Change-Point Detection with Local Trend Adjustment

Shengji Jia[1], Chunming Zhang[2*] and Yiming Tang[1*]

[1]*Shanghai Lixin University of Accounting and Finance*

[2]*University of Wisconsin-Madison*

Supplementary Material

S1 Conditions and proofs of main results

The following technical conditions are imposed. They are not the weakest possible conditions, but they are imposed to facilitate the proofs. Recall $\mathbf{X}_1 = (\mathbf{x}_1, \ldots, \mathbf{x}_{n-1})$, and suppose the response vector $\mathbf{y}$ and all the covariates (columns) of $\mathbf{X}_1$ are centered, which are denoted by $\overline{\mathbf{y}}$ and $\overline{\mathbf{X}}_1 = (\overline{\mathbf{x}}_1, \ldots, \overline{\mathbf{x}}_{n-1})$. For any subset $A \subseteq \{1, 2, \ldots, n-1\}$, let $\overline{\mathbf{X}}_{1,A}$ be the matrix formed by concatenating the columns $\{\overline{\mathbf{x}}_i : i \in A\}$ indexed by A. Let $\mathbf{x}_0^*, \ldots, \mathbf{x}_{n-1}^*$ be the discrete Fourier transform of $\mathbf{x}_0, \ldots, \mathbf{x}_{n-1}$, respectively.

(C1) There exists an integer n_0 such that

$$\left\{n/(\log(\log(n)))^4\right\}^{-1} \sum_{i=n_0}^{n/(\log(\log(n)))^4} x_{ij}^{*2} = O(1), \text{ for } j = 0, 1, \ldots, n-1,$$

* Corresponding author.

where x_{ij}^* is the jth element of the vector $\mathbf{x}_i^*$;

(C2) $\widehat{\sigma}^2 = \sigma^2 + O_{\mathrm{P}}(n^{1/2})$;

(C3) There exists some $c_1 \in (0, 1]$ such that

$$\|\overline{\mathbf{X}}_{1,S^c}^T \overline{\mathbf{X}}_{1,S} (\overline{\mathbf{X}}_{1,S}^T \overline{\mathbf{X}}_{1,S})^{-1}\|_\infty \leq 1 - c_1,$$

where $\|\mathbf{M}\|_\infty := \max_{i=1,\ldots,m} \sum_{j=1}^n |M_{ij}|$ for an $m \times n$ matrix $\mathbf{M}$;

(C4) There exists some $c_2 > 0$ such that $\lambda_{\min}\big(\frac{1}{n}\overline{\mathbf{X}}_{1,S}^T \overline{\mathbf{X}}_{1,S}\big) \geq c_2$, where $\lambda_{\min}(\cdot)$ denotes the minimal eigenvalue of a matrix;

(C5) The minimum value of the regression coefficient vector on its support S satisfies

$$\min_{i \in S} |\beta_i| \geq \lambda \|(\frac{1}{n}\overline{\mathbf{X}}_{1,S}^T \overline{\mathbf{X}}_{1,S})^{-1}\|_\infty + \frac{4\lambda\sigma}{\sqrt{c_2}};$$

(C6) The tuning parameter λ satisfies

$$\lambda > \frac{2\sqrt{2}\sigma}{c_1} \sqrt{\frac{\log(n)}{n}}.$$

As for the common change-point detection for array-based data, let $\mathbb{X} \equiv \mathbf{I}_d \otimes \overline{\mathbf{X}}_1$, where $\mathbf{I}_d$ denotes a $d \times d$ identity matrix and $\otimes$ denotes the Kronecker product. For any subset $A \subseteq \{1, 2, \ldots, n-1\}$, let $\mathbb{X}_A$ be the matrix formed by concatenating the columns of $\mathbb{X}$ indexed by $\{a + b(n-1) : a \in A,\ b = 0, 1, \ldots, d-1\}$. Now the conditions (C2)–(C6) will be replaced by the following ones:

(C2′) $\widehat{\sigma}_i^2 = \sigma_i^2 + O_{\mathrm{P}}(n^{1/2})$, for $i = 1, \ldots, d$;

(C3′) There exists some $c_1 \in (0, 1)$ such that $\forall i \in S^c$,

$$\|\mathbb{X}_{\{i\}}^T \mathbb{X}_S (\mathbb{X}_S^T \mathbb{X}_S)^{-1}\|_2 \leq \frac{1 - c_1}{\sqrt{|S|}};$$

