Bernoulli **24**(1), 2018, 616–660
DOI: 10.3150/16-BEJ890

Change-point estimators with true identification property

CHI TIM NG[1], WOOJOO LEE[2] and YOUNGJO LEE[3]

[1]*Department of Statistics, Chonnam National University, Gwangju 500-757, Korea.*
E-mail: easterlyng@gmail.com
[2]*Department of Statistics, Inha University, Incheon 402-751, Korea. E-mail:* lwj221@gmail.com
[3]*Department of Statistics, Seoul National University, Seoul 151-747, Korea. E-mail:* youngjo@snu.ac.kr

The change-point problem is reformulated as a penalized likelihood estimation problem. A new non-convex penalty function is introduced to allow consistent estimation of the number of change points, and their locations and sizes. Penalized likelihood methods based on LASSO and SCAD penalties may not satisfy such a property. The asymptotic properties for the local solutions are established and numerical studies are conducted to highlight their performance. An application to copy number variation is discussed.

Keywords: change point; consistency; penalized likelihood

1. Introduction

The change-point problems for the sequence of independent random variables have received considerable attention and found applications in various fields including econometrics, genetics, meteorology studies, engineering etc.; see, for example, [3,10,16,20,27,28,33], and [12]. In the change-point problems with the sample size n, the total number of possible configurations is 2^{n-1} as there are $n-1$ potential change points. If n is large, it becomes computationally intractable to apply the best subset approach and investigate all 2^{n-1} possible change-point models. Lai and Xing [21] reviewed the Bayesian approach for the change-point problems. By assuming that the number of change-points is known, [1] and [2] consider a maximum likelihood approach and develop an algorithm based on dynamic programming. For the unknown number of change-point cases, dynamic programming methods are also developed in [5,10,20,28], and [4] etc. However, it should be noted that the "at most" $O(n^2)$ computational complexity of various dynamic programming methods can be achieved only if the objective function can be updated in $O(1)$ time when a new observation is included. This may not be true in general unless the objective function is chosen as the sum of squares or likelihood function for the Gaussian random variables in change-in-mean and change-in-variance cases. In general, all terms in the log-likelihood have to be re-estimated based on the new estimated parameters. Therefore, the updating time may depend on the running sample size since the last change-point. Li and Sieling [26] proposed an $O(n)$ algorithm for the change-point detection based on the idea of FDR-control. However, the computation of the FDR requires Monte-Carlo simulation that is computational intensive.

Recently, [14,16], and [29] consider a penalized likelihood approach of the change-point problems based on the LASSO penalty of [31]. Such penalized likelihood approach requires neither the prior knowledge of the number of change points nor computationally intensive Monte-Carlo

1350-7265

simulation. Moreover, due to the convexity, the negative penalized likelihood function admits a unique local solution. This property allows the researchers to establish the asymptotic theory easily under the assumption of diminishing variance as described in [29]. However, the estimation of the number of change points is not guaranteed to be consistent if the variance is not diminishing.

In the literature of regression analysis, various non-convex penalties have been introduced, for example, the SCAD penalty of [8], the bridge penalty of [9] and [11], and the unbounded penalty of [24]. For the change-point detection problems, if suitable local quadratic approximation strategy (see, e.g., [17]) is used, the Hessian matrix of the penalized likelihood becomes tridiagonal, unlike those in the usual high-dimensional regression problems. Therefore, the resulting change-point detection method can be implemented efficiently in iterative $O(n)$ steps. However, there is a lack of literature on the application of the non-convex penalties to the change-point problems. One difficulty related to the use of the non-convex penalties is the non-uniqueness of the local solutions. If the SCAD penalty is used in the change-point problems, the oracle property that the true model is one of the local solutions can be established in a similar manner as in many literatures of regression, for example, [8]. However, simulation results suggest that the local solution obtained numerically does not always give consistent estimation of the number of change points. Suppose that there is one change point at 500 in Figure 1. It is found that the LASSO and SCAD penalties could select a consecutive change pattern, while the Bridge and unbounded penalties tend to detect a single sharp change. This illustrates that in the change-point problems, the LASSO and SCAD estimators could have difficulty in identifying the true number of change points. In this paper, new penalty functions called modified unbounded (modified bridge) penalty are constructed by combining the non-convex unbounded (bridge) penalty and the convex LASSO penalty. These new penalty functions allow all local solutions within a search space exhibit a trinity of consistencies of (i) the number of change points, (ii) their locations, and (iii) their sizes. Such a trinity is termed the true identification property in this paper.

The paper is organized as follows. In Section 2, the change-point problem is stated and the penalized likelihood method is described. The main results of the local solutions to the penalized likelihood function are presented in Section 3. In addition, the asymptotic properties of the local solutions under different kinds of penalty functions are compared. An algorithm for obtaining the proposed estimator is described in Section 4. A simulation study is conducted in Section 5 to investigate the finite-sample performance of the proposed method. Section 6 contains a real data example, followed by the concluding remarks in Section 7. In Appendix A, we examine the conditions for each of the 2^{n-1} possible configurations to give a local solution to the penalized likelihood function. For such purpose, the idea of restricted local solution is introduced. In Appendix B, the proofs of the theorems in Section 3 are provided, making use of the concepts and propositions developed in Appendix A.

2. Change-point problems

In this section, a change-point model is described and a penalized likelihood method is proposed. The modified unbounded (modified bridge) penalty function is introduced in Section 2.2. To support our choice of the proposed modified unbounded (modified bridge) penalty function, the theoretical properties of the change-point estimators based on the modified unbounded penalty

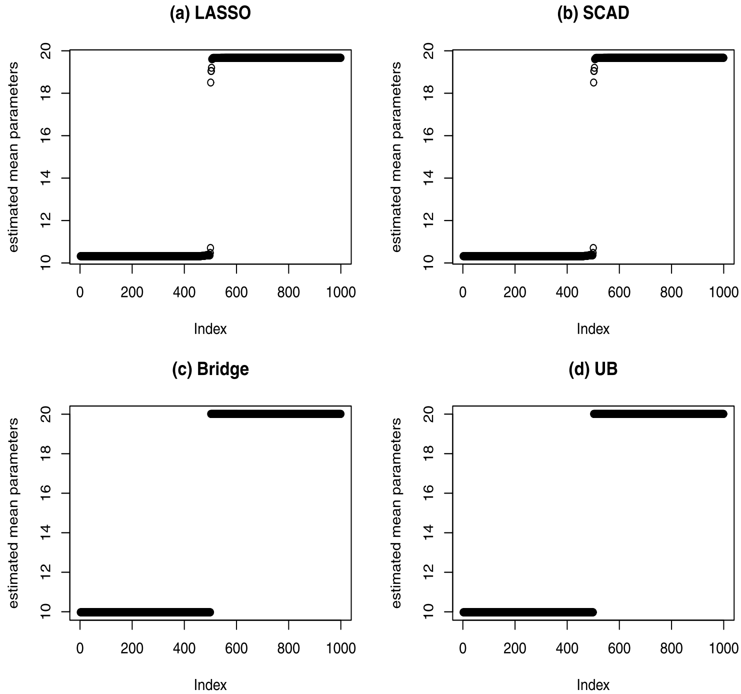

Figure 1. Plots of consecutive changes of LASSO (a) and SCAD (b) and sharp change of bridge (c) and unbounded penalty (d). The true model has one change point at 500.

(modified bridge) function and other penalty functions are further compared in Section 3. In Section 2.3, a search space is introduced. In Section 3, the asymptotic behaviors of the local solutions inside the search space are studied in details.

2.1. Model specification

Let X_i, $i = 1, 2, \ldots, n$ be independent random variables from the density function $f(x; \boldsymbol{\theta}_i, \boldsymbol{\phi})$. Here, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_n^T)^T$, where $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{ip})^T \in \Theta \subset R^p$ is allowed to be time-varying while $\boldsymbol{\phi} \in \Phi \subset R^q$ is assumed to be constant.

The data-generating process is described as follows. Suppose that $X_1, X_2, \ldots, X_n$ are equally-spaced observations collected over the time interval [0, 1]. There are k true change points at $0 = q^{(0)} < q^{(1)} < q^{(2)} < \cdots < q^{(k)} < q^{(k+1)} = 1$. Here, k is finite and fixed, as commonly assumed in the literatures on the change-point problems, for example, in [2]. The true value of $\boldsymbol{\phi}$ is $\boldsymbol{\phi}_0$. For $\ell = 1, 2, \ldots, k+1$, and $[nq^{(\ell-1)}] \le i < [nq^{(\ell)}]$, the true value of parameter $\boldsymbol{\theta}_i$ is $\boldsymbol{\theta}_0^{(\ell)}$. Here, the notation $[x]$ refers to the smallest integer $\ge x$.

The following regularity assumptions are used throughout the paper.

(R1) Both Θ and Φ are compact. The search space (to be described in Section 2.3) is a subset of $\Theta^n \times \Phi$ containing the true parameter vector as an interior point.

(R2) For all $i = 1, 2, \ldots, n$, the following quantity is bounded:

$$\sup_{\boldsymbol{\theta}_i \in \Theta, \boldsymbol{\phi} \in \Phi} \left| E\partial^3 \log f(X_i, \boldsymbol{\theta}_i, \boldsymbol{\phi}) \right|.$$

Here, ∂^3 refers to any third order partial derivative with respect to the components of $(\boldsymbol{\theta}_i, \boldsymbol{\phi})$. In addition, for any $\kappa > 0$ and $0 < \varepsilon < \kappa/2$,

$$\sup_{\boldsymbol{\theta}_i \in \Theta, \boldsymbol{\phi} \in \Phi} \max_{1 \le a < b \le n, b-a \ge n^\kappa} (b-a)^{-1/2} \left| \sum_{i=a}^{b} \left[\partial^m \log f(X_i, \boldsymbol{\theta}_i, \boldsymbol{\phi}) - E\partial^m \log f(X_i, \boldsymbol{\theta}_i, \boldsymbol{\phi}) \right] \right|$$
$$= o_p\left(n^\varepsilon\right).$$

Here, $m = 1, 2, 3$ and ∂^m refers to any mth order partial derivative with respect to the components of $(\boldsymbol{\theta}_i, \boldsymbol{\phi})$.

(R3) For the matrices $E\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \log f(X_i, \boldsymbol{\theta}_i, \boldsymbol{\phi})$, $i = 1, 2, \ldots, n$, $\boldsymbol{\theta}_i \in \Theta$, $\boldsymbol{\phi} \in \Phi$, the smallest eigenvalue is bounded below and the largest eigenvalue is bounded above by some positive constants.

(R4) $EX_i X_i^T$ is finite for all $i = 1, 2, \ldots, n$.

The proposition below gives examples where condition (R2) holds. Condition (R3) requires that the model is identifiable. If X_i is multivariate normal and is generated by a factor model, detecting changes in the factor loading can be problematic unless extra identifiability constraints are imposed. Other conditions are standard and are not discussed here.

Proposition 2.1. *Condition* (R2) *holds for sequence of exponential family random variables* X_i *with density function of the form*

$$f(X_i, \boldsymbol{\theta}_i, \boldsymbol{\phi}) = \exp\left\{G^T(\boldsymbol{\theta}_i, \boldsymbol{\phi}) H(X_i)\right\}$$

if $E[H_a(X)]_i^s < \infty$ *for any* $s > 0$. *Here,* $G(\cdot)$ *and* $H(\cdot)$ *are vector-valued functions and are continuous and differentiable up to order* 3 *over* $\Theta \times \Phi$. *The notation* H_a *refers to a component of* H.

The proposition can be established using Lemma 3.1 of [22] by choosing $h_s = 1$ and s to be sufficiently large. Condition R2 guarantees that if $\kappa > 0$ and $0 < \varepsilon < \kappa/2$ are chosen, the standard arguments based on Taylor expansion are applicable to the function $\sum_{i=a}^{b} \log f(X_i, \boldsymbol{\theta}_i, \boldsymbol{\phi})$ corresponding to any sub-series. The Hessian matrix can always be approximated by its expected value provided that a and b are separated by at least a distance of n^κ. Moreover, the third order terms in the Taylor expansion are ignorable.

2.2. Penalized likelihood estimation

To allow piecewise structure in the estimation of $(\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_n^T)^T$, consider the following negative penalized likelihood function:

$$Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\phi}) = -\sum_{i=1}^{n} \log f(X_i; \boldsymbol{\theta}_i, \boldsymbol{\phi}) + \sum_{i=1}^{n-1} \mathcal{P}_\lambda\big(\|\boldsymbol{\xi}_i\|\big),$$

where $\mathcal{P}_\lambda(\cdot)$ is a penalty function, $\boldsymbol{\xi}_i = \boldsymbol{\theta}_i - \boldsymbol{\theta}_{i+1}$, and $\|\boldsymbol{\xi}_i\|$ is the L_2 norm.

The asymptotic properties of the penalized likelihood estimation are closely related to the choice of the penalty function $\mathcal{P}_\lambda(\cdot)$. Comparing to the regression problem, the role played by the penalty function $\mathcal{P}_\lambda(\cdot)$ is more important than the likelihood as illustrated in the following example. To gain some intuitions, it is interesting to compare $\boldsymbol{\theta}^a$ and $\boldsymbol{\theta}^b$ described below. For $\boldsymbol{\theta}^a$, $\boldsymbol{\theta}_1^a = \boldsymbol{\theta}_2^a = \cdots = \boldsymbol{\theta}_{[n/2]}^a = 1$ and $\boldsymbol{\theta}_{[n/2]+1}^a = \boldsymbol{\theta}_{[n/2]+2}^a = \cdots = \boldsymbol{\theta}_n^a = 2$. For $\boldsymbol{\theta}^b$, $\boldsymbol{\theta}_1^b = \boldsymbol{\theta}_2^b = \cdots = \boldsymbol{\theta}_{[n/2]}^b = 1$, $\boldsymbol{\theta}_{[n/2]+1}^b = 1.5$, and $\boldsymbol{\theta}_{[n/2]+2}^b = \boldsymbol{\theta}_{[n/2]+3}^b = \cdots = \boldsymbol{\theta}_n^b = 2$. Though there are two detected change-points in $\boldsymbol{\theta}^b$, they are adjacent to each other. Comparing to $\boldsymbol{\theta}^b$, $\boldsymbol{\theta}^a$ gives sparser representation of the solution. Different penalty functions $\mathcal{P}_\lambda(\cdot)$ exhibit different preferences of $\boldsymbol{\theta}^a$ and $\boldsymbol{\theta}^b$. In the regression problems, if two of the components in the coefficient vector differ by some $O(1)$ quantities, standard arguments can be used to show that the likelihoods differ by an $O_p(n)$ quantity under certain regularity conditions on the model matrix. If n is sufficiently large, the sign of such $O_p(n)$ can further be determined according to the ergodic theorem or the law of large numbers. However, this is not true in the change-point problems. Here, the likelihoods at $(\boldsymbol{\theta}^a, \boldsymbol{\phi})$ and $(\boldsymbol{\theta}^b, \boldsymbol{\phi})$ differ only by an $O_p(1)$ quantity

$$-\log f(X_{[n/2]+1}; 2, \boldsymbol{\phi}) + \log f(X_{[n/2]+1}; 1.5, \boldsymbol{\phi}).$$

Moreover, the sign of such $O_p(1)$ quantity cannot be predicted using the law of large numbers even though n is sufficiently large. Therefore, the difference in the penalty functions, $\mathcal{P}_\lambda(1) - 2\mathcal{P}_\lambda(0.5)$ actually plays a more important role in distinguishing $\boldsymbol{\theta}^a$ from $\boldsymbol{\theta}^b$. Four penalty functions that are commonly used in regression analysis are discussed in what follows. The abilities of such penalty functions to distinguish $\boldsymbol{\theta}^a$ from $\boldsymbol{\theta}^b$ are studied in particular. Below, we see that these four penalty functions have both pros and cons. To overcome the difficulties in the change-point problem, a new penalty function that we call modified unbounded penalty is introduced.

LASSO: $\mathcal{P}_\lambda(z) = \lambda|z|$. It has been introduced in [31] for variable selection in the regression analysis. Note that the penalty terms corresponding to $\boldsymbol{\theta}^a$ and $\boldsymbol{\theta}^b$ are the same. Therefore, $\boldsymbol{\theta}^a$ cannot be distinguished from $\boldsymbol{\theta}^b$ based on the difference $\mathcal{P}_\lambda(1) - 2\mathcal{P}_\lambda(0.5)$. Unlike the regression problem where the sign of the difference in the likelihoods can be predicted, it is far more difficult to distinguish $\boldsymbol{\theta}^a$ from $\boldsymbol{\theta}^b$ in the change-point problem if the LASSO penalty is used. This explains the consecutive change patterns observed in the simulation studies. The LASSO penalty does not prefer a sparse representation of the solution.

SCAD: It is defined as

$$\mathcal{P}_\lambda(z) = \begin{cases} \lambda|z|, & |z| \le n^{-1}\lambda, \\ -\big(nz^2 - 2a\lambda|z| + n^{-1}\lambda^2\big)/\big[2(a-1)\big], & n^{-1}\lambda < |z| \le an^{-1}\lambda, \\ (a+1)n^{-1}\lambda^2/2, & |z| > an^{-1}\lambda \end{cases}$$

for some $a > 2$. Usually, $a = 3.7$ is chosen according to the suggestion in [8]. As in the case of LASSO, the penalty terms corresponding to $\boldsymbol{\theta}^a$ and $\boldsymbol{\theta}^b$ are the same if n is sufficiently large by the definition of the SCAD penalty. Therefore, like the LASSO penalty, the SCAD penalty does not show preference of the sparse representation of the solution.

