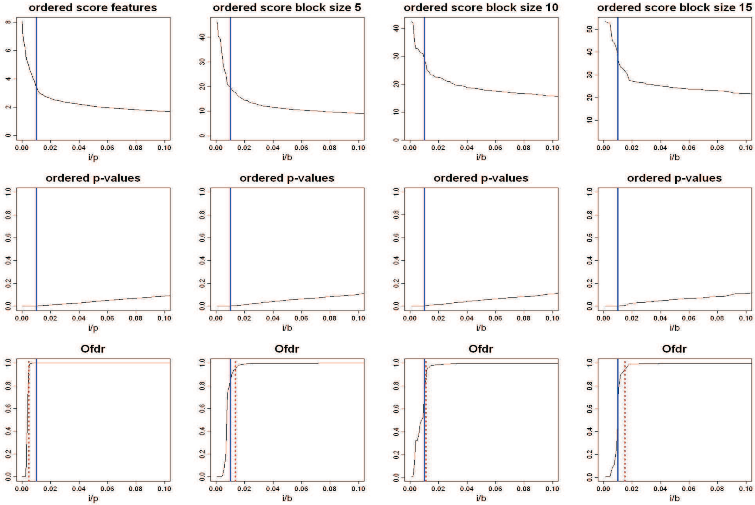
Simulated data



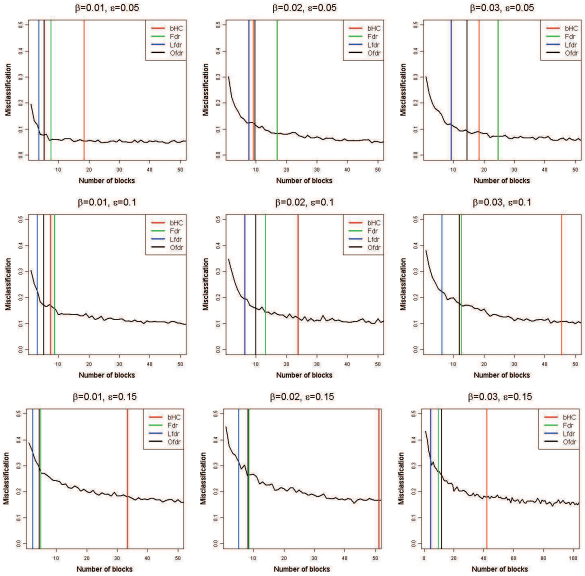
Simulated data



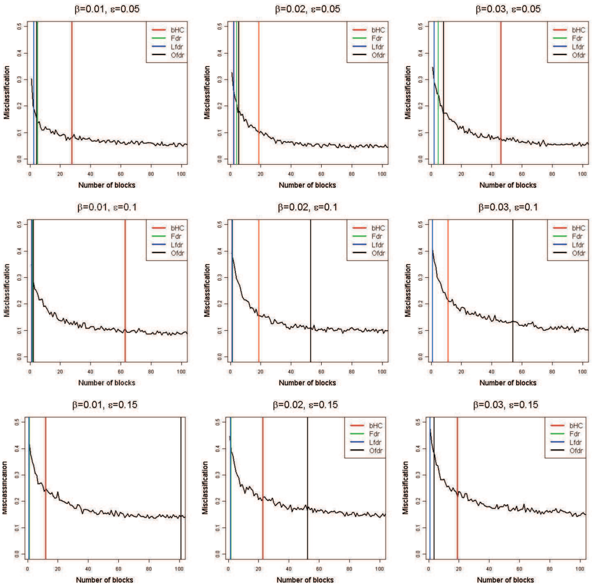
Simulated data



Simulated data



Simulated data



Simulated data

		bHC						Fdr					
		$\tilde{b}$		*fpr*		*mc*		$\tilde{b}$		*fpr*		*mc*	
β	ϵ	*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
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
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
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
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
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
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
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
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
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
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
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
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
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
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

Table: 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 $p_0 = 20$.

Real data

Block size	No.selected blocks bHC	Fdr	Lfdr	Misclassification rate 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							

Real data

Block size	No.inf.blocks bHC	Fdr	Lfdr	Misclassification rate bHC	Fdr	Lfdr	All
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							

Thank You

Benjamini, Y. & Hochberg, Y. (1995), ‘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.
URL: *http://www.jstor.org/stable/2346101*

Cuthill, E. & McKee, J. (1969), Reducing the bandwidth of sparse symmetric matrices, *in* ‘Proceedings of the 1969 24th national conference’, ACM ’69, ACM, New York, NY, USA, pp. 157–172.
URL: *http://doi.acm.org/10.1145/800195.805928*

Donoho, D. & Jin, J. (2004), ‘Higher criticism for detecting sparse heterogeneous mixtures’, *Ann. Statist* pp. 962–994.

Donoho, D. & Jin, J. (2008), ‘Higher criticism thresholding: Optimal feature selection when useful features are rare and weak’, *Proceedings of the National Academy of Sciences* **105**(39), 14790–14795.
URL: *http://www.pnas.org/content/105/39/14790.abstract*

Donoho, D. & Jin, J. (2009), ‘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.
URL: *http://rsta.royalsocietypublishing.org/content/367/1906/4449.abstract*

Efron, B. (2004), Local false discovery rates, Technical report, Department of Statistics, Stanford University.

Efron, B., Storey, J. D. & Tibshirani, R. (2001), ‘Microarrays, empirical bayes methods, and false discovery rates’, *Genet. Epidemiol* **23**, 70–86.

Friedman, J., Hastie, T. & Tibshirani, R. (2009), *Graphical lasso- estimation of Gaussian graphical models*. Manual to the R-package glasso.

Ingster, Y. I. (1999), ‘Minimax detection of a signal for l_n^p balls’, *Mathematical Methods of Statistics* **7**, 401–428.

Klaus, B. & Strimmer, K. (2013), 'Signal identification for rare and weak features: higher criticism or false discovery rates?', *Biostatistics* **14**, 129.

Meinshausen, N. & Rice, J. (2006), 'Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses', *Annals of Statistics* **34**(1), 373–393.

Pawitan, Y., Bjohle, J., Amler, L., Borg, A., Egyhazi, S., Hall, P., Han, X., Holmberg, L., Huang, F., Klaar, S., Liu, E. T., Miller, L., Nordgren, H., Ploner, A., Sandelin, K., Shaw, P. M., Smeds, J., Skoog, L., Wedren, S. & Bergh, J. (2005), '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.

Rothman, A., Bickel, P., Levina, E. & Zhu, J. (2008), 'Sparse permutation invariant covariance estimation', *Electronic Journal of Statistics* **2**, 494–515.

Storey, J. D. & Tibshirani, R. (2003), 'Statistical significance for genomewide studies', *Proceedings of the National Academy of Sciences* **100**(16), 9440–9445.
URL: *http://www.pnas.org/content/100/16/9440.abstract*

Sun, W. & Cai, T. T. (2007), 'Oracle and adaptive compound decision rules for false discovery rate control', *Journal of the American Statistical Association* **102**(479), 901–912.
URL: *http://amstat.tandfonline.com/doi/abs/10.1198/016214507000000545*

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J. A., Marks, J. R. & Nevins, J. R. (2001), 'Predicting the clinical status of human breast cancer by using gene expression profiles', *PNAS* **98**(20), 11462–11467.