(C4′) There exists some $c_2 > 0$ such that $\lambda_{\min}\big(\frac{1}{n}\mathbb{X}_S^T \mathbb{X}_S\big) \geq c_2$;

(C5′) $\alpha := \min_{j \in S} \|\boldsymbol{\beta}_{\bullet,j}\|_2$ satisfies

$$\frac{1}{\alpha}\Big\{\sqrt{\frac{\log(d|S|)}{n}} + d\lambda\sqrt{|S|}\Big\} \to 0 \quad \text{as } n \to \infty;$$

(C6′) The tuning parameter λ satisfies

$$\frac{1}{\lambda}\sqrt{\frac{\log(dn - \sum_{j \in S} \|\boldsymbol{\beta}_{\bullet,j}\|_0)}{n}} \to 0 \quad \text{as } n \to \infty.$$

Proof of Theorem 1. After centering the response $\mathbf{y}$ and the columns of $\mathbf{X}_1$, the penalized least squares problem (2.5) reduces to the following standard Lasso problem:

$$\min_{\boldsymbol{\beta}_1} \frac{1}{2n}\|\overline{\mathbf{y}} - \overline{\mathbf{X}}_1 \boldsymbol{\beta}_1\|_2^2 + \lambda\|\boldsymbol{\beta}_1\|_1.$$

According to Theorem 1(b) of Wainwright (2009), conditions (C3)–(C6) imply that $p_n := \mathrm{P}(\widehat{S}^{\mathrm{ini}} = S) \geq 1 - 4\exp(-c_3\lambda^2 n)$ for some constant $c_3 > 0$. Under the event $\{\widehat{S}^{\mathrm{ini}} = S\}$, conditions (C1)–(C2) and Theorem 1 of Fan and Huang (2001) imply that

$$\mathrm{P}(T_n \leq x \mid \widehat{S}^{\mathrm{ini}} = S) \to \exp(-\exp(-x)), \quad \text{as } n \to \infty.$$

Therefore, we have

$$\begin{aligned}
\mathrm{P}(T_n \le x) &= \mathrm{P}(T_n \le x \mid \widehat{S}^{\mathrm{ini}} = S)\mathrm{P}(\widehat{S}^{\mathrm{ini}} = S) + \mathrm{P}(T_n \le x \mid \widehat{S}^{\mathrm{ini}} \neq S)\mathrm{P}(\widehat{S}^{\mathrm{ini}} \neq S) \\
&= p_n \mathrm{P}(T_n \le x \mid \widehat{S}^{\mathrm{ini}} = S) + (1 - p_n)\mathrm{P}(T_n \le x \mid \widehat{S}^{\mathrm{ini}} \neq S) \\
&\to \exp(-\exp(-x)), \quad \text{as } n \to \infty,
\end{aligned}$$

which completes the proof.

Proof of Theorem 2. After centering the response $\mathbf{y}_i$ and the columns of $\mathbf{X}_1$, (3.19) reduces to minimizing the following standard group-Lasso problem:

$$\frac{1}{2n}\sum_{i=1}^{d} \| \overline{\mathbf{y}}_i - \overline{\mathbf{X}}_1 \boldsymbol{\beta}_{i,\bullet} \|_2^2 + \lambda \sum_{j=1}^{n-1} \|\boldsymbol{\beta}_{\bullet,j}\|_2.$$

According to conditions (C3′)–(C6′) and Theorem 4.2 of Nardi and Rinaldo (2008), we have $\mathrm{P}(\widehat{S}^{\mathrm{ini}} = S) \to 1$ as $n \to \infty$. Similar to the proof of Theorem 1, we only need to show that

$$\mathrm{P}(T_n \le x \mid \widehat{S}^{\mathrm{ini}} = S) \to \exp(-\exp(-x)), \quad \text{as } n \to \infty. \tag{A.1}$$

Let $\{z_{i,j} : \ i = 1, \ldots, d; \ j = 1, \ldots, n\}$ be the independent random variables with $z_{i,j} \overset{\text{i.i.d.}}{\sim} N(0, \sigma_i^2)$. Following the proof of Theorem 1 of Fan and Huang (2001), under the event $\{\widehat{S}^{\mathrm{ini}} = S\}$, together with conditions (C1) and (C2′), we have

$$k^{-1/2}\sum_{j=n_0}^{k} \widehat{\varepsilon}_{i,j}^{*2} = k^{-1/2}\sum_{j=n_0}^{k} z_{i,j}^2 + o_{\mathrm{P}}\{(\log(\log(n)))^{-3/2}\}, \quad i = 1, \ldots, d. \tag{A.2}$$