Bridge: $\mathcal{P}_\lambda(z) = \lambda|z|^\gamma$ $(0 < \gamma < 1)$. It was introduced in [9].

Unbounded penalty: $\mathcal{P}_\lambda(z)$ is defined as

$$\lambda\left\{\log\Gamma(1/\tau) + \frac{\log\tau}{\tau} + \frac{z^2}{2\nu g(z^2;\tau,\nu)} + \frac{(\tau-2)\log g(z^2;\tau,\nu)}{2\tau} + \frac{g(z^2;\tau,\nu)}{\tau}\right\} \qquad (\tau > 2, \nu > 0),$$

where

$$g\big(z^2;\tau,\nu\big) = \frac{1}{4}\left\{2 - \tau + \sqrt{(2-\tau)^2 + \frac{8\tau z^2}{\nu}}\right\}.$$

This penalty function is derived from a random effect model by [24]. Unlike the LASSO and SCAD penalties, the bridge and unbounded penalties prefers $\boldsymbol{\theta}^a$ over $\boldsymbol{\theta}^b$ if the tuning parameter is large enough so that the difference in the likelihoods is dominated. However, both bridge and unbounded penalties are non-differentiable at the origin, giving challenges in both theory and computation.

Modified unbounded (modified bridge) penalty: To achieve the true identification property in the change-point problems, we introduce a new class of penalty functions of the form

$$\mathcal{P}_\lambda(z) = \begin{cases} P_\lambda(z), & \text{if } |z| > B, \\ P_\lambda(B) - \lambda^*\big(B - |z|\big), & \text{otherwise,} \end{cases}$$

where $P_\lambda(z)$ is chosen as the unbounded penalty or bridge penalty. For simplicity, the notation $\mathcal{P}_\lambda(z)$ is used instead of $\mathcal{P}_{\lambda,\lambda^*}(z)$ though λ^* is involved. Figure 2 shows a schematic diagram of the modified unbounded penalty. For $z \le B$, the LASSO penalty is used while for $z > B$, the unbounded penalty is used. Here, λ and λ^* are two tuning parameters and B is chosen such that $n^{-1}P'_\lambda(B) \to \infty$. In [35], the elastic net is constructed by combining the quadratic penalty and the LASSO penalty. In this paper, the similar idea is applied and the unbounded penalty is replaced by the LASSO penalty near the origin. It should be noted that the SCAD penalty also has LASSO portion near the origin.

In this paper, the following definition of local minimum is adopted. This definition is also applicable for the bridge and unbounded penalties, where $\mathcal{P}'_\lambda(0_+) = \infty$.

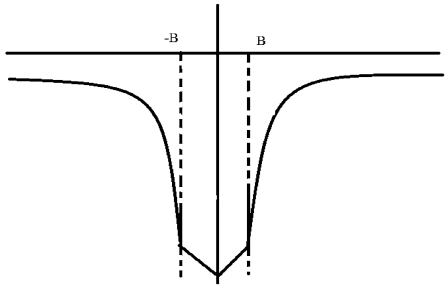

Figure 2. Plot of the modified unbounded penalty function.

Definition 2.1. *Let ∇_i be the gradient operator with respect to $\boldsymbol{\theta}_i$. Then, a point $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}})$ is said to be a local minimum if there exists a neighborhood $\mathcal{N}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}})$ such that for all*

$$(\boldsymbol{\theta}, \boldsymbol{\phi}) \in \mathcal{N}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}) - \left\{(\boldsymbol{\theta}, \boldsymbol{\phi}) \in \Theta^n \times \Phi : \boldsymbol{\theta}_i = 0 \textit{ for some } i = 1, 2, \ldots, n\right\},$$

we have

$$(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}})^T \nabla_{\boldsymbol{\phi}} Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\phi}) + \sum_{i=1}^{n} (\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i)^T \nabla_i Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\phi}) < 0. \tag{2.1}$$

2.3. Search space

To establish the estimation theory of the non-convex penalties, the main difficulty is the non-uniqueness of the local solution. Moreover, the possibilities that there exist both consistent and inconsistent local solutions cannot be ruled out easily. To overcome such difficulties, a search space is proposed. In Section 3, the consistency theory of the local solutions inside the search space is established. The properties of the local solutions outside the search space are also discussed in Section 3.

Before describing the search space, some notations are introduced. For $(\boldsymbol{\theta}, \boldsymbol{\phi}) \in \Theta^n \times \Phi$, let $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(k'+1)}$, $\ell = 1, 2, \ldots, k' + 1$ be the distinct values of $\boldsymbol{\theta}$. The detected number of change points is defined as k'. For $\ell = 1, 2, \ldots, k' + 1$, define $\Delta^{(\ell)}$, the length of the detected regime as $n^{(\ell)}/n$ where $n^{(\ell)}$ is the number of $\boldsymbol{\theta}_i$ that share the same value of $\boldsymbol{\theta}^{(\ell)}$. For $\ell = 1, 2, \ldots, k'$, set $\boldsymbol{\xi}^{(\ell)} = \boldsymbol{\theta}^{(\ell)} - \boldsymbol{\theta}^{(\ell+1)}$. A change point in a local solution refers to the first observation in a regime.

Consider the search space

$$\aleph = \left\{(\boldsymbol{\theta}, \boldsymbol{\phi}) \in \Theta^n \times \Phi : \min_{\ell \in \{1,2,\ldots,k'+1\}} \Delta^{(\ell)} > n^{\kappa - 1} \text{ and } \min_{\ell \in \{1,2,\ldots,k'\}} \left\| \boldsymbol{\xi}^{(\ell)} \right\| > n^{-\delta}\right\}.$$

Here, $\kappa > 0$ and $\delta > 0$ are chosen fulfilling the following conditions.

Conditions on SCAD penalty

(SCAD1) $\lambda = n^{\alpha}$ for some $1/2 < \alpha < 1$.

(SCAD2) $\alpha < \kappa < 1$.
(SCAD3) $0 < \delta < \kappa - \alpha$.

Conditions on bridge penalty

(BR1) $\lambda = n^{\alpha}$ for some $\gamma/2 < \alpha < 1$.
(BR2) $\alpha < \kappa < 1$.
(BR3) $0 < \delta < (\kappa - \alpha)/(2 - \gamma)$.

Conditions on unbounded penalty

(UB1) $\lambda = n^{\alpha}$ for some $0 < \alpha < 1$.
(UB2) $\alpha < \kappa < 1$.
(UB3) $0 < \delta < (\kappa - \alpha)/2$.

Conditions on modified unbounded penalty

(MUB1) $\lambda = n^{\alpha}$ for some $0 < \alpha < 1$ and $\lambda^* = n^{\beta}$ for some $\max\{1/2, \alpha\} < \beta < 1$ such that $2\beta - \alpha < 1$.
(MUB2) $2\beta - \alpha < \kappa < 1$.
(MUB3) $0 < \delta < (\kappa - \alpha)/2$.

Conditions on modified bridge penalty

(MBR1) $\lambda = n^{\alpha}$ for some $\gamma/2 < \alpha < 1$ and $\lambda^* = n^{\beta}$ for some $\max\{1/2, \alpha\} < \beta < 1$ such that $(2 - \gamma)\beta - \alpha < 1 - \gamma$.
(MBR2) $(2 - \gamma)\beta - \alpha < \kappa(1 - \gamma) < 1 - \gamma$.
(MBR3) $0 < \delta < (\kappa - \alpha)/(2 - \gamma)$.
(MBR4) If $\dim(\Theta) > 1$, further assume that

$$-\frac{\alpha}{\gamma} + \frac{1 - \kappa}{2} < -\frac{1 - \alpha}{2 - \gamma}.$$

Remark on condition MBR4: Suppose that $\alpha = 1/2$ is chosen. When κ is chosen close to one, the bound for γ is close to one too. When κ is chosen close to $1/2$, that is, the minimum value allowed, the bound for γ becomes $\gamma < 3 - \sqrt{5} \approx 0.7639$. Then, one can see that $\alpha = 1/2$ and $\gamma < 0.7639$ guarantee condition MBR4.

The permissible ranges of the parameters $\kappa > 0$ and $\delta > 0$ are closely related to the chance that the local solution obtained numerically is consistent. It is natural to prefer a penalty that allows a wider space $\aleph$. This will further be explained in Section 3. The numerical issues are discussed in Section 4.

Through the search space $\aleph$, we impose restrictions on the minimum distance between two consecutive detected change points. In doing so, the number of detected change points is not allowed to increase too fast as the sample size grows. This rules out inconsistent local solutions. On the other hand, k' is allowed to increase as the sample size grows so as to ensure that consistent solutions with finitely many detected change points are included in $\aleph$.

Though the local solution to the non-convex penalty function is not necessarily unique, we show that this search space rules out most of inconsistent local solutions while keeping consistent solutions. The properties of an accepted local solution are established in Theorems 3.1

and 3.2. Discussion on a certain class of local solutions outside the search space is also provided in Section 3.3.

It is an interesting future research direction to develop rigorous methods for the constrained optimization within the search space and study the influences of δ and κ on the probability that the optimal value is attained on the boundary.

3. Main results

In this section, we investigate the asymptotic behavior of the penalized likelihood estimation. In Section 3.1, the results of the local solutions lying inside $\aleph$ are presented. In Section 3.2, the estimation bias due to the penalty terms is discussed. Going beyond $\aleph$, we show that if the SCAD penalty is used, there exists a class of local solutions exhibiting the consecutive change patterns. That means there is a sequence of consecutive change points in the estimation. On the contrary, the bridge, unbounded, modified bridge, and modified unbounded penalties discourage the consecutive change pattern and therefore allow higher degree of sparsity in the local solution. The consecutive change patterns in the local solution are discussed in Section 3.3.

3.1. True identification property of the non-convex penalties

The main theorems on the true identification property are presented under various kinds of non-convex penalties.

Theorem 3.1 shows that for SCAD, bridge, and unbounded penalties, the number of change points in the local solutions within the search space $\aleph$ is bounded by $2k$. In Theorem 3.2, the upper limit $2k$ can further be reduced to k if the modification to the unbounded penalty function or bridge penalty function is considered instead. The true identification property can then be established.

Theorem 3.1. *For the SCAD* (*bridge*, *unbounded*) *penalty and the search space* $\aleph$ *satisfying SCAD*1–*SCAD*3 (*BR*1–*BR*3, *UB*1–*UB*3), *properties* (i) *and* (ii) *in the following hold*:

(i) *The oracle property is satisfied*,

$$P_k = P\Bigg(\textit{There exists a local solution to } Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\phi}) \textit{ with}$$

$$k' = k \textit{ change points at } \frac{[nq^{(1)}]}{n}, \dots, \frac{[nq^{(k)}]}{n} \Bigg| k\Bigg) \to 1.$$

This guarantees that the search space $\aleph$ *is non-empty.*

(ii) *A bound of* k', *the number of the detected change points is given as*

$$P\big(\textit{All local solutions of } Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\phi}) \textit{ in } \aleph \textit{ have } k' \le 2k | k\big) \to 1.$$

For the LASSO penalty, property (iii) *holds*:

(iii) *Take* $\lambda = O(n^{\alpha})$ *for some* $0 < \alpha < 1$. *If* $\alpha = 1/2$, $P_0 \to C$ *for some* $C \in (0, 1)$. *If* $1/2 < \alpha < 1$, $P_0 \to 1$. *If* $1/2 \le \alpha < 1$ *and* $k > 0$, $P_k \to 0$. *That means the oracle property does not hold for the LASSO penalty unless* $k = 0$.

Though all of the SCAD, bridge, and unbounded penalties exhibit oracle property, the proposition below suggests that the unbounded penalty allows a wider search space $\aleph$ than other two penalties.

Proposition 3.1. (i) *For the SCAD penalty, SCAD*1*–SCAD*3 *are all satisfied if* $0 < \delta < 1/2$, $1/2 + \delta < \kappa < 1$, *and* $1/2 < \alpha < \kappa - \delta$ *are chosen.* (ii) *For the bridge penalty, BR*1*–BR*3 *are all satisfied if* $0 < \delta < 1/2$, $\gamma/2 + (2 - \gamma)\delta < \kappa < 1$, *and* $\gamma/2 < \alpha < \kappa - (2 - \gamma)\delta$ *are chosen.* (iii) *For the unbounded penalty, UB*1*–UB*3 *are all satisfied if* $0 < \delta < 1/2$, $2\delta < \kappa < 1$, *and* $0 < \alpha < \kappa - 2\delta$ *are chosen.*

Proposition 3.1 suggests that for a given $0 < \delta < 1/2$, the lower bounds for the choice of κ under SCAD, bridge, and unbounded penalties are $1/2+\delta$, $\gamma/2+(2-\gamma)\delta$, and 2δ, respectively. Such bound is the smallest under the unbounded penalty. That means, the restriction on the search space is the weakest under the unbounded penalty. For example, if the unbounded penalty is used, $\delta = 1/4$, $\kappa = 3/4$, and $\alpha = 1/8$ can be chosen to give $\aleph = \{(\boldsymbol{\theta}, \boldsymbol{\phi}) \in \Theta^n \times \Phi : \min \Delta^{(\ell)} > n^{-1/4}$ and $\min\|\boldsymbol{\xi}^{(\ell)}\| > n^{-1/4}, \ell = 1, 2, \ldots, k'\}$. If the SCAD penalty is used, $\delta = 1/4$, $\kappa = 7/8$, and $\alpha = 9/16$ can be selected to give $\aleph = \{(\boldsymbol{\theta}, \boldsymbol{\phi}) \in \Theta^n \times \Phi : \min \Delta^{(\ell)} > n^{-1/8}$ and $\min \|\boldsymbol{\xi}^{(\ell)}\| > n^{-1/4}\}$. From $\aleph$, we see that any fixed small minimum length of the detected regime and any fixed small minimum change in the true model can be identified if the sample size n is sufficiently large.

Theorem 3.1 implies that if the tuning parameters are chosen appropriately, the SCAD, bridge, and unbounded estimators satisfy

$$P\big(\text{All local solutions of } Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\phi}) \text{ in } \aleph \text{ have } k' = 0 | k = 0\big) \to 1.$$

For these three penalties, when $k = 0$, $k' = 0$ is the unique solution within $\aleph$ as sample size tends to infinity. The numerical study in Section 5 shows that when $k = 0$ the unbounded penalty seems to achieve the true identification property in the fastest rate as the sample size grows. Note that for $k > 0$, the consistency is not guaranteed by Theorem 3.1. However, if the modified bridge and the modified unbounded penalties are used instead, Theorem 3.2 below further guarantees the true identification property that all local solutions within a search space exhibit a trinity of consistent estimations of (i) the number of change points, (ii) their locations, and (iii) their sizes.

Theorem 3.2. *For the modified unbounded* (*modified bridge*) *penalty and the search space* $\aleph$ *satisfying conditions MUB*1*–MUB*3 (*MBR*1*–MBR*4), *the estimator with the modified unbounded* (*modified bridge*) *penalty satisfies the true identification property*:

(i) *The oracle property is satisfied,*

$$P_k = P\bigg(\text{There exists a local solution to } Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\phi}) \text{ with } k' = k \text{ at } \frac{[nq^{(1)}]}{n}, \ldots, \frac{[nq^{(k)}]}{n} \bigg| k > 0\bigg) \to 1.$$

This guarantees that the search space $\aleph$ *is non-empty.*

(ii) *The estimation of the number of change points is consistent*,

$$P\big(\text{All local solutions of } Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\phi}) \text{ in } \aleph \text{ have } k' = k | k\big) \to 1.$$

(iii) *The estimations of the locations and the sizes of change points are consistent. Let* β *be a constant in condition MUB*1 *for modified unbounded penalty* (*MBR*1 *for modified bridge penalty*) *and* $\varepsilon > 0$ *be arbitrarily small. For* $\ell = 1, 2, \ldots, k$, *let* $\hat{q}^{(\ell)}$ *be the location of the* ℓ*th change point in the local solution*, *that is*, *the first time point with a new value of* $\hat{\boldsymbol{\theta}}$. *Then*

$$\big|\hat{q}^{(\ell)} - q^{(\ell)}\big| \le O_p\big(n^{\beta+\varepsilon-1}\big) \quad \text{and} \quad \big\|\hat{\boldsymbol{\theta}}^{(\ell)} - \boldsymbol{\theta}_0^{(\ell)}\big\| \le O_p\big(n^{\beta+\varepsilon-1}\big)$$

with probability going to one.

3.2. Estimation bias due to the non-convex penalty terms

In this subsection, we study the bias in the estimation due to the penalty term. To avoid the difficulty related to the non-uniqueness of the local solution, we only compare the oracle penalized likelihood estimator and the oracle maximum likelihood estimator defined below.