Let $Z_j^{(i)} = (z_{i,j}^2 - \sigma_i^2)/(\sqrt{2}\sigma_i^2)$ and

$$\widetilde{T}_n = \max_{1\le k\le n} \Big\{ \sum_{i=1}^{d} \Big(\sum_{j=1}^{k} \frac{z_{i,j}^2 - \sigma_i^2}{\sqrt{2k\sigma_i^4}} \Big)^2 \Big\}^{1/2}.$$

Then, by Lemma 2.2 of Horváth (1993) and the fact that $Z_j^{(i)}$ has a zero mean and unit variance, we have

$$\mathrm{P}\Big(\sqrt{2\log(\log(n))}\widetilde{T}_n - \Big\{ 2\log(\log(n)) + \frac{d}{2}\log(\log(\log(n))) - \log\big(\Gamma\big(\frac{d}{2}\big)\big) \Big\} \le x \Big) \to \exp(-\exp(-x)). \tag{A.3}$$

Hence,

$$\begin{aligned} \widetilde{T}_n &= \sqrt{2\log(\log(n))}\{1 + o_{\mathrm{P}}(1)\}, \\ \widetilde{T}_{\log(n)} &= \sqrt{2\log(\log(\log(n)))}\{1 + o_{\mathrm{P}}(1)\}, \end{aligned}$$

which implies that the maximum of $\widetilde{T}_n$ cannot be achieved at $k < \log(n)$. Thus, using (3.20) and (A.2), we have

$$T_n^* = \widetilde{T}_n + o_{\mathrm{P}}\{(\log(\log(n)))^{-3/2}\}. \tag{A.4}$$

Therefore, (A.1) follows from (A.3) and (A.4), which completes the proof.

S2 Additional simulation

In this section, we evaluate the performance of nonconvex penalty in the partial penalized least squares (PPLS) algorithm. Suppose the true data

generating process is as follows:

$$y_j = f(x_j) + \sum_{k=1}^{K} \beta_k \mathrm{I}(x_j \geq \tau_k) + \varepsilon_j, \quad j = 1, \ldots, n,$$

where the sample size $n = 500$, and $x_j = j$, $j = 1, \ldots, n$. We set the number of change points $K = 4$, with corresponding locations of change points $S = (\tau_1, \tau_2, \tau_3, \tau_4) = (150, 200, 400, 450)$, and the differences in mean levels $(\beta_1, \beta_2, \beta_3, \beta_4) = (0.5, -1, 1, -0.5)$. Assume $\varepsilon_j \overset{\text{i.i.d.}}{\sim} N(0, 0.15)$, and the mechanism for generating the wave pattern $f(\cdot)$ is as follows:

Scenario I Setting $m = 2$ and configuring all Fourier coefficients $\{a_1, a_2, b_1, b_2\}$ in (3.8) to be equal to 0.1.

Scenario II Setting $m = 7$, with the Fourier coefficients $\{a_3, a_4, a_5, b_3, b_4, b_5\}$ in (3.8) equal to 0, while the other Fourier coefficients are set to 0.1.

Scenario III Defining $f(\cdot)$ as in (4.24) with $\theta = 0.1$.

We conduct the simulation 100 times. For each simulated dataset, we employ the following methods to estimate the support set S:

Method 1 PPLS estimator (3.14) with $m = 5$.

Method 2 PPLS estimator (3.14) with $m = 10$.

Method 3 PPLS estimator (3.14) with $m = 5$ and SCAD penalty.

Method 4 PPLS estimator (3.14) with $m = 10$ and SCAD penalty.

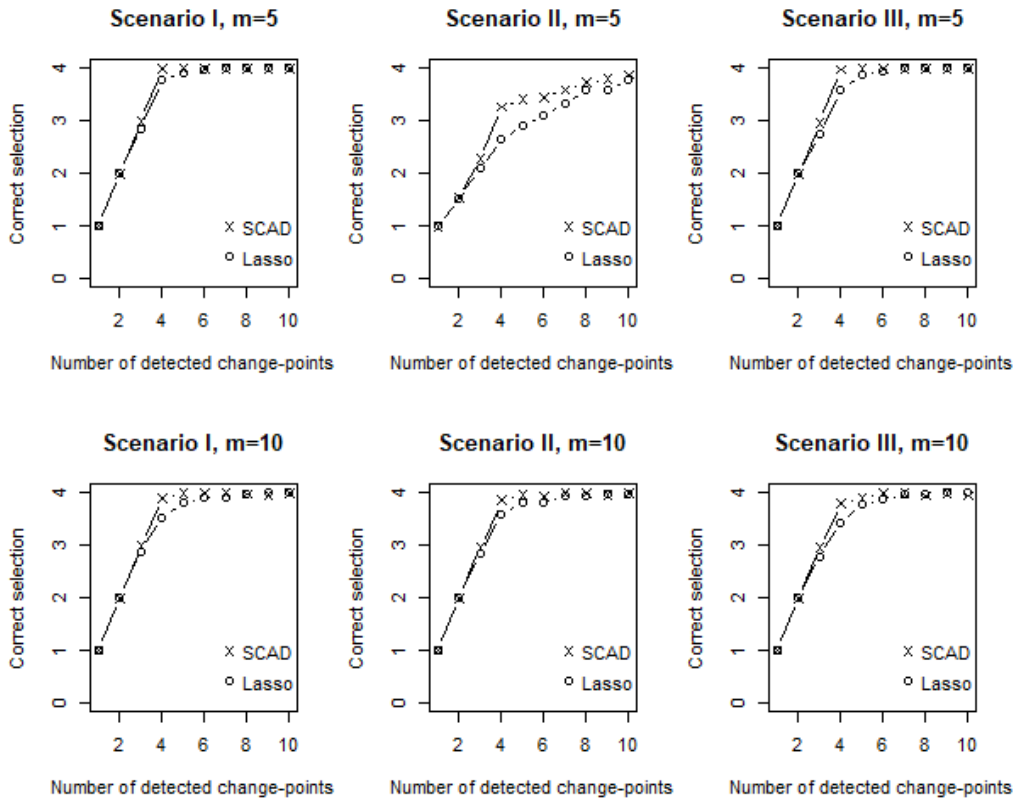

Figure 1: *Average number of true change-points detected by different methods over* 100 *simulations with Lasso* (∘) *or SCAD* (×) *penalty. Top panels: m=5; bottom panels: m=10. Left panels: Scenario I; middle panels: Scenario II; right panels: Scenario III.*

Figure 1 displays the average number of true change-points correctly detected by the estimated support set $\widehat{S}$ using different methods across 100 simulations, as λ decreases (or equivalently, as the estimated number of change-points $\widehat{K} = |\widehat{S}|$ increases from 1 to 10). A true change-point τ_k is considered correctly detected if there exists $\widehat{\tau} \in \widehat{S}$ such that $\widehat{\tau} \in$

$[\tau_k - 2, \tau_k + 2]$. The results demonstrate that the PPLS algorithms with Lasso and SCAD penalties are comparable in most scenarios, except in the top middle panel, where we assume $m = 5$ in the fitting procedure but the true value is $m = 7$. Even for the most complex wave pattern (Scenario III) which deviates from the standard Fourier series expansion, our proposed method with either Lasso or SCAD penalty performs remarkably well. Notably, the Lasso-based algorithm is significantly faster than the SCAD-based counterpart due to the nonconvex nature of the SCAD penalty.

S3 Additional real data analysis

In this section, we show the performance of our proposed algorithms using a publicly available aCGH dataset obtained from the UCSF Cancer Center Array CGH Core Facility (`http://microarrays.curie.fr/publications/oncologie_moleculaire/bladder_TCM/`). The data comprise bladder tumor samples analyzed on CGH microarrays, with more than 2000 bacterial artificial chromosome (BAC) clones covering the human genome, offering an average resolution of 1.3 Mb (HumArray 2.0). Spots located in regions with spatial bias, characterized by abnormally high $\log_2$ ratios measured in specific areas of the array, typically due to edge or corner effect, have been excluded (Stransky et al., 2006). For subsequent transcriptome and CGH

correlations, we consider the $\log_2$ ratios of positions spanning 2171559 to 37334583 kilobases from four samples (X1410, X1533-1, X506 and X2259-1). This selection results in a final list of 2300 probes for each sample. Notably, segments with high or low $\log_2$ ratios correspond to gains or losses of copy numbers.