Consider the oracle penalized likelihood solution $(\hat{\boldsymbol{\theta}}^{(1)}, \ldots, \hat{\boldsymbol{\theta}}^{(k+1)}, \hat{\boldsymbol{\phi}})$ as if all the locations of change points are known in advance, which minimizes

$$-\sum_{\ell=1}^{k+1} \sum_{i=[nq^{(\ell-1)}]}^{[nq^{(\ell)}]-1} \log f\big(X_i; \boldsymbol{\theta}^{(\ell)}, \boldsymbol{\phi}\big) + \sum_{\ell=1}^{k} \mathcal{P}_\lambda\big(\big\|\boldsymbol{\theta}^{(\ell)} - \boldsymbol{\theta}^{(\ell+1)}\big\|\big).$$

Note that the existence of such oracle penalized likelihood solution is guaranteed by (i) of Theorems 3.1 and 3.2.

Consider the oracle maximum likelihood estimator $(\hat{\boldsymbol{\theta}}^{(1)}_{\lambda=0}, \ldots, \hat{\boldsymbol{\theta}}^{(k+1)}_{\lambda=0}, \hat{\boldsymbol{\phi}}_{\lambda=0})$, which minimizes

$$-\sum_{\ell=1}^{k+1} \sum_{i=[nq^{(\ell-1)}]}^{[nq^{(\ell)}]-1} \log f\big(X_i; \boldsymbol{\theta}^{(\ell)}, \boldsymbol{\phi}\big).$$

In the following theorem, we study the order of the bias $\hat{\boldsymbol{\theta}}^{(\ell)} - \hat{\boldsymbol{\theta}}^{(\ell)}_{\lambda=0}$ for various penalties.

Theorem 3.3. (i) *For the bridge*, *modified bridge*, *unbounded*, *and modified unbounded penalties*, $\hat{\boldsymbol{\theta}}^{(\ell)} - \hat{\boldsymbol{\theta}}^{(\ell)}_{\lambda=0} = O_p(\lambda/n)$.

(ii) *For the SCAD penalty*, $\hat{\boldsymbol{\theta}}^{(\ell)} - \hat{\boldsymbol{\theta}}^{(\ell)}_{\lambda=0} = 0$ *with probability going to one*.

In the regression problems, [8] showed that if the SCAD penalty is used, at least one of the local solutions is asymptotically equivalent to the oracle maximum likelihood estimator obtained as if the true subset of relevant covariates is known in advance. Theorem 3.3(ii) suggests that this holds in the change-point problems too. For the unbounded and bridge penalties, if $\lambda = O(n^{\alpha})$ is chosen, the bias is $O_p(1/n^{1-\alpha})$. From the assumptions BR1 abd UB1, $\alpha > \gamma/2$ has to be chosen for the bridge penalty and $\alpha > 0$ has to be chosen for the unbounded penalty. Thus, the unbounded penalty can be chosen to allow less asymptotic shrinkage, and therefore less bias than the bridge penalty.

3.3. Beyond the search space $\aleph$

It is possible that there exist local solutions outside the search space $\aleph$. In this section, a class of local solutions with consecutive change pattern is studied. Such kind of local solutions falls outside $\aleph$. We show that for sufficiently large sample size n, the bridge, unbounded, modified bridge, and modified unbounded penalties are all able to rule out such kind of local solutions while LASSO and SCAD penalty are not. Therefore, bridge, unbounded, modified bridge, and modified unbounded penalties have tendencies of encouraging sparsity.

Suppose that the true model has one change point ($k = 1$). Let $K_L = 0, \pm 1, \pm 2, \dots$ be a given constant. Let $k' > 1$. The probability of detecting a consecutive change with k' consecutive change points is given as

$$P_{k'} = P\Bigg(\text{There exists a local solution to } Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\phi}) \\ \text{with } k' > 1 \text{ at } \frac{[nq^{(1)}] + K_L}{n}, \frac{[nq^{(1)}] + K_L + 1}{n}, \dots, \frac{[nq^{(1)}] + K_L + k' - 1}{n} \Bigg| k = 1\Bigg).$$

Theorem 3.4. (i) *For the LASSO penalty, if* $\lambda = O(n^{\alpha})$ *and* $1/2 \le \alpha < 1$, $P_{k'} \to C$ *for some constant* $C < 1$.

(ii) *For the SCAD penalty, if* $n^{-1/2}\lambda \to \infty$ *and* $n^{-1}\lambda \to 0$, $P_{k'} \to 1$.

(iii) *For the unbounded and bridge penalties, if* $\lambda \to \infty$, $P_{k'} \to 0$.

(iv) *For the modified unbounded and modified bridge penalties, if* $\lambda \to \infty$, $n^{-1}\lambda^* \to 0$, *and* $\lambda^*/\max\{n^{1/2}, \lambda\} \to \infty$, *then*, $P_{k'} \to 0$.

This implies that a model with a consecutive change with $k' > 1$ can be a unique LASSO solution when $k = 1$. For the SCAD penalty to allow an oracle solution, it requires $n^{-1/2}\lambda \to \infty$. Thus, it cannot rule out the consecutive change when $k = 1$.

4. Algorithm

The majorization-minimization algorithm in [17] can be used to obtain a local solution to the penalized likelihood function under all of the LASSO, SCAD, bridge, unbounded, modified bridge, and modified unbounded penalties discussed in Sections 2 and 3.

Let $\ell = \sum_{i=1}^{n} \log f(X_i, \boldsymbol{\theta}_i, \boldsymbol{\phi})$ and

$$\mathbf{A}_\lambda = \sum_{i=1}^{n-1} \frac{\mathcal{P}'_\lambda(\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i+1}\| + \varepsilon)}{\sqrt{\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i+1}\|^2 + \varepsilon}} \mathbf{v}_i \mathbf{v}_i^T,$$

where $\mathbf{v}_i = (0, \ldots, 0, 1, -1, 0, \ldots, 0)^T$ with ith and $i+1$th elements 1 and -1, respectively. Following [17], a small ε is used to avoid the singularity problem in the computation. Here, we choose $\varepsilon = 10^{-8}$. Then, the approximated solution is close to the solution of the penalized likelihood $Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\phi})$ and thus, the trinity of consistent estimation can be achieved. However, it should be noted that the local solution obtained from such approximation strategy never include exact zeros. Therefore, a threshold value is needed. In our simulation studies, a change of size less than 10^{-5} is regarded as zero.

The updating formula is given by

$$\hat{\boldsymbol{\theta}}_{\text{new}} = \hat{\boldsymbol{\theta}}_{\text{old}} - \left(-\partial^2\ell/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T + \mathbf{A}_\lambda\right)^{-1}(-\partial\ell/\partial\boldsymbol{\theta} + \mathbf{A}_\lambda\hat{\boldsymbol{\theta}}_{\text{old}}).$$

All quantities on the right-hand side are evaluated at $\hat{\boldsymbol{\theta}}_{\text{old}}$. Note that $\partial^2\ell/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$ is a diagonal matrix and $\mathbf{A}_\lambda$ is a banded matrix with bandwidth 1. Thus, this updating step can be carried out very efficiently with $O(n)$ computational burden by using the sparse matrix technique in [13]. We choose the tuning parameter λ by minimizing the Bayesian information criterion:

$$\text{BIC}(\lambda) = -2\sum_{i=1}^{n} \log f(X_i; \hat{\boldsymbol{\theta}}_i, \hat{\boldsymbol{\phi}}) + e(\lambda)\log(n),$$

where $(\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\phi}})$ is the solution from the Newton–Raphson algorithm and $e(\lambda)$ is the number of effective parameters:

$$e(\lambda) = \text{tr}\left(\left(-\partial^2\ell/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T + \mathbf{A}_\lambda\right)^{-1}\left(-\partial^2\ell/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T\right)\right).$$

Alternatively, we may use the Akaike information criterion or cross-validation. But, since cross-validation is a time-consuming procedure and AIC tends to choose too many change points, we will not consider them further.

The numerical procedure is summarized in Algorithm 1.

Algorithm 1 Pseudo code for the penalized likelihood estimation

- Initial values

while $\max|\hat{\boldsymbol{\theta}}_{\text{new}} - \hat{\boldsymbol{\theta}}_{\text{old}}| > 10^{-7}$ **do**

- Compute A_λ
- Solve

$$\left(-\partial^2\ell/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T + \mathbf{A}_\lambda\right)\hat{\boldsymbol{\delta\theta}} = (-\partial\ell/\partial\boldsymbol{\theta} + \mathbf{A}_\lambda\hat{\boldsymbol{\theta}}_{\text{old}})$$

by implementing forward substitution: $O(n)$ for the banded matrix.

- $\hat{\boldsymbol{\theta}}_{\text{new}} = \hat{\boldsymbol{\theta}}_{\text{old}} - \hat{\boldsymbol{\delta\theta}}$

end while

- Compute BIC

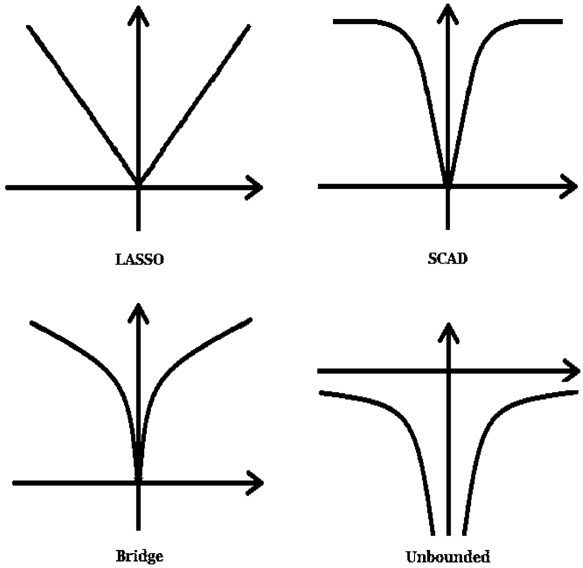

Figure 3. Plot of four penalty functions: LASSO, SCAD, bridge and unbounded penalty.

5. Numerical studies

In this section, the finite-sample behavior of various change-point detection methods are investigated.

We consider the LASSO, SCAD, bridge, unbounded, modified unbounded, and modified bridge penalties. A plot of the first four commonly used penalty functions is given in Figure 3. In the simulation study, the tuning parameters λ in all of the six penalty functions considered are chosen based on the Bayesian information criterion. Other tuning parameters are chosen according to the regularity conditions in Section 2.3. For the bridge and the modified bridge penalty, we choose $\gamma = 1/2$. For both the unbounded penalty and the modified unbounded penalty, $\tau = 30$ and $\nu = 1$ are considered. For the modified bridge and modified unbounded penalty, given the sample size n, $B = 1/n$ and $\lambda^* = n^{0.6}$ are used.

The performance of the proposed methods are also compared with those of three existing R-packages of change-point detection, namely *changepoint* (PELT) by [19], *Segmentor3IsBack* (SEG) by [6], and *stepR* (SMUCE) by [15]. These packages are developed based on the pruned exact linear time algorithm of [20], the pruned dynamic programming segmentation algorithm of [28], and the simultaneous multiscale change-point estimator of [10], respectively. The method of [20] is developed based on the algorithm of [18] by using a pruning step and the computational cost is linear in n. Such a method leads to a substantially more accurate estimation than binary segmentation, one of the most widely used change-point search methods. Frick *et al.* [10] propose to estimate an unknown step function by minimizing the number of change points over the acceptance region of a multiscale test. They provide asymptotic confidence sets for the unknown step function and its change points. In addition, simultaneous multiscale change-point estimator is shown to exhibit optimal detection rate of vanishing signals as $n \to \infty$. Rigaill [28] exploits on a pruned dynamic programming algorithm for detecting multiple change points of an independent random process and demonstrates that its computational complexity is linear in n.

To use PELT and SEG, penalty on the model complexity has to be specified. For PELT package, BIC is chosen. For SEG package, oracle penalty of [23], BIC, and modified BIC of [34] can be considered. Here, the oracle penalty is used. Note that when the same penalty function is used, PELT and SEG produce the same output for the change-point problem because they optimize a common objective function. Since BIC penalty is chosen for PELT, the results of SEG with BIC is not reported here. As we shall see, PELT is slightly better than SEG in various cases, which would imply that the penalty used by PELT is preferred to that of SEG. The differences between PELT and SEG are caused by penalty. The initial value for the penalized likelihood estimators is chosen as the SEG estimator. To measure performance for parameter estimation, mean root mean squared errors of estimators are considered:

$$\sqrt{\sum_{i=1}^{n} \|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\|^2/n}.$$

In the simulation studies, the number of true change points varies from 0 to 4 and the length of each segment between change points varies from 100 to 300. The jump size at change points is 1. The simulation results are presented in Tables 1–4. In the first column of each table, rep(x, y) means that y observations have mean x. In each of the four examples shown below, a replication of 100 times is considered.

Example 5.1. Case $k = 0$: $X_1, \dots, X_n$ ($n = 300, 700$) are generated independently from $N(0, 1)$. Table 1 shows the frequency table of the detected number of change points. We see that the penalized likelihood method with the modified unbounded, the modified bridge and the bridge select the true model with remarkably high frequency when $n = 300$ and perfectly when $n = 700$. The unbounded also identifies the true model perfectly when $n = 700$. Under this setting, the modified unbounded, modified bridge, bridge and SCAD show that the asymptotic property described in Theorems 3.1 and 3.2 holds well. Overall, the modified unbounded, the modified bridge and the bridge methods perform the best.

Example 5.2. Case $k = 1$: $X_1, \dots, X_n$ ($n = 200, 600$) are generated independently from $N(\theta_i, 1)$ with $\theta_i = 0$ for $1 \le i \le n/2$, and $\theta_i = 1$ for $(n/2 + 1) \le i \le n$. Simulation results in Table 2 show that SMUCE, the modified unbounded, the modified bridge and the bridge methods work the best. The influence of tuning parameters γ and τ in the modified bridge and modified unbounded penalties on the performance of change point detection are examined. The segmentation is relatively insensitive to the choice of the parameters. A typical result is reported in Figure 4 where $\gamma = 0.3, 0.4, 0.5, 0.6$ and $\tau = 20, 30, 40, 50$.

Example 5.3. Case $k = 3$: $X_1, \dots, X_n$ ($n = 400, 1200$) are generated independently from $N(\theta_i, 1)$ with $\theta_i = 0$ for $1 \le i \le n/4$, $\theta_i = 1$ for $(n/4 + 1) \le i \le n/2$, $\theta_i = 0$ for $(n/2 + 1) \le i \le 3n/4$ and $\theta_i = 1$ for $(3n/4 + 1) \le i \le n$. The simulation results in Table 3 show that the modified

Table 1. Frequency of the estimated number of change points, and the means of the root mean square error values for estimators when the true model does not have a change point (Example 5.1)

Simul 1	Method	0	1	2	3	≥4	mean RMSE($\hat{\mu}$)
rep(0, 300)	Lasso	91	5	1	0	3	0.003
	SCAD	94	1	3	1	1	0.007
	Bridge	99	1	0	0	0	0.004
	Unbounded	95	5	0	0	0	0.004
	Modified bridge	99	1	0	0	0	0.004
	Modified unbounded	98	2	0	0	0	0.004
	PELT	96	4	0	0	0	0.005
	SEG	94	3	2	0	1	0.007
	SMUCE	93	7	0	0	0	0.006
Simul 2	**Method**	**0**	**1**	**2**	**3**	**≥4**	**mean RMSE($\hat{\mu}$)**
rep(0, 700)	Lasso	94	3	0	0	3	0.002
	SCAD	99	1	0	0	0	0.002
	Bridge	100	0	0	0	0	0.001
	Unbounded	100	0	0	0	0	0.002
	Modified bridge	100	0	0	0	0	0.001
	Modified unbounded	100	0	0	0	0	0.001
	PELT	98	2	0	0	0	0.002
	SEG	97	2	1	0	0	0.002
	SMUCE	98	2	0	0	0	0.001

unbounded is the best when $n = 400$ while the SMUCE is the best when $n = 1200$. Overall, the modified unbounded is the best.

Example 5.4. Case $k = 4$: $X_1, \dots, X_n$ ($n = 500, 1500$) are generated independently from $N(\theta_i, 1)$ with $\theta_i = 0$ for $1 \le i \le n/5$, $\theta_i = 1$ for $(n/5 + 1) \le i \le 2n/5$, $\theta_i = 2$ for $(2n/5 + 1) \le i \le 3n/5$, $\theta_i = 3$ for $(3n/5 + 1) \le i \le 4n/5$, and $\theta_i = 4$ for $(4n/5 + 1) \le i \le n$. The simulation results in Table 4 show that when $n = 500$, the modified unbounded, the modified bridge and the bridge penalties perform better than all other methods. However, when $n = 1500$, SMUCE performs as good as the modified unbounded, the modified bridge and the bridge methods.

Overall, the modified unbounded and the modified bridge methods work the best for detecting change points in mean. In general, LASSO and SCAD methods tend to select too many change points, and have larger RMSE($\hat{\mu}$). SMUCE method works well when n is large. However, it works poorer than SEG method when $k = 4$ with small sample. SEG method tends to select too many change points when k is small. In summary, the modified unbounded and the modified bridge methods outperform the existing methods in the change-point problems.

In addition to the mean change scenarios, two interesting change point scenarios are considered: (1) a change in variance and (2) a change in the shape parameter of a gamma distribution.

Table 2. Frequency of the estimated number of change points, and the means of the root mean square error values for estimators when the true model has a change point (Example 5.2)

Simul3	Method	0	1	2	3	≥4	mean RMSE($\hat{\mu}$)
rep(0, 100), rep(1, 100)	Lasso	0	76	4	2	28	0.037
	SCAD	0	84	13	2	1	0.037
	Bridge	1	98	1	0	0	0.028
	Unbounded	0	95	5	0	0	0.036
	Modified bridge	0	99	1	0	0	0.028
	Modified unbounded	0	99	1	0	0	0.029
	PELT	0	92	8	0	0	0.032
	SEG	0	88	9	2	1	0.035
	SMUCE	0	99	1	0	0	0.027
Simul4	**Method**	**0**	**1**	**2**	**3**	**≥4**	**mean RMSE($\hat{\mu}$)**
rep(0, 300), rep(1, 300)	Lasso	0	70	2	0	28	0.013
	SCAD	0	98	2	0	0	0.009
	Bridge	0	100	0	0	0	0.008
	Unbounded	0	95	5	0	0	0.015
	Modified bridge	0	99	1	0	0	0.009
	Modified unbounded	0	99	1	0	0	0.008
	PELT	0	97	3	0	0	0.009
	SEG	0	96	4	0	0	0.009
	SMUCE	0	99	1	0	0	0.008

In this simulation study, we consider the modified unbounded and the modified bridge methods only because they perform the best in the mean change scenarios and their performances will be compared to those of the three existing change-point detection methods.

Example 5.5. Change in variance: $X_1, \dots, X_n$ ($n = 500$) are generated independently from $N(0, \sigma_i^2)$ with $\sigma_i = 1$ for $1 \le i \le n/2$, and $\sigma_i = 5$ for $(n/2 + 1) \le i \le n$. Simulation results in Table 5 show that the modified unbounded method work the best in terms of RMSE.

Note that the oracle penalty of [23] involves the variance of X_i which is supposed to be constant. The generalization of the oracle penalty to the changing variance cases is not trivial. Therefore, the results of SEG method are omitted here.

Example 5.6. Change in the shape parameter of a gamma distribution: $X_1, \dots, X_n$ ($n = 600$) are generated independently from Gamma$(\alpha_i, 1)$ with $\alpha_i = 1$ for $1 \le i \le n/2$, and $\alpha_i = 5$ for $(n/2 + 1) \le i \le n$. Due to a lack of available packages implementing the dynamic programming methods for gamma random variables, the PELT, SEG, and SMUCE are run as if the data is from change-in-mean Gaussian model. Simulation results in Table 6 show that the modified bridge and modified unbounded methods work the best. For illustration purpose, histograms of the identified change points for four methods are provided in Figure 5. Compared to PELT, SEG and SMUCE,

Table 3. Frequency of the estimated number of change points, and the means of the root mean square error values for estimators when the true model has three change points (Example 5.3)

Simul5	Method	0	1	2	3	4	≥5	mean RMSE($\hat{\mu}$)
rep(0, 100), rep(1, 100)	Lasso	0	0	0	96	1	3	0.112
rep(0, 100), rep(1, 100)	SCAD	0	5	2	88	3	2	0.047
	Bridge	0	0	0	84	14	2	0.034
	Unbounded	0	2	4	93	1	0	0.064
	Modified bridge	0	0	0	96	4	0	0.040
	Modified unbounded	0	0	0	97	3	0	0.034
	PELT	0	0	0	91	9	0	0.034
	SEG	0	0	0	84	14	2	0.035
	SMUCE	0	0	7	93	0	0	0.039

Simul6	Method	0	1	2	3	4	≥5	mean RMSE($\hat{\mu}$)
rep(0, 300), rep(1, 300)	Lasso	0	0	0	92	4	4	0.039
rep(0, 300), rep(1, 300)	SCAD	0	0	0	88	10	2	0.013
	Bridge	0	0	0	95	5	0	0.011
	Unbounded	0	0	0	95	5	0	0.012
	Modified bridge	0	0	0	98	2	0	0.014
	Modified unbounded	0	0	0	99	1	0	0.011
	PELT	0	0	0	98	2	0	0.010
	SEG	0	0	0	95	5	0	0.011
	SMUCE	0	0	0	100	0	0	0.011

the modified unbounded method shows a much sharper peak, which implies its superiority in detecting the true change point.

It should be noted that the "at most" $O(n^2)$ computational complexity of various dynamic programming methods, including PELT, SEG, and SMUCE can be achieved only if the maximum log-likelihood can be updated in $O(1)$ time when a new observation is included. This is true for the Gaussian random variables in both change-in-mean and change-in-variance cases. However, the log-likelihood of gamma random variables cannot be updated in deterministic $O(1)$ time because the corresponding maximum likelihood estimation problem does not have analytic solution. Therefore, the deterministic $O(n^2)$ computational burden is not guaranteed. On the other hand, the proposed algorithm presented in Section 4 is still iterative $O(n)$ complexity.

In the following example, the computational speeds of different methods are compared. To avoid difficulties in comparing algorithms developed from different programming languages, computational time is normalized so that the times of running 100 replications with $n = 1400$ are unity for all methods under comparison.

Example 5.7. Consider the change-point model with $k = 1$ in Example 5.2. In the simulation, sample sizes are chosen as $n = 1400 + 600 * (i - 1)$ for $i = 1, \dots, 5$. Here, in addition to PELT,

Table 4. Frequency of the estimated number of change points, and the means of the root mean square error values for estimators when the true model has four change points (Example 5.4)

Simul7	Method	1	2	3	4	≥5	mean RMSE($\hat{\mu}$)
rep(0, 100), rep(1, 100)	Lasso	0	0	0	41	59	0.050
rep(2, 100), rep(3, 100)	SCAD	0	0	0	97	3	0.034
rep(4,100)	Bridge	0	0	1	98	1	0.034
	Unbounded	0	0	0	95	5	0.062
	Modified bridge	0	0	0	99	1	0.034
	Modified unbounded	0	0	0	99	1	0.034
	PELT	0	0	0	88	12	0.035
	SEG	0	0	0	95	5	0.034
	SMUCE	0	0	24	76	0	0.061
Simul8	**Method**	**1**	**2**	**3**	**4**	**≥5**	**mean RMSE($\hat{\mu}$)**
rep(0, 300), rep(1, 300)	Lasso	0	0	0	17	83	0.018
rep(2, 300), rep(3, 300)	SCAD	0	0	0	98	2	0.010
rep(4, 300)	Bridge	0	0	0	100	0	0.010
	Unbounded	0	0	0	95	5	0.030
	Modified bridge	0	0	0	100	0	0.010
	Modified unbounded	0	0	0	100	0	0.010
	PELT	0	0	0	97	3	0.012
	SEG	0	0	0	97	3	0.011
	SMUCE	0	0	0	100	0	0.010

SEG, SMUCE and MUB, [25] FDR control method (FDRSeg) is considered in our comparative study. The R-package *FDRSeg* by [26] is used and its current version employs a computationally intensive Monte-Carlo step to obtain the null distribution of the multiscale statistic proposed in their paper. In this study, 30 repetitions are used in the Monte-Carlo step. For each of the above five methods and five sample sizes, the simulation is repeated for 100 times. The normalized computational times are obtained by dividing the average run-times by the average run-time for $n = 1400$ so that the normalized times are unit-free. Figure 6 shows a plot of the normalized time against the sample size n. SEG shows the slowest increasing rate and the proposed method is comparable to PELT, SMUCE and FDRSeg.

In the above example, a Gaussian case that favors SEG, PELT and SMUCE is considered. However, it is not clear how to extend these methods to general non-Gaussian models without increasing computational complexity, which is an interesting future research.

6. Application to copy number variations

In this section, we apply the proposed method to the copy number variation for the dataset GM13330 of Corriel human tumor cell lines in [30]. After filtering out missing data, the dataset

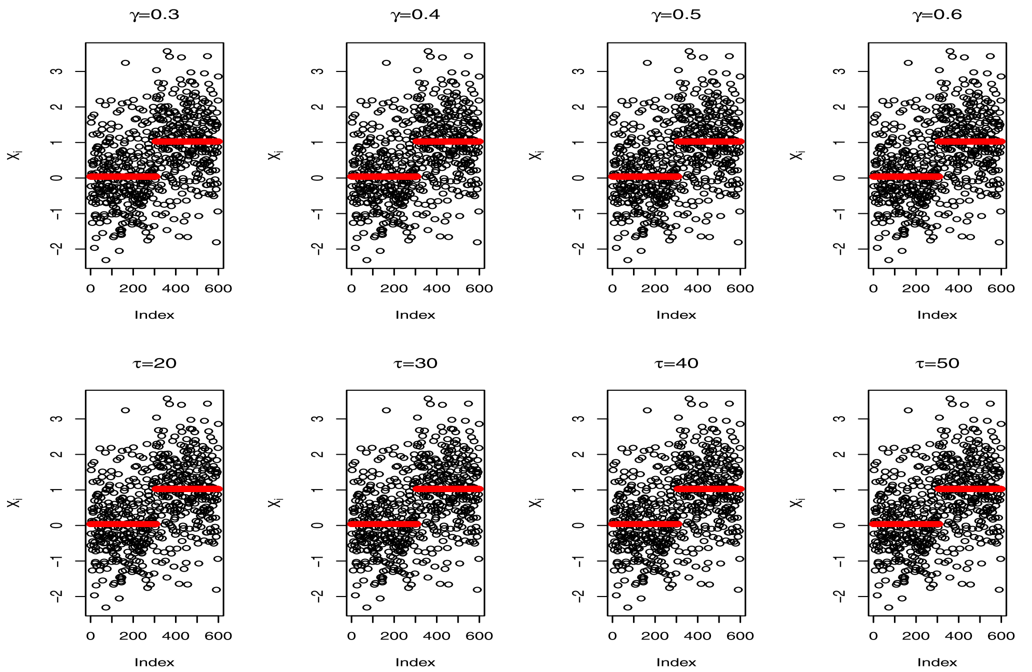

Figure 4. Change point detection (Case $k = 1$) using modified bridge and modified unbounded penalties under different values of γ (upper panel) and τ (lower panel). The solid lines denote the estimated mean parameters θ_i.

consists of 2077 log2-intensity ratio measurements of copy number variation on the autosomal chromosomes 1-22 and X chromosome obtained from microarray-based comparative genomic hybridization of tumor cells and reference cells. Let $X_i = \theta_i + \varepsilon_i$, $i = 1, 2, \ldots, 2077$ be the log2-intensity ratio measurements. Here, θ_i is the mean function and ε_i are independent and identically distributed normal random variables that account for measurement errors. As in [16] and [32], assume that θ_i is a piecewise constant function. The change points of X_i can be interpreted as

Table 5. Frequency of the estimated number of change points for variance, and the means of the root mean square error values for estimators when the true model has a change point in variance (Example 5.5)

Simul9	Method	0	1	2	3	≥4	mean RMSE($\widehat{\sigma^2}$)
rep(0, 250), rep(1, 250)	Modified bridge	0	95	4	1	0	3.162
	Modified unbounded	0	96	3	1	0	2.873
	PELT	0	96	2	2	0	4.278
	SMUCE	0	96	4	0	0	3.271

Table 6. Frequency of the estimated number of change points for the shape parameter of gamma distribution, and the means of the root mean square error values for estimators when the true model has a change point in the shape parameter (Example 5.6)

Simul10	Method	0	1	2	3	≥4	mean RMSE($\hat{\mu}$)
rep(1, 300), rep(5, 300)	Modified bridge	0	98	2	0	0	0.021
	Modified unbounded	0	98	2	0	0	0.019
	PELT	0	0	0	0	100	1.316
	SEG	0	72	4	15	9	0.113
	SMUCE	0	2	1	1	96	0.434

the start and end positions of the mutated regions, which is informative in medical genetics and cancer diagnosis [30].

Figure 7 shows the plots of X_i overlaid with the estimated varying coefficient $\hat{\theta}_i$ by using the (a) LASSO, (b) modified unbounded, (c) modified bridge methods, (d) PELT, (e) SEG and (f) SMUCE, respectively. The modified unbounded penalty method detects only 4 change points at $i = 82, 129, 429$, and 446. The first two and the last two correspond to chromosome 1 (156 276, 240 000) and chromosome 4 (173 943,184 000), respectively, where the numbers in the parentheses are the positions in the chromosomes. The sizes of the change points obtained by the modified unbounded penalty are 0.493, −0.563, −0.779 and 0.827, respectively. The modified

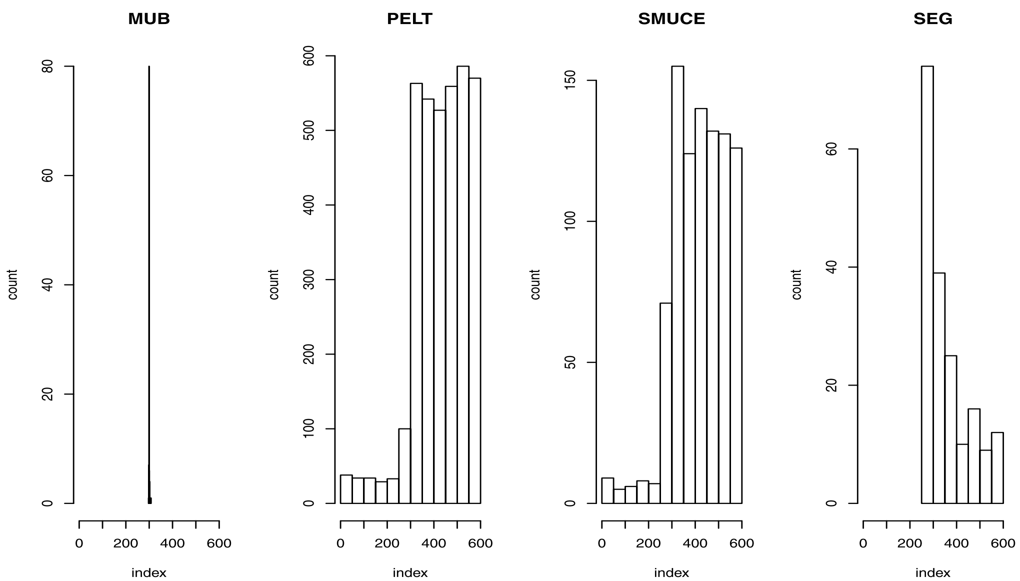

Figure 5. Histograms of the identified change points for the modified unbounded penalty, PELT, SMUCE, and SEG.

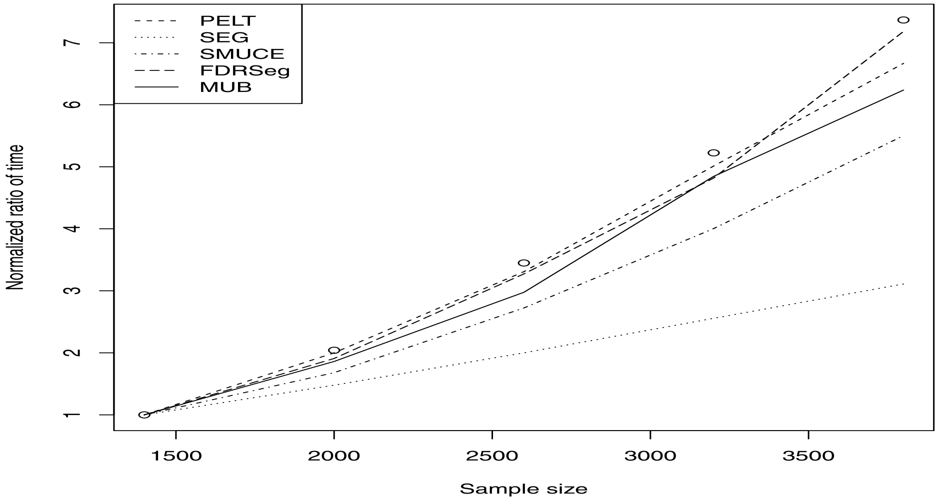

Figure 6. Comparison of (normalized) computational times for five change-point detection methods: PELT, SEG, SMUCE, MUB and FDRSeg.