We utilize the CBS (Olshen *et al.*, 2004), cumSeg (Muggeo and Adelfio, 2011), LB (Huang *et al.*, 2005) algorithms, and our proposed partial penalized least squares method to detect change points in the bladder tumor data. In our proposed method, both the tuning parameter λ and the number m of basis functions in the Fourier expansion (3.8) are chosen by minimizing the following generalized Bayesian information criterion (gBIC):

$$\text{gBIC} = \log(\hat{\sigma}^2) + \text{edf}\frac{\log(n)}{n}C_n,$$

where $\hat{\sigma}^2$ is the residual variance estimator, edf is the actual model dimension quantified by the number of estimated parameters, and $C_n = \log(\log(n))$, as described in Muggeo and Adelfio (2011).

Table 1 displays the estimated number of change points obtained from different methods for each sample. Additionally, the adaptive Neyman test statistics (3.12) were computed for all four samples, and as all of them are large, we rejected the null hypothesis H_0 in (3.9) at a significance level $\alpha = 0.05$. The CBS and LB methods tend to identify more change points

Table 1: *Estimated number of change points from different methods and the adaptive Neyman test statistic* T_n *in* (3.12) *for each sample.*

Sample	CBS	cumSeg	LB	Proposed	T_n
X1410	26	14	18	16	38.52
X1533-1	46	16	41	29	48.71
X506	41	15	28	15	35.97
X2259-1	29	15	27	22	18.32

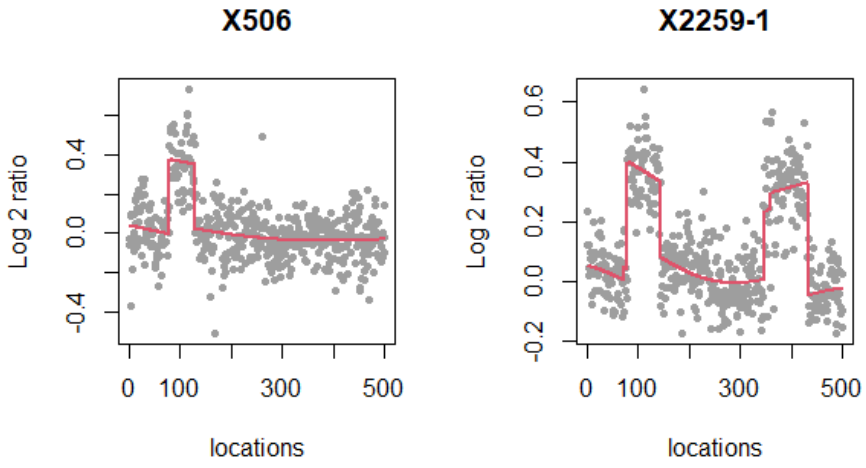

Figure 2: *Analysis of the* $\log_2$ *ratios for sample* $X506$ *and* $X2259$-1*: data (the first* 500 *probes) and the fitted lines using the PPLS estimator* (3.14).

compared to the other methods. However, most of these detected change points are likely to be false positives due to the presence of local trends. Given the lengthy original sequences, it is challenging to visually discern the wave patterns directly from the plots. In Figure 2, we provide a partial view