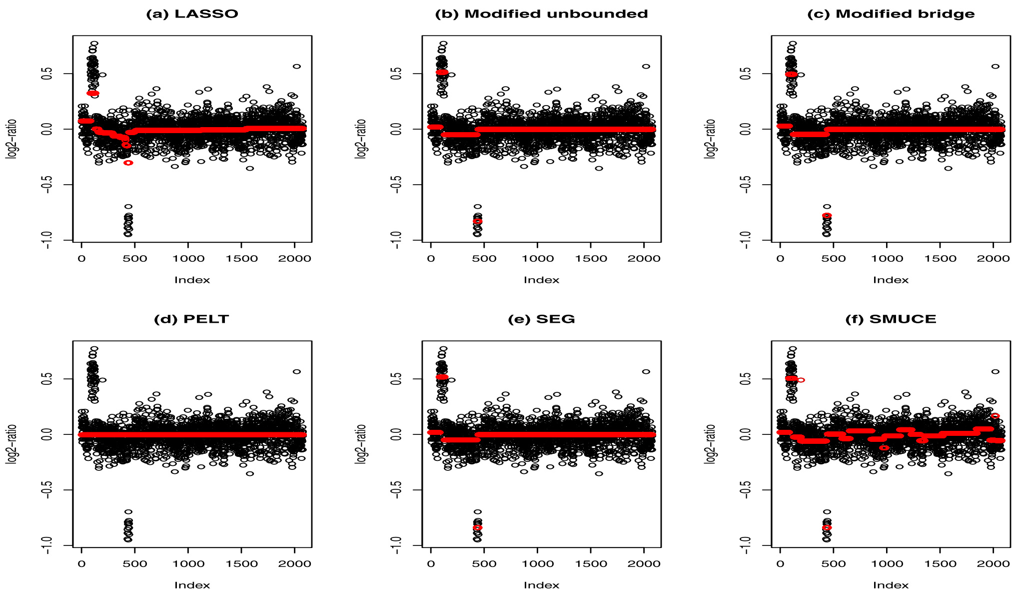

Figure 7. Plots of log2-intensity ratios (black circles) from an array CGH experiment [30] overlaid with detected change points (solid lines) by (a) LASSO, (b) the modified unbounded penalty, (c) the modified bridge penalty, (d) PELT, (e) SEG and (f) SMUCE.

bridge penalty method gives very similar results. It detects 4 change points at the same 4 points. The sizes of the change points by the modified bridge method are 0.482, −0.554, −0.759 and 0.807 in order, so the shrinkages are heavier than those of the modified unbounded penalty. On the other hand, the LASSO method detects more change points, with several consecutive changes, for examples, from $i = 155$ to 160, from $i = 296$ to 301 and from $i = 407$ to 414. The greatest distance between two adjacent change points within a consecutive change is 8. Compared to the modified unbounded and modified bridge methods, the LASSO seems to split interesting regimes into too many pieces. It is also interesting to observe that the LASSO tends to over-smooth the mean function comparing to the modified bridge penalty and the modified unbounded penalty counterparts. For example, the difference in the means of the first two regimes are smaller for the LASSO method. This phenomenon can be explained intuitively below. Compare two time-varying coefficients $\boldsymbol{\theta}^a$ and $\boldsymbol{\theta}^b$ that share the same change points and differ in the components of only one of the regimes. First, the difference in the log likelihoods depends on the length of this regime. If this regime is short, it is possible that the difference in the log likelihoods is dominated by the difference in the penalty terms. For example, if both the size of change point and the length of the regime are $O(1)$, the penalty always dominates the log likelihoods if $\lambda \to \infty$ is chosen. Then, we can see the so-called over-smoothing. For the modified bridge and modified unbounded penalties, $\mathcal{P}'_\lambda(z)$ is much smaller than λ if z is far away from zero. This makes the LASSO penalty different from other penalties.

PELT detects no change point while SMUCE gives too many change points. SEG gives 4 change points at the same locations as those of the modified unbounded penalty method, but the sizes of the change points are 0.500, −0.568, −0.790 and 0.837, which are greater than those of the modified unbounded penalty method.

7. Concluding remarks

The oracle property means that there exists a local solution that is consistent with the true model. However, it cannot guarantee that the local solution obtained numerically is the consistent one. In the change-point problems, we show that by combining the non-convex unbounded (bridge) penalty with the convex LASSO penalty, the true identification property that all local solutions are consistent can be achieved within a search space.

Our current theory is very general so it can be used to detect the changes in any kinds of parameters in sequence of independent random vectors, for example, the change in variance and the change in the regression coefficients. For the regression problem, it would be useful to impose the extra assumptions that the covariates are finite-dimensional, random, independent, and identically distributed. These assumptions are used to rule out the possibility of time-varying nuisance parameters. Our current theory does not allow time-varying nuisance parameters. However, the detailed theory and algorithm should be investigated thoroughly in the future research.

Appendix A: Restricted local solution

To study the theoretical properties of the local minimums of $Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\phi})$, we examine the possibility of each of the 2^{n-1} configurations to give a local solution to the penalized likelihood function.

For a given configuration, we call a local minimum point of $Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\phi})$ under this configuration, if exists a restricted local solution. Below, the precise definitions of restricted local solutions are given. The conditions for a restricted local solution to be a local solution are provided in the following propositions. In Appendix B, we further show that such conditions hold only if the configuration is consistent with the true model.

Definition A.1. Restricted local solution without change point: *Consider the function*

$$Q_\lambda^*(\boldsymbol{\theta}^{(1)}, \boldsymbol{\phi}) = -\sum_{i=1}^{n} \log f(X_i; \boldsymbol{\theta}^{(1)}, \boldsymbol{\phi}).$$

Let $(\hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\phi}})$ *be the local minimum of* $Q_\lambda^*(\boldsymbol{\theta}^{(1)}, \boldsymbol{\phi})$. *The restricted local solution corresponding to the above negative likelihood function is defined as* $\hat{\boldsymbol{\theta}}_1 = \cdots = \hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}^{(1)}$.

Definition A.2. Restricted local solution with k' change points: *Let* $1 = t^{(0)} < t^{(1)} < t^{(2)} < \cdots < t^{(k')} < t^{(k'+1)} = n + 1$ *be given indexes. Consider the function*

$$\begin{aligned} &Q_\lambda^*(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(k'+1)}, \boldsymbol{\phi}) \\ &\quad = -\sum_{\ell=1}^{k'+1} \sum_{i=t^{(\ell-1)}}^{t^{(\ell)}-1} \log f(X_i; \boldsymbol{\theta}^{(\ell)}, \boldsymbol{\phi}) + \sum_{\ell=1}^{k'} \mathcal{P}_\lambda(\|\boldsymbol{\xi}^{(\ell)}\|). \end{aligned}$$

Let $(\hat{\boldsymbol{\theta}}^{(1)}, \ldots, \hat{\boldsymbol{\theta}}^{(k'+1)}, \hat{\boldsymbol{\phi}})$ *be the local solution with* $\hat{\boldsymbol{\theta}}^{(\ell)} \neq \hat{\boldsymbol{\theta}}^{(\ell+1)}$ *for all* $\ell = 1, 2, \ldots, k'$. *The restricted local solution corresponding to the above function is defined as* $\hat{\boldsymbol{\theta}}_{t^{(\ell-1)}} = \cdots = \hat{\boldsymbol{\theta}}_{t^{(\ell)}-1} = \hat{\boldsymbol{\theta}}^{(\ell)}$ *for* $\ell = 1, 2, \ldots, k' + 1$.

Definition A.3. Restricted local solution with consecutive change: *Let* $1 \neq t^{(1)} \leq n$ *be a given index. The restricted local solution with* k' *change points at* $t^{(1)}, t^{(1)} + 1, \ldots, t^{(1)} + k' - 1$ *is called a restricted local solution with consecutive change.*

It should be noted that a local solution to $Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\phi})$ is a restricted local solution while a restricted local solution is not necessarily a local solution to $Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\phi})$. The conditions for the local minimality of restricted local solution are given in the following proposition.

Proposition A.1. *The restricted local solution without change point* (*see Definition* A.1) *is a local minimum of* $Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\phi})$ *if and only if* $\max_i \|\mathbf{S}_i^{(1)}\| < \mathcal{P}_\lambda'(0_+)$, *where* $\mathbf{S}_0^{(1)} = 0$ *and* $\mathbf{S}_i^{(1)} = \mathbf{S}_{i-1}^{(1)} - \nabla_{\boldsymbol{\theta}} \log f(X_i; \hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\phi}})$ *for* $i = 1, 2, \ldots, n - 1$.

Proof. Let $\mathbf{a} = (\mathbf{a}_1^T, \dots, \mathbf{a}_n^T, \mathbf{b}^T)^T$ be a unit vector. To establish the sufficiency, consider the directional derivative of Q_λ (as a function of $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \dots, \boldsymbol{\xi}_{n-1}, \boldsymbol{\theta}^{(1)}, \boldsymbol{\phi}$) along the direction $\mathbf{a}$,

$$
\begin{aligned}
&D_{\mathbf{a}} Q_\lambda\big(\mathbf{0}, \dots, \mathbf{0}, \hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\phi}}\big) \\
&\quad= \lim_{\varepsilon \to 0} \mathbf{b}^T \nabla_{\boldsymbol{\phi}} Q_\lambda\big(\varepsilon \mathbf{a}_1, \dots, \varepsilon \mathbf{a}_{n-1}, \hat{\boldsymbol{\theta}}^{(1)} + \varepsilon \mathbf{a}_n, \hat{\boldsymbol{\phi}} + \varepsilon \mathbf{b}\big) \\
&\qquad + \sum_{i=1}^{n} \mathbf{a}_i^T \nabla_i Q_\lambda\big(\varepsilon \mathbf{a}_1, \dots, \varepsilon \mathbf{a}_{n-1}, \hat{\boldsymbol{\theta}}^{(1)} + \varepsilon \mathbf{a}_n, \hat{\boldsymbol{\phi}} + \varepsilon \mathbf{b}\big) \\
&\quad= \mathbf{b}^T \nabla_{\boldsymbol{\phi}} L\big(\mathbf{0}, \dots, \mathbf{0}, \hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\phi}}\big) + \sum_{i=1}^{n} \mathbf{a}_i^T \nabla_i L\big(\mathbf{0}, \dots, \mathbf{0}, \hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\phi}}\big) + \mathcal{P}_\lambda'(0_+) \sum_{i=1}^{n-1} \|\mathbf{a}_i\| \\
&\quad= \sum_{i=1}^{n-1} \mathbf{a}_i^T \nabla_i L\big(\mathbf{0}, \dots, \mathbf{0}, \hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\phi}}\big) + \mathcal{P}_\lambda'(0_+) \sum_{i=1}^{n-1} \|\mathbf{a}_i\| \\
&\quad\geq -\sum_{i=1}^{n-1} \|\mathbf{a}_i\| \cdot \big\|\nabla_i L\big(\mathbf{0}, \dots, \mathbf{0}, \hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\phi}}\big)\big\| + \mathcal{P}_\lambda'(0_+) \sum_{i=1}^{n-1} \|\mathbf{a}_i\|.
\end{aligned}
$$

Here, we have used the fact that

$$
\nabla_{\boldsymbol{\phi}} L\big(\mathbf{0}, \dots, \mathbf{0}, \hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\phi}}\big) = \nabla_n L\big(\mathbf{0}, \dots, \mathbf{0}, \hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\phi}}\big) = 0.
$$

Therefore,

$$
\big\|\nabla_i L\big(\mathbf{0}, \dots, \mathbf{0}, \hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\phi}}\big)\big\| < \mathcal{P}_\lambda'(0_+) \qquad \text{for all } i = 1, 2, \dots, n-1
$$

is sufficient for the local minimality. Necessity is obvious. □

The general results of $k' > 0$ are stated without proof as follows.

Proposition A.2. *Let* $1 \leq t^{(1)} < t^{(2)} < \cdots < t^{(k')} < t^{(k'+1)} = n + 1$ *be given indexes. Suppose that* $(\hat{\boldsymbol{\theta}}^{(1)}, \dots, \hat{\boldsymbol{\theta}}^{(k'+1)}, \hat{\boldsymbol{\phi}})$ *is a differentiable local solution of the function* $Q_\lambda^*(\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(k'+1)}, \boldsymbol{\phi})$ (*see Definition* A.2). *Then, the corresponding restricted local solution is a local minimum of* $Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\phi})$ *if and only if*

(i) *the function*

$$
Q_\lambda^*\big(\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(k'+1)}, \boldsymbol{\phi}\big) = -\sum_{\ell=1}^{k'+1} \sum_{i=t^{(\ell-1)}}^{t^{(\ell)}-1} \log f\big(X_i; \boldsymbol{\theta}^{(\ell)}, \boldsymbol{\phi}\big) + \sum_{\ell=1}^{k'} \mathcal{P}_\lambda\big(\big\|\boldsymbol{\xi}^{(\ell)}\big\|\big)
$$

admits a local minimum $(\hat{\boldsymbol{\theta}}^{(1)}, \dots, \hat{\boldsymbol{\theta}}^{(k'+1)}, \hat{\boldsymbol{\phi}})$ *fulfilling* $\hat{\boldsymbol{\theta}}^{(\ell)} \neq \hat{\boldsymbol{\theta}}^{(\ell+1)}$ *for* $\ell = 1, 2, \dots, k'$,

(ii) $\max_{i=1,\dots,t^{(\ell)}-t^{(\ell-1)}-1}\|\mathbf{S}_i^{\ell}\| < \mathcal{P}'_\lambda(0_+)$ *for* $\ell = 1, 2, \dots, k'+1$, *where*

$$\mathbf{S}_0^{(1)} = \mathbf{0}, \qquad \mathbf{S}_0^{(2)} = -\mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(1)}\|\big)\cdot \mathbf{j}^{(1)}, \qquad \dots, \qquad \mathbf{S}_0^{(k'+1)} = -\mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(k')}\|\big)\cdot \hat{\mathbf{j}}^{(k')},$$

and

$$\mathbf{S}_i^{(\ell)} = \mathbf{S}_{i-1}^{(\ell)} - \nabla_{\boldsymbol{\theta}} \log f\big(X_{t^{(\ell)}+i-1}; \hat{\boldsymbol{\theta}}^{(\ell)}, \hat{\boldsymbol{\phi}}\big).$$

Here $\hat{\mathbf{j}}^{(\ell)} = \hat{\boldsymbol{\xi}}^{(\ell)}/\|\hat{\boldsymbol{\xi}}^{(\ell)}\|$ *for* $\ell = 1, 2, \dots, k'$.

If (i) *holds*, *the local minimum point of* $Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\phi})$, $(\hat{\boldsymbol{\theta}}^{(1)}, \dots, \hat{\boldsymbol{\theta}}^{(k'+1)}, \hat{\boldsymbol{\phi}})$ *satisfies the following first-order conditions*:

$$\begin{aligned}
&-\sum_{i=1}^{t^{(1)}-1} \nabla_{\boldsymbol{\theta}} \log f\big(X_i; \hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\phi}}\big) = -\mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(1)}\|\big)\cdot \hat{\mathbf{j}}^{(1)},\\
&-\sum_{i=t^{(1)}}^{t^{(2)}-1} \nabla_{\boldsymbol{\theta}_i} \log f\big(X_i; \hat{\boldsymbol{\theta}}^{(2)}, \boldsymbol{\phi}\big) = -\big[\mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(2)}\|\big)\cdot \hat{\mathbf{j}}^{(2)} - \mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(1)}\|\big)\cdot \hat{\mathbf{j}}^{(1)}\big],\\
&\vdots\\
&-\sum_{i=t^{(k'-1)}}^{t^{(k')}-1} \nabla_{\boldsymbol{\theta}} \log f\big(X_i; \hat{\boldsymbol{\theta}}^{(k')}, \hat{\boldsymbol{\phi}}\big) = -\big[\mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(k')}\|\big)\cdot \hat{\mathbf{j}}^{(k')} - \mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(k'-1)}\|\big)\cdot \hat{\mathbf{j}}^{(k'-1)}\big],\\
&-\sum_{i=t^{(k')}}^{n} \nabla_{\boldsymbol{\theta}} \log f\big(X_i; \hat{\boldsymbol{\theta}}^{(k'+1)}\big) = \mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(k')}\|\big)\cdot \hat{\mathbf{j}}^{(k')},\\
&-\sum_{i=1}^{t^{(1)}-1} \nabla_{\boldsymbol{\phi}} \log f\big(X_i; \hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\phi}}\big) - \dots - \sum_{i=t^{(k')}}^{n} \nabla_{\boldsymbol{\phi}} \log f\big(X_i; \hat{\boldsymbol{\theta}}^{(k'+1)}, \hat{\boldsymbol{\phi}}\big) = 0.
\end{aligned}$$

Remark A.1. Proposition A.2 requires that $(\hat{\boldsymbol{\theta}}^{(1)}, \dots, \hat{\boldsymbol{\theta}}^{(k'+1)}, \hat{\boldsymbol{\phi}})$ is a differentiable point of $Q^*_\lambda(\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(k'+1)}, \boldsymbol{\phi})$. For the modified unbounded (modified bridge) penalty, the search space $\aleph$ does not include any solutions with a change point of size B. Therefore, Proposition A.2 is applicable within $\aleph$. Going beyond $\aleph$, extra arguments are needed to rule out the possibility of non-differentiability in the proof of Theorem 3.4.

Appendix B: Proofs of the main theorems

B.1. Conventions

Throughout this paper, the following notations are used. The notations for the true model are as follows. Let k be the true number of change points. The true values of the parameters are $\boldsymbol{\theta}_i^0 = \boldsymbol{\theta}_0^{(\ell)}$ for $i = [nq^{(\ell-1)}], [nq^{(\ell-1)}] + 1, \dots, [nq^{(\ell)}] - 1$, $\ell = 1, 2, \dots, k + 1$. Define $\boldsymbol{\xi}_0^{(\ell)} = \boldsymbol{\theta}_0^{(\ell)} - \boldsymbol{\theta}_0^{(\ell+1)}$ for $\ell = 1, 2, \dots, k$. The notations for the estimated model are as follows. Let $k' + 1$ be the number of non-zero $\hat{\boldsymbol{\xi}}_i$ in vector $\hat{\boldsymbol{\xi}}$. The parameters are $\hat{\boldsymbol{\theta}}_i = \hat{\boldsymbol{\theta}}^{(\ell)}$ for $i = t^{(\ell-1)}, t^{(\ell-1)} + 1, \dots, t^{(\ell)} - 1$, $\ell = 1, 2, \dots, k' + 1$. Let $n^{(\ell)} = t^{(\ell)} - t^{(\ell-1)}$, for $\ell = 1, 2, \dots, k' + 1$. Define $\hat{\boldsymbol{\xi}}^{(\ell)} = \hat{\boldsymbol{\theta}}^{(\ell)} - \hat{\boldsymbol{\theta}}^{(\ell+1)}$ for $\ell = 1, 2, \dots, k'$. To describe the regimes in vector $\hat{\boldsymbol{\xi}}$ that contain at least one true change point, the following conventions are needed.

Convention B.1. *For any two sequences x_n and y_n, that x_n dominates y_n or $x_n \gg y_n$ means $\lim_{n\to\infty} P(|y_n/x_n| > M) \to 0$ for any constants $M > 0$.*

Convention B.2. *Suppose that there are k_ℓ true change points in the estimated regime ℓ, where $\ell = 1, 2, \dots, k' + 1$. Let $n_1^{(\ell)}, n_2^{(\ell)}, \dots, n_{k_\ell+1}^{(\ell)}$ be the lengths of the sub-regimes split by such k_ℓ true change points, $E_1^{(\ell)}, \dots, E_{k^\ell+1}^{(\ell)}$ and $\mathrm{Var}_1^{(\ell)}, \dots, E_{k^\ell+1}^{(\ell)}$ be the expectations and variances taken under the true parameter value of $(\boldsymbol{\theta}, \boldsymbol{\phi})$ in the sub-regimes of estimated regime ℓ. Define $(\bar{\boldsymbol{\theta}}^{(1)}, \bar{\boldsymbol{\theta}}^{(2)}, \dots, \bar{\boldsymbol{\theta}}^{(k'+1)})$ and $\bar{\boldsymbol{\phi}}$ using equations*

$$n_1^{(\ell)} E_1^{(\ell)} \nabla_{\boldsymbol{\theta}} \log f\big(X; \bar{\boldsymbol{\theta}}^{(\ell)}, \bar{\boldsymbol{\phi}}\big) + \cdots + n_{k^\ell+1}^{(\ell)} E_{k^\ell+1}^{(\ell)} \nabla_{\boldsymbol{\theta}} \log f\big(X; \bar{\boldsymbol{\theta}}^{(\ell)}, \bar{\boldsymbol{\phi}}\big) = 0$$

for all $\ell = 1, 2, \dots, k' + 1$ and

$$\sum_{\ell=1}^{k'+1} \big\{ n_1^{(\ell)} E_1^{(\ell)} \nabla_{\boldsymbol{\theta}} \log f\big(X; \bar{\boldsymbol{\theta}}^{(\ell)}, \bar{\boldsymbol{\phi}}\big) + \cdots + n_{k^\ell+1}^{(\ell)} E_{k^\ell+1}^{(\ell)} \nabla_{\boldsymbol{\theta}} \log f\big(X; \bar{\boldsymbol{\theta}}^{(\ell)}, \bar{\boldsymbol{\phi}}\big) \big\} = 0.$$

Define

$$\bar{\mathbf{H}}^{(\ell)} = \frac{n_1^{(\ell)}}{n^{(\ell)}} \mathrm{Var}_1^{(\ell)} \nabla \log f\big(X; \bar{\boldsymbol{\theta}}^{(\ell)}, \bar{\boldsymbol{\phi}}\big) + \cdots + \frac{n_{k_\ell+1}^{(\ell)}}{n^{(\ell)}} \mathrm{Var}_{k_\ell+1}^{(\ell)} \nabla \log f\big(X; \bar{\boldsymbol{\theta}}^{(\ell)}, \bar{\boldsymbol{\phi}}\big).$$

Convention B.3. *Define the following types of regimes in the estimated mean function $\hat{\boldsymbol{\theta}}_i$ according to the number of true change point(s) included:*

1. *N: no true change point,*
2. *R: one change point, the length of the right portion dominates that of the left,*
3. *L: one change point, the length of the left portion dominates that of the right,*
4. *M: two change points, the length of the middle portion dominates both the left and the right.*

Remark B.1. Note that Lemma B.4 guarantees that as $n \to \infty$, with probability goes to one all regimes in all local solutions belong to one of the above-mentioned four types of regimes.

Convention B.4. *For the unbounded penalty, if* $\|\boldsymbol{\xi}\| \to 0$, *the following approximations are used throughout the proof,*

$$\mathcal{P}_\lambda\big(\|\boldsymbol{\xi}\|\big) \approx \lambda\tau^{-1}(\tau-2)\log\|\boldsymbol{\xi}\|,$$
$$\nabla\mathcal{P}_\lambda\big(\|\boldsymbol{\xi}\|\big) \approx \lambda\tau^{-1}(\tau-2)\|\boldsymbol{\xi}\|^{-1}\mathbf{j},$$
$$\nabla^2\mathcal{P}_\lambda\big(\|\boldsymbol{\xi}\|\big) \approx \lambda\tau^{-1}(\tau-2)\|\boldsymbol{\xi}\|^{-2}\big(I-2\mathbf{j}\mathbf{j}^T\big),$$

where $\mathbf{j} = \boldsymbol{\xi}/\|\boldsymbol{\xi}\|$.

B.2. Proofs

Proof of Theorem 3.1. Results (i) and (iii) follow immediately from Lemmas B.5 and B.6. Let a be the number of Type N regimes and b be the number of non-Type N regimes. To prove (ii), it suffices to show that Type N regimes (see Convention B.3) cannot be neighbors of each other. If this is so, $a \le b+1$. Since a non-Type N regime must contains at least one true change point, $b \le k$. Then, $k' = a+b-1 \le 2k$. By contradiction suppose that regimes ℓ and $\ell+1$ are Type N. Let δ and κ be chosen as in conditions SCAD1–SCAD3 for SCAD penalty, UB1–UB3 for unbounded penalty, and BR1–BR3 for bridge penalty. From Lemma B.1,

$$\|\hat{\boldsymbol{\xi}}^{(\ell)}\| = \|\hat{\boldsymbol{\theta}}^{(\ell)} - \hat{\boldsymbol{\theta}}^{(\ell+1)}\| < \|\hat{\boldsymbol{\theta}}^{(\ell)} - \bar{\boldsymbol{\theta}}^{(\ell)}\| + \|\hat{\boldsymbol{\theta}}^{(\ell+1)} - \bar{\boldsymbol{\theta}}^{(\ell+1)}\| \le 2Cn^{-\delta}.$$

Here, we have used the fact that $\bar{\boldsymbol{\theta}}^{(\ell)} = \bar{\boldsymbol{\theta}}^{(\ell+1)}$ as both regimes ℓ and $\ell+1$ are type N regimes. If constant C in Lemma B.1 is chosen to be less than $1/2$, condition SCAD3 for SCAD penalty (UB3 for unbounded penalty, BR3 for bridge penalty) is violated. □

Proof of Theorem 3.2(i). The result follows from Lemmas B.5 and B.6. □

Proof of Theorem 3.2(ii). Consider the four types of regimes described in Convention B.3. Lemma B.4 guarantees with probability goes to one that all regimes fall into these four types. To complete the proof, it suffices to establish (a) adjacent to Type R or N regime must be regime of Type R or M on the right, (b) adjacent to Type L or N regime must be regime of Type L or M on the left, (c) adjacent to Type L or M regime must be regime of Type L or N on the right, (d) adjacent to Type R or M regime must be regime of Type R or N on the left. These are for ensuring that each detected change point has one and only one true change point in its proximity.

Results (c) and (d) are trivial as true change points must be separated by a distance of at least $O_p(n)$. The proof of (b) is symmetric to that of (a). To prove (a), we need to rule out the possibilities of Type L and N regimes on the right using Lemma B.7 for modified unbounded penalty and Lemma B.8 for modified bridge penalty.

First, ℓ_1 and ℓ_2 required in Lemmas B.7 and B.8 are chosen. Let regime ℓ_1 be either a Type R regime or a Type N next to neither Type N nor Type R regime on the right. If the closest non-Type N regime on the right of regime ℓ is of Type R, M, or none, set ℓ_2 as the last of these consecutive Type N regimes, otherwise, set ℓ_2 as the Type L regime.

The above-mentioned choice of ℓ_1 and ℓ_2 guarantees that the dominating portions of the regime ℓ_1 and the adjacent regime on the left (regime ℓ_2 and the adjacent regime on the right) if exists are from two different true regimes. This guarantees that conditions (i), (iii), and (iv) in Lemmas B.7 and B.8 hold. For condition (ii), consider the case where regime ℓ_1 is of Type R, regime ℓ_2 is of Type N, and $\ell_2 = k'+1$. Other cases can be handled in the same manner. First, we have $\bar{\boldsymbol{\theta}}^{(k')} - \bar{\boldsymbol{\theta}}^{(k'+1)} = 0$. Second, $\bar{\boldsymbol{\theta}}^{(\ell_1)} - \bar{\boldsymbol{\theta}}^{(\ell_1+1)} = O_p(n_1^{(\ell_1)}/n^{(\ell_1)})$. For the modified unbounded penalty, Lemma B.4, $n_1^{(\ell_1)}$ (the dominated portion in regime ℓ_1) is $\le O_p(\lambda^*)$. Using condition MUB2, that is $2\beta - \alpha < \kappa$, we have $n_1^{(\ell_1)}/n^{(\ell_1)}$ is dominated by $(\lambda/n^{(\ell_1)})^{1/2}$. Similarly, for the modified bridge penalty, $n_1^{(\ell_1)}/n^{(\ell_1)}$ is dominated by $(\lambda/n^{(\ell_1)})^{1/(2-\gamma)}$. □

Proof of Theorem 3.2(iii). That $|\hat{q}^{(\ell)} - q^{(\ell)}| \le \lambda^*/n \le O_p(n^{\beta+\varepsilon-1})$ is a direct consequence of Lemma B.4. Consider

$$\hat{\boldsymbol{\theta}}^{(\ell)} - \boldsymbol{\theta}_0^{(\ell)} = \hat{\boldsymbol{\theta}}^{(\ell)} - \bar{\boldsymbol{\theta}}^{(\ell)} + \bar{\boldsymbol{\theta}}^{(\ell)} - \boldsymbol{\theta}_0^{(\ell)}.$$

From the consistency of k' in Theorem 3.2(ii) and the consistency of $\hat{q}^{(\ell)}$, all regimes in $\hat{\boldsymbol{\theta}}$ must have lengths $O_p(n)$. Then, Lemma B.2 together with standard arguments based on the central limit theorem suggest that $\|\hat{\boldsymbol{\theta}}^{(\ell)} - \bar{\boldsymbol{\theta}}^{(\ell)}\| \le O_p(n^{\max\{\alpha,1/2\}-1})$ which is $\le O_p(n^{\beta+\varepsilon-1})$ under the condition MUB1 (MBR1). In addition, using $|\hat{q}^{(\ell)} - q^{(\ell)}| < O_p(n^{\beta+\varepsilon})$ and standard arguments based on Taylor expansion, $\|\bar{\boldsymbol{\theta}}^{(\ell)} - \boldsymbol{\theta}_0^{(\ell)}\| \le O_p(n^{\beta+\varepsilon-1})$. □

Proof of Theorem 3.3. It is a direct consequence of Lemma B.2. □

Proof of Theorem 3.4. Consider Definition A.3 and the first-order conditions described in Proposition A.2:

$$\begin{aligned}
&-\sum_{i=1}^{t^{(1)}-1} \nabla_{\boldsymbol{\theta}} \log f\big(X_i; \hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\phi}}\big) = -\mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(1)}\|\big) \cdot \hat{\mathbf{j}}^{(1)}, \\
&-\nabla_{\boldsymbol{\theta}} \log f\big(X_{t^{(1)}}; \hat{\boldsymbol{\theta}}^{(2)}, \hat{\boldsymbol{\phi}}\big) = -\big[\mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(2)}\|\big) \cdot \hat{\mathbf{j}}^{(2)} - \mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(1)}\|\big) \cdot \hat{\mathbf{j}}^{(1)}\big], \\
&\vdots \\
&-\nabla_{\boldsymbol{\theta}} \log f\big(X_{t^{(1)}+k'-2}; \hat{\boldsymbol{\theta}}^{(k')}, \hat{\boldsymbol{\phi}}\big) = -\big[\mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(k')}\|\big) \cdot \hat{\mathbf{j}}^{(k')} - \mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(k'-1)}\|\big) \cdot \hat{\mathbf{j}}^{(k-1)}\big], \\
&-\sum_{i=t^{(1)}+k'-1}^{n} \nabla_{\boldsymbol{\theta}} \log f\big(X_i; \hat{\boldsymbol{\theta}}^{(k'+1)}, \hat{\boldsymbol{\phi}}\big) = \mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(k')}\|\big) \cdot \hat{\mathbf{j}}^{(k')}, \\
&-\sum_{i=1}^{t^{(1)}-1} \nabla_{\boldsymbol{\phi}} \log f\big(X_i; \hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\phi}}\big) - \cdots - \sum_{i=t^{(1)}+k'-1}^{n} \nabla_{\boldsymbol{\phi}} \log f\big(X_i; \hat{\boldsymbol{\theta}}^{(k+1)}, \hat{\boldsymbol{\phi}}\big) = 0.
\end{aligned} \tag{B.1}$$

Let Ω be the event that condition (i) in Proposition A.2 holds. To obtain $P(\Omega)$, we consider approximations to $\hat{\mathbf{j}}^{(1)}, \hat{\mathbf{j}}^{(2)}, \ldots, \hat{\mathbf{j}}^{(k')}$, $\hat{\boldsymbol{\theta}}^{(1)}, \ldots, \hat{\boldsymbol{\theta}}^{(k'+1)}$, and $\hat{\boldsymbol{\phi}}$. Below, we discuss $P(\Omega)$ under different penalties. For the modified unbounded (modified bridge) penalty, further arguments are provided to rule out the possibility of $\|\hat{\boldsymbol{\theta}}^{(\ell)}\| = B$ for some $\ell = 1, 2, \ldots, k'$ (see Remark A.1).

LASSO penalty: Suppose that condition (i) in Proposition A.2 holds. That means, the equation (B.1) admits a solution. First, from the first and the k'th equations in (B.1), $\hat{\boldsymbol{\theta}}^{(1)}$ and $\hat{\boldsymbol{\theta}}^{(k'+1)}$ consistently estimates $\boldsymbol{\theta}_0^{(1)}$ and $\boldsymbol{\theta}_0^{(2)}$ since $\mathcal{P}'_\lambda(\cdot)$ is constant λ and $n^{-1}\lambda \to 0$. This can be shown using standard arguments based on Taylor expansion are applicable. Then, from the second to $(k'-1)$th equations in (B.1),

$$\hat{\mathbf{j}}^{(1)} \approx \hat{\mathbf{j}}^{(2)} \approx \cdots \approx \hat{\mathbf{j}}^{(k')} \approx \frac{\hat{\boldsymbol{\theta}}^{(1)} - \hat{\boldsymbol{\theta}}^{(k'+1)}}{\|\hat{\boldsymbol{\theta}}^{(1)} - \hat{\boldsymbol{\theta}}^{(k'+1)}\|} \approx \frac{\boldsymbol{\theta}_0^{(1)} - \boldsymbol{\theta}_0^{(2)}}{\|\boldsymbol{\theta}_0^{(1)} - \boldsymbol{\theta}_0^{(2)}\|}.$$

Otherwise, the right-hand sides diverge while the left-hand sides are finite. Here, we claim that $\lambda\hat{\mathbf{j}}^{(k'+1)} - \lambda\hat{\mathbf{j}}^{(1)}$ is nearly perpendicular to $\hat{\mathbf{j}}^{(1)}$, that is $\boldsymbol{\theta}_0^{(1)} - \boldsymbol{\theta}_0^{(k'+1)}$ approximately. To see this, consider $\lambda\hat{\mathbf{j}}^{(k'+1)} - \lambda\hat{\mathbf{j}}^{(1)}$ as the base of an isosceles triangle formed by $\lambda\hat{\mathbf{j}}^{(1)}$ and $\lambda\hat{\mathbf{j}}^{(k'+1)}$. If $\hat{\mathbf{j}}^{(1)} \approx \hat{\mathbf{j}}^{(k'+1)}$, the angle between $\lambda\hat{\mathbf{j}}^{(1)}$ and $\lambda\hat{\mathbf{j}}^{(k'+1)}$ is small. Therefore, the base angle is close to a right angle. Let $\boldsymbol{\theta}(X)$ be a point lying on the line segment from $\boldsymbol{\theta}_0^{(1)}$ to $\boldsymbol{\theta}_0^{(2)}$ such that $\nabla \log f(X; \boldsymbol{\theta}(X), \boldsymbol{\phi}_0)$ is perpendicular to $\boldsymbol{\theta}_0^{(1)} - \boldsymbol{\theta}_0^{(2)}$. Asymptotically, we have $\hat{\boldsymbol{\theta}}^{(k'+1)} \approx \boldsymbol{\theta}(X_{t^{(1)}})$. Consider $\hat{\boldsymbol{\theta}}^{(1)} \approx \boldsymbol{\theta}_0^{(1)}, \hat{\boldsymbol{\theta}}^{(k'+1)} \approx \boldsymbol{\theta}_0^{(2)},$ and $\hat{\boldsymbol{\phi}} \approx \boldsymbol{\phi}_0$. Event Ω can be approximated as

$$\boldsymbol{\theta}_0^{(1)} < \boldsymbol{\theta}(X_{t^{(1)}}) < \cdots < \boldsymbol{\theta}(X_{t^{(1)}+k'-1}) < \boldsymbol{\theta}_0^{(k'+1)}.$$

It can be seen that $P(\Omega)$ converges to neither zero nor one.