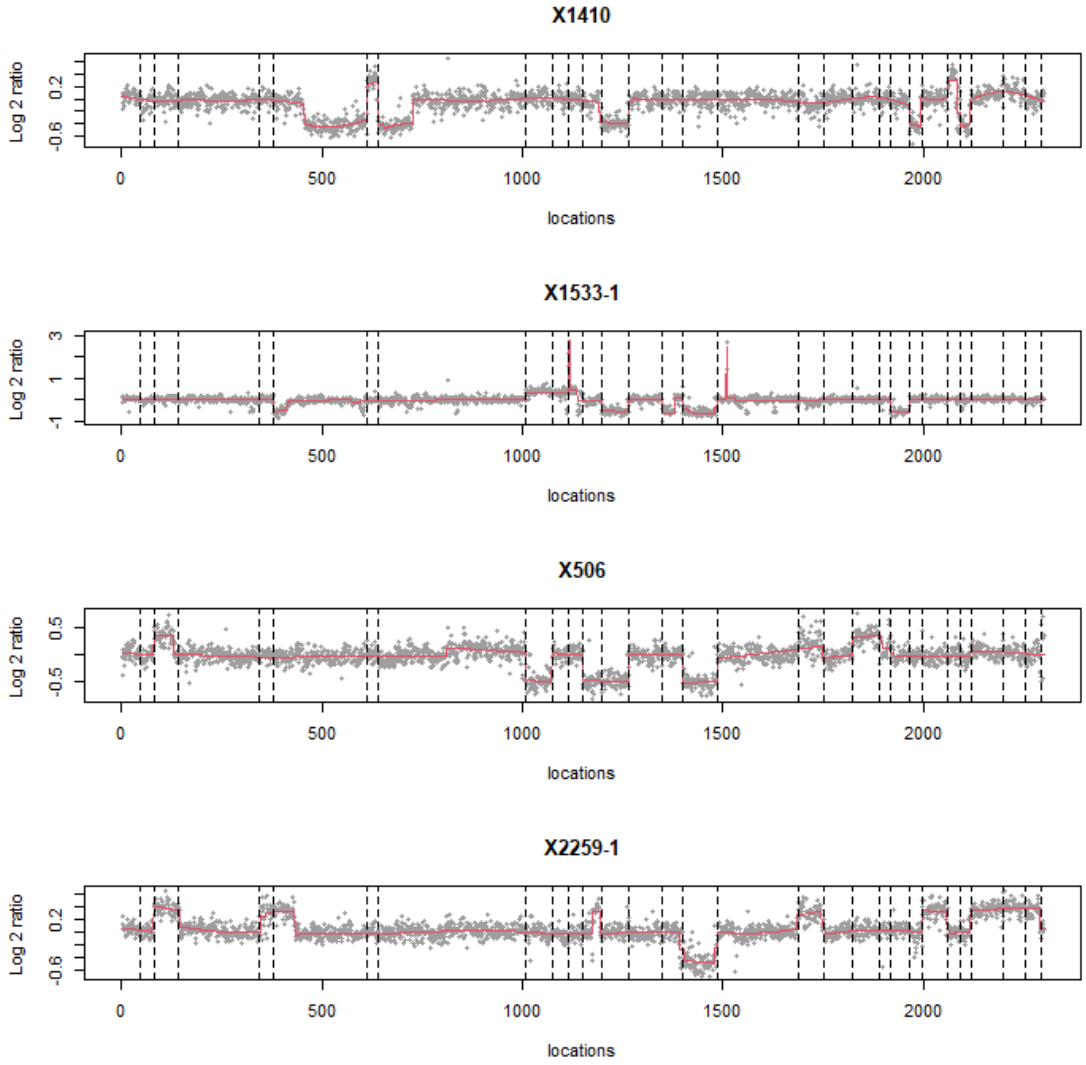

Figure 3: *Analysis of the* $\log_2$ *ratios for the four samples: the (solid) lines fitted by the proposed estimator* (3.14) *for a single sequence and the common change points (vertical dashed lines) according to the proposed estimator* (3.22) *for multiple sequences.*

of the data along with the corresponding lines fitted using PPLS method. When comparing this approach with classic piecewise-constant regression model, we observe that significant change points are effectively preserved, while our proposed method also provides a good fit for the wave patterns.

In Figure 3, we present the estimates of common change points (indicated by vertical dashed lines) obtained via the GPPLS method (3.22). Unlike change point estimators for a single sequence, this method detects some probes with weak signals, such as locations 46 and 2254. This benefit arises from the utilization of multiple samples, which enhances the statistical power for change point detection. Additionally, certain change point estimates, for example, at location 810, which are significant for only one or two sequences and might be related to diseases other than bladder tumors, are not identified as common change points by our proposed method.

References

[1] Horváth, L. (1993). The maximum likelihood method for testing changes in the parameters of normal observations. *Ann. Statist.*, 21, 671–680.

[2] Nardi, Y. and Rinaldo, A. (2008). On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics*, 2, 605–633.

[3] Stransky, N., Vallot, C., Reyal, F., Bernard-Pierrot, I., Diez de Medina, S. G., Segraves,

R., de Rycke, Y., Elvin, P., Cassidy, A., Spraggon, C., Graham, A., Southgate, J., Asselain, B., Allory, Y., Abbou, C. C., Albertson, D. G., Thiery, J. P., Chopin, D. K., Pinkel, D. and Radvanyi. F. (2006). Regional copy number-independent deregulation of transcription in cancer. *Nat. Genet.*, 38, 1386–1396.

[4] Wainwright, M. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using l_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55, 2183–2202.