SCAD penalty: Under the regularity condition (R3), the functions

$$\log f\big(X_{t^{(1)}}; \boldsymbol{\theta}^{(2)}, \boldsymbol{\phi}\big), \ldots, \log f\big(X_{t^{(1)}+k'-2}; \boldsymbol{\theta}^{(k')}, \boldsymbol{\phi}\big)$$

must have maximums fulfilling the following equations

$$\nabla \log f\big(X_{t^{(1)}}; \hat{\boldsymbol{\theta}}^{(2)}, \hat{\boldsymbol{\phi}}\big) = \cdots = \nabla \log f\big(X_{t^{(1)}+k'-2}; \hat{\boldsymbol{\theta}}^{(k')}, \hat{\boldsymbol{\phi}}\big) = 0.$$

Since k' is finite, the quantities $\hat{\boldsymbol{\xi}}^{(1)}, \hat{\boldsymbol{\xi}}^{(2)}, \ldots, \hat{\boldsymbol{\xi}}^{(k')}$ are all $O_p(1)$. Consequently,

$$\mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(1)}\|\big) = \mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(2)}\|\big) = \cdots = \mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(k')}\|\big) = 0.$$

Such quantities $\hat{\boldsymbol{\xi}}^{(1)}, \hat{\boldsymbol{\xi}}^{(2)}, \ldots, \hat{\boldsymbol{\xi}}^{(k')}$ fulfills equation (B.1). Therefore, $P(\Omega) \to 1$.

Bridge and unbounded penalties: If Ω holds, the second to $(k'-1)$th equations in (B.1) implies both

$$\mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(1)}\|\big) \approx \mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(2)}\|\big) \approx \cdots \approx \mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(k')}\|\big)$$

and

$$\hat{\mathbf{j}}^{(1)} \approx \hat{\mathbf{j}}^{(2)} \approx \cdots \approx \hat{\mathbf{j}}^{(k')}.$$

Otherwise, the right-hand sides diverge while the left-hand sides are finite. If $\hat{\boldsymbol{\theta}}^{(2)}$ bisects the line segment joining $\hat{\boldsymbol{\theta}}^{(1)}$ and $\hat{\boldsymbol{\theta}}^{(3)}$, although the stationary point may exist, it is not a local minimum. To see this, consider the second derivatives of Q^*_λ with respect to $\boldsymbol{\theta}^{(2)}$, which can be approximated as

$$-\nabla^{(2)}_{\boldsymbol{\theta}\boldsymbol{\theta}} \log f\big(X_{t^{(1)}}; \hat{\boldsymbol{\theta}}^{(2)}, \hat{\boldsymbol{\phi}}\big) + \lambda\tau^{-1}(\tau-2)\big\|\hat{\boldsymbol{\xi}}^{(2)}\big\|^{-2}\big[I - 2\hat{\mathbf{j}}^{(2)}\big(\hat{\mathbf{j}}^{(2)}\big)^T\big]$$

for the unbounded penalty and

$$-\nabla^{(2)}_{\boldsymbol{\theta}\boldsymbol{\theta}} \log f\big(X_{t^{(1)}}; \hat{\boldsymbol{\theta}}^{(2)}, \hat{\boldsymbol{\phi}}\big) + \lambda\gamma\big\|\hat{\boldsymbol{\xi}}^{(2)}\big\|^{\gamma-2}\big[I - (2-\gamma)\hat{\mathbf{j}}^{(2)}\big(\hat{\mathbf{j}}^{(2)}\big)^T\big]$$

for bridge, see Convention B.4. Left multiplying by $(\hat{\mathbf{j}}^{(2)})^T$ and right multiplying by $\hat{\mathbf{j}}^{(2)}$, the penalty-related part which is negative dominates. The second derivative is no longer positive definite.

Modified bridge and modified unbounded penalties: Consider the following four exhaustive and mutually exclusive cases.

(i) $\|\hat{\boldsymbol{\xi}}^{(\ell)}\| \le B$ for all $\ell = 1, 2, \ldots, k'$. Under the assumption that k' is finite, $\hat{\boldsymbol{\theta}}_i$, $i = 1, 2, \ldots, n$ differ each other only by some $o_p(1)$ quantities. Then, it can be checked using standard arguments based on Taylor expansion that all $\hat{\boldsymbol{\theta}}_i$, $i = 1, 2, \ldots, n$ have at least $O_p(1)$ distance from either $\boldsymbol{\theta}_0^{(1)}$ or $\boldsymbol{\theta}_0^{(2)}$. Then, $\max_{i=1,\ldots,t^{(1)}-1} \|\mathbf{S}_i^{(1)}\| = O_p(n)$ cannot be bounded by λ^*. Indeed, this is necessary due to the local minimality of $Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\phi})$ along the directions of $\boldsymbol{\theta}_i$ for $i = 1, \ldots, t^{(1)} - 1$.

(ii) $\|\hat{\boldsymbol{\xi}}^{(\ell)}\| > B$ for at least one $\ell = 1, 2, \ldots, k'$ and $\|\hat{\boldsymbol{\xi}}^{(\ell)}\| = B$ for at least one $\ell = 1, 2, \ldots, k'$. Without loss of generality assume that $\|\hat{\boldsymbol{\xi}}^{(1)}\| = B$ and $\|\hat{\boldsymbol{\xi}}^{(2)}\| > B$. Then, we have the first-order condition

$$-\sum_{i=1}^{t^{(1)}-1} \nabla_{\boldsymbol{\theta}} \log f\big(X_i; \hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\phi}}\big) - \nabla_{\boldsymbol{\theta}} \log f\big(X_{t^{(1)}}; \hat{\boldsymbol{\theta}}^{(2)}, \hat{\boldsymbol{\phi}}\big) = -\mathcal{P}'_\lambda\big(\big\|\hat{\boldsymbol{\xi}}^{(2)}\big\|\big) \cdot \hat{\mathbf{j}}^{(2)}. \qquad \text{(B.2)}$$

Standard arguments show that $\hat{\boldsymbol{\theta}}^{(1)}$ consistently estimates $\boldsymbol{\theta}_0^{(1)}$. In addition,

$$\max_{i=1,\ldots,t^{(1)}-1} \big\|\mathbf{S}_i^{(1)}\big\| = O_p\big(n^{\max\{1/2,\alpha\}}\big) = o_p\big(\lambda^*\big).$$

However, the local minimality of $Q_\lambda(\boldsymbol{\theta}, \boldsymbol{\phi})$ along the directions of $\boldsymbol{\theta}_{t^{(1)}}$ requires that

$$\lambda^* < \big|\mathbf{S}^{(1)}_{t^{(1)}-1} - \nabla_{\boldsymbol{\theta}} \log f\big(X_{t^{(1)}}; \hat{\boldsymbol{\theta}}^{(2)}, \hat{\boldsymbol{\phi}}\big)\big| < P'_\lambda(B).$$

Here, $P_\lambda(\cdot)$ is the unbounded (bridge) penalty function. This is impossible because $\nabla_{\boldsymbol{\theta}} \log f(X_{t^{(1)}}; \hat{\boldsymbol{\theta}}^{(2)}, \hat{\boldsymbol{\phi}})$ is $O_p(1)$ quantity but $\lambda^* \gg \lambda$ from Condition MUB1 (MBR1).

(iii) $\|\hat{\boldsymbol{\xi}}^{(\ell)}\| > B$ for one of $\ell = 1, 2, \ldots, k'$ and $\|\hat{\boldsymbol{\xi}}^{(\ell)}\| \neq B$ for all $\ell = 1, 2, \ldots, k'$. Without loss of generality, assume that $\|\hat{\boldsymbol{\xi}}^{(1)}\| > B$. Then, $\mathcal{P}'_\lambda(\|\hat{\boldsymbol{\xi}}^{(k')}\|) = \lambda$. This suggests that $\hat{\boldsymbol{\theta}}^{(k'+1)}$ consistently estimates $\boldsymbol{\theta}_0^{(2)}$. Consider the sum of the first two equations in (B.1),

$$-\sum_{i=1}^{t^{(1)}-1} \nabla_{\boldsymbol{\theta}} \log f\big(X_i; \hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\phi}}\big) - \nabla_{\boldsymbol{\theta}} \log f\big(X_{t^{(1)}}; \hat{\boldsymbol{\theta}}^{(2)}, \hat{\boldsymbol{\phi}}\big) = -\mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(2)}\|\big) \cdot \hat{\mathbf{j}}^{(2)}.$$

Similarly, $\hat{\boldsymbol{\theta}}^{(1)}$ consistently estimates $\boldsymbol{\theta}_0^{(1)}$. However, this is impossible because the left-hand-side of the second equation in (B.1) is $O_p(1)$ while the right-hand side is $O_p(\lambda^*)$.

(iv) $\|\hat{\boldsymbol{\xi}}^{(\ell)}\| > B$ for two or more $\ell = 1, 2, \ldots, k'$ and $\|\hat{\boldsymbol{\xi}}^{(\ell)}\| \neq B$ for all $\ell = 1, 2, \ldots, k'$. Without loss of generality assume that $\|\hat{\boldsymbol{\xi}}^{(\ell)}\| > B$ for $\ell = 1$ and $\ell = k'$. Adding up the second to $(k'-1)$th equations in (B.1),

$$-\sum_{\ell=2}^{k'} \nabla_{\boldsymbol{\theta}} \log f\big(X_{t^{(1)}+\ell-2}; \hat{\boldsymbol{\theta}}^{(\ell)}, \hat{\boldsymbol{\phi}}\big) = -\big[\mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(k')}\|\big) \cdot \hat{\mathbf{j}}^{(k')} - \mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(1)}\|\big) \cdot \hat{\mathbf{j}}^{(1)}\big].$$

Note that the left-hand side is $O_p(1)$. Then, the same arguments as in the proofs of bridge and unbounded penalties are then applicable to derive contradiction.

□

B.3. Technical lemmas

In this subsection, the regularity conditions R1–R4 are assumed without mentioning.

Lemma B.1. *Let $C > 0$ be arbitrary constant and δ be defined in condition SCAD3 for SCAD penalty (UB3 for the unbounded penalty, BR3 for the bridge penalty, MUB3 for the modified unbounded penalty, and MBR3 for the modified bridge penalty). Then, under conditions SCAD1–SCAD3 (UB1–UB3, BR1–BR3, MUB1–MUB3, MBR1–MBR3), with probability going to one, all restricted local solution in $\aleph$ has $\|\hat{\boldsymbol{\theta}}^{(\ell)} - \bar{\boldsymbol{\theta}}^{(\ell)}\| < Cn^{-\delta}$ for all ℓ as $n \to \infty$ (see Convention B.2 for the notation of $\bar{\boldsymbol{\theta}}^{(\ell)}$).*

Proof. Let ℓ^* be a regime with $\|\hat{\boldsymbol{\theta}}^{(\ell)} - \bar{\boldsymbol{\theta}}^{(\ell)}\| \geq Cn^{-\delta}$. Proposition A.2 suggests that

$$\begin{aligned}
&-\sum_{i=t^{(\ell^*-1)}}^{t^{(\ell^*)}-1} \nabla_{\boldsymbol{\theta}} \log f\big(X_i; \hat{\boldsymbol{\theta}}^{(\ell^*)}, \hat{\boldsymbol{\phi}}\big) \\
&\quad = -\big[\mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(\ell^*)}\|\big) \cdot \hat{\mathbf{j}}^{(\ell^*)} - \mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(\ell^*-1)}\|\big) \cdot \hat{\mathbf{j}}^{(\ell^*-1)}\big].
\end{aligned} \tag{B.3}$$

Consider the Taylor expansion of the left-hand-side of (B.3) around $(\bar{\boldsymbol{\theta}}^{(\ell^*)}, \bar{\boldsymbol{\phi}})$.

For the SCAD (bridge, unbounded, modified bridge, modified unbounded) penalty, SCAD1–SCAD3 (BR1–BR3, UB1–UB3, MBR1–MBR3, MUB1–MUB3) guarantees that $n^{(\ell^*)} \to \infty$ and $\delta < \kappa/2$. Therefore, the left-hand side, with the second-order term of Taylor expansion dominating the first order term, has order $\geq O_p(n^{\kappa-\delta})$. Let $\varepsilon > 0$ be arbitrarily small. Condition R2 and triangular inequality suggests that

$$\max\{\mathcal{P}'_\lambda(\|\hat{\boldsymbol{\xi}}^{(\ell^*)}\|), \mathcal{P}'_\lambda(\|\hat{\boldsymbol{\xi}}^{(\ell^*-1)}\|)\} \geq \frac{1}{2}\left\|\sum_{i=t^{(\ell^*-1)}}^{t^{(\ell^*)}-1} \nabla_{\boldsymbol{\theta}} \log f\big(X_i; \hat{\boldsymbol{\theta}}^{(\ell^*)}, \hat{\boldsymbol{\phi}}\big)\right\| \geq O_p\big(n^{\kappa-\delta+\varepsilon}\big).$$

Such bound does not depend on the location of the change points in $\hat{\boldsymbol{\theta}}$. Without loss of generality, suppose that $\mathcal{P}'_\lambda(\|\hat{\boldsymbol{\xi}}^{(\ell^*-1)}\|) \geq O_p(n^{\kappa-\delta})$. For the SCAD cases, $\mathcal{P}'_\lambda(\|\hat{\boldsymbol{\xi}}^{(\ell^*-1)}\|) \leq \lambda = O_p(n^\alpha)$. Then, $O_p(n^{\kappa-\delta}) \leq O_p(n^\alpha)$, violating condition SCAD3. For the unbounded penalty cases, $\mathcal{P}'_\lambda(\|\hat{\boldsymbol{\xi}}^{(\ell^*-1)}\|) \geq O_p(n^{\kappa-\delta})$ implies that $\|\hat{\boldsymbol{\xi}}^{(\ell^*-1)}\| \leq O_p(n^{\alpha-\kappa+\delta})$. However, this violates UB3 since $\|\hat{\boldsymbol{\xi}}^{(\ell^*-1)}\| \geq O(n^{-\delta})$. Similarly, for the bridge penalty cases, $\|\hat{\boldsymbol{\xi}}^{(\ell^*-1)}\| \leq O_p(n^{(\alpha-\kappa+\delta)/(1-\gamma)})$ violates BR3. Modified bridge and modified unbounded penalties can be handled using the same arguments as above. □

Lemma B.2. *Suppose that conditions SCAD*1*–SCAD*3 *for SCAD penalty* (*UB*1*–UB*3 *for the unbounded penalty*, *BR*1*–BR*3 *for the bridge penalty*, *MUB*1*–MUB*3 *for the modified unbounded penalty*, *and MBR*1*–MBR*3 *for the modified bridge penalty*) *hold. Then*, *for any restricted local solution* $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}})$ *in* $\aleph$, $\hat{\boldsymbol{\phi}}$ *and the distinct values of* $\hat{\boldsymbol{\theta}}$ *can be approximated as follows*:

$$\begin{aligned}
\hat{\boldsymbol{\phi}} - \bar{\boldsymbol{\phi}} &\approx \frac{1}{n}\bar{\mathbf{H}}^{-1}_{\boldsymbol{\phi}\boldsymbol{\phi}|\boldsymbol{\theta}} L_{\boldsymbol{\phi}|\boldsymbol{\theta}}, \\
\hat{\boldsymbol{\theta}}^{(1)} - \bar{\boldsymbol{\theta}}^{(1)} &\approx \frac{1}{n^{(1)}}\big(\bar{\mathbf{H}}^{(1)}_{\boldsymbol{\theta}\boldsymbol{\theta}}\big)^{-1} L^{(1)}_{\boldsymbol{\theta}} - \frac{1}{n}\big(\bar{\mathbf{H}}^{(1)}_{\boldsymbol{\theta}\boldsymbol{\theta}}\big)^{-1}\bar{\mathbf{H}}^{(1)}_{\boldsymbol{\theta}\boldsymbol{\phi}}\bar{\mathbf{H}}^{-1}_{\boldsymbol{\phi}\boldsymbol{\phi}|\boldsymbol{\theta}} L_{\boldsymbol{\phi}|\boldsymbol{\theta}} \\
&\quad - \frac{1}{n^{(1)}}\big[\bar{\mathbf{H}}^{(1)}_{\boldsymbol{\theta}\boldsymbol{\theta}}\big]^{-1}\mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(1)}\|\big)\hat{\mathbf{j}}^{(1)}, \\
\hat{\boldsymbol{\theta}}^{(2)} - \bar{\boldsymbol{\theta}}^{(2)} &\approx \frac{1}{n^{(2)}}\big(\bar{\mathbf{H}}^{(2)}_{\boldsymbol{\theta}\boldsymbol{\theta}}\big)^{-1} L^{(2)}_{\boldsymbol{\theta}} - \frac{1}{n}\big(\bar{\mathbf{H}}^{(2)}_{\boldsymbol{\theta}\boldsymbol{\theta}}\big)^{-1}\bar{\mathbf{H}}^{(2)}_{\boldsymbol{\theta}\boldsymbol{\phi}}\bar{\mathbf{H}}^{-1}_{\boldsymbol{\phi}\boldsymbol{\phi}|\boldsymbol{\theta}} L_{\boldsymbol{\phi}|\boldsymbol{\theta}} \\
&\quad - \frac{1}{n^{(2)}}\big[\bar{\mathbf{H}}^{(2)}_{\boldsymbol{\theta}\boldsymbol{\theta}}\big]^{-1}\big[\mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(2)}\|\big)\hat{\mathbf{j}}^{(2)} - \mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(1)}\|\big)\hat{\mathbf{j}}^{(1)}\big], \\
&\vdots \approx \vdots \\
\hat{\boldsymbol{\theta}}^{(k')} - \bar{\boldsymbol{\theta}}^{(k')} &\approx \frac{1}{n^{(k')}}\big(\bar{\mathbf{H}}^{(k')}_{\boldsymbol{\theta}\boldsymbol{\theta}}\big)^{-1} L^{(k')}_{\boldsymbol{\theta}} - \frac{1}{n}\big(\bar{\mathbf{H}}^{(k')}_{\boldsymbol{\theta}\boldsymbol{\theta}}\big)^{-1}\bar{\mathbf{H}}^{(k')}_{\boldsymbol{\theta}\boldsymbol{\phi}}\bar{\mathbf{H}}^{-1}_{\boldsymbol{\phi}\boldsymbol{\phi}|\boldsymbol{\theta}} L_{\boldsymbol{\phi}|\boldsymbol{\theta}} \\
&\quad - \frac{1}{n^{(k')}}\big[\bar{\mathbf{H}}^{(k')}_{\boldsymbol{\theta}\boldsymbol{\theta}}\big]^{-1}\big[\mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(k')}\|\big)\hat{\mathbf{j}}^{(k')} - \mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(k'-1)}\|\big)\hat{\mathbf{j}}^{(k'-1)}\big], \\
\hat{\boldsymbol{\theta}}^{(k'+1)} - \bar{\boldsymbol{\theta}}^{(k'+1)} &\approx \frac{1}{n^{(k'+1)}}\big(\bar{\mathbf{H}}^{(k'+1)}_{\boldsymbol{\theta}\boldsymbol{\theta}}\big)^{-1} L^{(k'+1)}_{\boldsymbol{\theta}} - \frac{1}{n}\big(\bar{\mathbf{H}}^{(k'+1)}_{\boldsymbol{\theta}\boldsymbol{\theta}}\big)^{-1}\bar{\mathbf{H}}^{(k'+1)}_{\boldsymbol{\theta}\boldsymbol{\phi}}\bar{\mathbf{H}}^{-1}_{\boldsymbol{\phi}\boldsymbol{\phi}|\boldsymbol{\theta}} L_{\boldsymbol{\phi}|\boldsymbol{\theta}} \\
&\quad + \frac{1}{n^{(k'+1)}}\big[\bar{\mathbf{H}}^{(k'+1)}_{\boldsymbol{\theta}\boldsymbol{\theta}}\big]^{-1}\mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(k')}\|\big)\hat{\mathbf{j}}^{(k')}.
\end{aligned}$$

Here,

$$\bar{\mathbf{H}}_{\boldsymbol{\phi}\boldsymbol{\phi}|\boldsymbol{\theta}} = \frac{1}{n}\sum_{\ell=1}^{k'+1} n^{(\ell)}\big[\bar{\mathbf{H}}^{(\ell)}_{\boldsymbol{\phi}\boldsymbol{\phi}} - \bar{\mathbf{H}}^{(\ell)}_{\boldsymbol{\phi}\boldsymbol{\theta}}\big(\bar{\mathbf{H}}^{(\ell)}_{\boldsymbol{\theta}\boldsymbol{\theta}}\big)^{-1}\bar{\mathbf{H}}^{(\ell)}_{\boldsymbol{\theta}\boldsymbol{\phi}}\big],$$

$$L^{(\ell)}_{\boldsymbol{\theta}} = \sum_{i=t^{(\ell-1)}}^{t^{(\ell)}-1} \nabla_{\boldsymbol{\theta}} \log f\big(X_i;\bar{\boldsymbol{\theta}}^{(\ell)},\bar{\boldsymbol{\phi}}\big),$$

$$L_{\boldsymbol{\phi}|\boldsymbol{\theta}} = \sum_{\ell=1}^{k'+1}\sum_{i=t^{(\ell-1)}}^{t^{(\ell)}-1}\big[\nabla_{\boldsymbol{\phi}} - \bar{\mathbf{H}}^{(\ell)}_{\boldsymbol{\phi}\boldsymbol{\theta}}\big(\bar{\mathbf{H}}^{(\ell)}_{\boldsymbol{\theta}\boldsymbol{\theta}}\big)^{-1}\nabla_{\boldsymbol{\theta}}\big]\log f\big(X_i;\bar{\boldsymbol{\theta}}^{(\ell)},\bar{\boldsymbol{\phi}}\big).$$

Proof. For simplicity, consider only the case $k' = 1$. Consider the first-order conditions for restricted local solution (see Definition A.2 and Proposition A.2). Lemmas B.1 and condition SCAD1–SCAD2 (UB1–UB2, BR1–BR2, MUB1–MUB2, MBR1–MBR2) guarantees that $\min n^{(\ell)} \to \infty$ and that the bias introduced by the penalty term does not diverge. Then, the standard arguments based on the central limit theorem can be applied on such first-order conditions. In addition, the third-order remainder terms in the Taylor expansion of the log-likelihood function (as a function of $\boldsymbol{\theta}^{(1)},\dots,\boldsymbol{\theta}^{(k'+1)}$ and $\boldsymbol{\phi}$) around $(\bar{\boldsymbol{\theta}}^{(1)},\dots,\bar{\boldsymbol{\theta}}^{(k'+1)},\bar{\boldsymbol{\phi}})$ are ignorable. Then, we have the following approximations:

$$\begin{pmatrix}\hat{\boldsymbol{\theta}}^{(1)}-\bar{\boldsymbol{\theta}}^{(1)}\\ \hat{\boldsymbol{\theta}}^{(2)}-\bar{\boldsymbol{\theta}}^{(2)}\\ \hat{\boldsymbol{\phi}}-\bar{\boldsymbol{\phi}}\end{pmatrix} \approx \begin{pmatrix} n^{(1)}\bar{\mathbf{H}}^{(1)}_{\boldsymbol{\theta}\boldsymbol{\theta}} & 0 & n^{(1)}\bar{\mathbf{H}}^{(1)}_{\boldsymbol{\theta}\boldsymbol{\phi}}\\ 0 & n^{(2)}\bar{\mathbf{H}}^{(2)}_{\boldsymbol{\theta}\boldsymbol{\theta}} & n^{(2)}\bar{\mathbf{H}}^{(2)}_{\boldsymbol{\theta}\boldsymbol{\phi}}\\ n^{(1)}\bar{\mathbf{H}}^{(1)}_{\boldsymbol{\phi}\boldsymbol{\theta}} & n^{(2)}\bar{\mathbf{H}}^{(2)}_{\boldsymbol{\phi}\boldsymbol{\theta}} & n^{(1)}\bar{\mathbf{H}}^{(1)}_{\boldsymbol{\phi}\boldsymbol{\phi}}+n^{(2)}\bar{\mathbf{H}}^{(2)}_{\boldsymbol{\phi}\boldsymbol{\phi}}\end{pmatrix}^{-1}$$

$$\times\begin{pmatrix} -\mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(1)}\|\big)\cdot\hat{\mathbf{j}}^{(1)} + \sum_{i=1}^{t^{(1)}-1}\nabla_{\boldsymbol{\theta}}\log f\big(X_i;\bar{\boldsymbol{\theta}}^{(1)},\bar{\boldsymbol{\phi}}\big)\\ +\mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(1)}\|\big)\cdot\hat{\mathbf{j}}^{(1)} + \sum_{i=t^{(1)}}^{n}\nabla_{\boldsymbol{\theta}}\log f\big(X_i;\bar{\boldsymbol{\theta}}^{(2)},\bar{\boldsymbol{\phi}}\big)\\ \sum_{\ell=1}^{2}\sum_{i=t^{(\ell-1)}}^{t^{(\ell)}-1}\nabla_{\boldsymbol{\phi}}\log f\big(X_i;\bar{\boldsymbol{\theta}}^{(\ell)},\bar{\boldsymbol{\phi}}\big)\end{pmatrix}$$

(see the notations defined in Convention B.2). Here, Taylor expansion is applied to the log-likelihood function only. The penalty terms are not approximated. After algebraic manipulations, we get

$$\hat{\boldsymbol{\phi}}-\bar{\boldsymbol{\phi}} \approx \left\{\sum_{\ell=1}^{2} n^{(\ell)}\big[\bar{\mathbf{H}}^{(\ell)}_{\boldsymbol{\phi}\boldsymbol{\phi}} - \bar{\mathbf{H}}^{(\ell)}_{\boldsymbol{\phi}\boldsymbol{\theta}}\big(\bar{\mathbf{H}}^{(\ell)}_{\boldsymbol{\theta}\boldsymbol{\theta}}\big)^{-1}\bar{\mathbf{H}}^{(\ell)}_{\boldsymbol{\theta}\boldsymbol{\phi}}\big]\right\}^{-1}$$

$$\times\left\{\sum_{\ell=1}^{2}\sum_{i=t^{(\ell-1)}}^{t^{(\ell)}-1}\big[\nabla_{\boldsymbol{\phi}} - \bar{\mathbf{H}}^{(\ell)}_{\boldsymbol{\phi}\boldsymbol{\theta}}\big(\bar{\mathbf{H}}^{(\ell)}_{\boldsymbol{\theta}\boldsymbol{\theta}}\big)^{-1}\nabla_{\boldsymbol{\theta}}\big]\log f\big(X_i;\bar{\boldsymbol{\theta}}^{(\ell)},\bar{\boldsymbol{\phi}}\big)\right\},$$

$$\hat{\boldsymbol{\theta}}^{(1)} - \bar{\boldsymbol{\theta}}^{(1)} \approx \frac{1}{n^{(1)}}\big(\bar{\mathbf{H}}^{(1)}_{\boldsymbol{\theta}\boldsymbol{\theta}}\big)^{-1}\left\{-\mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(1)}\|\big)\cdot\hat{\mathbf{j}}^{(1)} + \sum_{i=1}^{t^{(1)}-1}\nabla_{\boldsymbol{\theta}}\log f\big(X_i;\bar{\boldsymbol{\theta}}^{(1)},\bar{\boldsymbol{\phi}}\big)\right\} - \big(\bar{\mathbf{H}}^{(1)}_{\boldsymbol{\theta}\boldsymbol{\theta}}\big)^{-1}\bar{\mathbf{H}}^{(1)}_{\boldsymbol{\theta}\boldsymbol{\phi}}(\hat{\boldsymbol{\phi}} - \bar{\boldsymbol{\phi}}),$$

$$\hat{\boldsymbol{\theta}}^{(2)} - \bar{\boldsymbol{\theta}}^{(2)} \approx \frac{1}{n^{(2)}}\big(\bar{\mathbf{H}}^{(2)}_{\boldsymbol{\theta}\boldsymbol{\theta}}\big)^{-1}\left\{+\mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(1)}\|\big)\cdot\hat{\mathbf{j}}^{(1)} + \sum_{i=t^{(1)}}^{n}\nabla_{\boldsymbol{\theta}}\log f\big(X_i;\bar{\boldsymbol{\theta}}^{(2)},\bar{\boldsymbol{\phi}}\big)\right\} - \big(\bar{\mathbf{H}}^{(2)}_{\boldsymbol{\theta}\boldsymbol{\theta}}\big)^{-1}\bar{\mathbf{H}}^{(2)}_{\boldsymbol{\theta}\boldsymbol{\phi}}(\hat{\boldsymbol{\phi}} - \bar{\boldsymbol{\phi}}).$$

The desired results follow immediately. □

Lemma B.3. *Suppose that conditions UB*1*–UB*3 *for the unbounded penalty* (*BR*1*–BR*3 *for the bridge penalty*, *MUB*1*–MUB*3 *for the modified unbounded penalty*, *and MBR*1*–MBR*3 *for the modified bridge penalty*) *hold. Then*, *all restricted local solutions* $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}})$ *in* $\aleph$ *satisfy*

$$\mathbf{S}^{(\ell)}_i = -\mathbf{y}^{(\ell)}_i - \mathbf{m}_i - \mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(\ell-1)}\|\big)\cdot\hat{\mathbf{j}}^{(\ell-1)} - \frac{i}{n^{(\ell)}}\cdot\big[\mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(\ell)}\|\big)\cdot\hat{\mathbf{j}}^{(\ell)} - \mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(\ell-1)}\|\big)\cdot\hat{\mathbf{j}}^{(\ell-1)}\big],$$

where

$$\mathbf{m}^{(\ell)}_i = \sum_{j=t^{(\ell-1)}}^{t^{(\ell-1)}+i-1} E\nabla_{\boldsymbol{\theta}}\log f(X_j;\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\phi}}) - \frac{i}{n^{(\ell)}}\sum_{j=t^{(\ell-1)}}^{t^{(\ell)}-1} E\nabla_{\boldsymbol{\theta}}\log f(X_j;\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\phi}})$$

and

$$\mathbf{y}^{(\ell)}_i = -\mathbf{m}^{(\ell)}_i + \sum_{j=t^{(\ell-1)}}^{t^{(\ell-1)}+i-1} \nabla_{\boldsymbol{\theta}}\log f(X_j;\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\phi}}) - \frac{i}{n^{(\ell)}}\sum_{j=t^{(\ell-1)}}^{t^{(\ell)}-1} \nabla_{\boldsymbol{\theta}}\log f(X_j;\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\phi}})$$

for $\ell = 1, 2, \ldots, k' + 1$, $i = 1, \ldots, n^{(\ell)} - 1$.

Proof. From Proposition A.2,

$$\mathbf{S}^{(\ell)}_i = -\mathcal{P}'_\lambda\big(\|\hat{\boldsymbol{\xi}}^{(\ell-1)}\|\big)\cdot\hat{\mathbf{j}}^{(\ell-1)} - \sum_{j=t^{(\ell-1)}}^{t^{(\ell-1)}+i-1}\nabla_{\boldsymbol{\theta}}\log f\big(X_j;\hat{\boldsymbol{\theta}}^{(\ell)},\hat{\boldsymbol{\phi}}\big).$$

Moreover, Lemma B.1 and conditions UB1–UB2 (BR1–BR2, MUB1–MUB2, MBR1–MBR2) guarantee that $n^{(\ell)} \to \infty$ and $\|\hat{\boldsymbol{\theta}}^{(\ell)} - \bar{\boldsymbol{\theta}}^{(\ell)}\| \to 0$. Then, the standard arguments based on Taylor approximation can be applied and the desired results follow immediately from Lemma B.2. □

Lemma B.4. (i) *For the modified unbounded* (*modified bridge*) *penalty*, *under conditions MUB*1*–MUB*3 (*MBR*1*–MBR*3), *with probability going to one*, *all regimes in any restricted local solution in* $\aleph$ *belong to one of the four types N*, *M*, *R*, *and L* (*see Convention* B.3). (ii) *For the* *first regime, only type N and type L are allowed.* (iii) *For the last regime, only type N and type R are allowed.* (iv) *For Type L, R, and M regimes, the length(s) of the dominated portion(s) is (are) bounded above by $O_p(\lambda^*)$ quantities.*

Proof. First, we see that three or more change points are not allowed. By contradiction assume that such solution exists and satisfies the first-order conditions in Proposition A.2. Consider the expression of $\mathbf{S}_i^{(\ell)}$ in Lemma B.3. Note that the maximum $\max_{i=t^{(\ell-1)},\dots,t^{(\ell)}-1}\|\mathbf{m}_i^{(\ell)}\|$ is $O_p(n)$, dominating $\|\mathbf{y}_i^{(\ell)}\| \le O_p(\sqrt{n^{(\ell)}})$, see [7]. Next bounds of $\mathbf{P}_{\text{start}} = \mathcal{P}'_\lambda(\|\hat{\boldsymbol{\xi}}^{(\ell-1)}\|)\cdot\hat{\mathbf{j}}^{(\ell-1)}$ and $\mathbf{P}_{\text{end}} = \mathcal{P}'_\lambda(\|\hat{\boldsymbol{\xi}}^{(\ell)}\|)\cdot\hat{\mathbf{j}}^{(\ell)}$ are given. From Proposition A.2, $\|\mathbf{S}_1^{(\ell)}\| < \lambda^*$ and $\|\mathbf{S}_{n^{(\ell)}-1}^{(\ell)}\| < \lambda^*$. That means, $\|\mathbf{P}_{\text{start}} + (\mathbf{P}_{\text{end}} - \mathbf{P}_{\text{start}})/n^{(\ell)} + O_p(1)\| < \lambda^*$ and $\|\mathbf{P}_{\text{end}} - (\mathbf{P}_{\text{end}} - \mathbf{P}_{\text{start}})/n^{(\ell)} + O_p(1)\| < \lambda^*$. Note that under MUB1 and MUB3 (MBR1 and MBR3), $n^{-\delta} \gg n^{-\kappa/2}$. Together with from Lemma B.1 and Lemma B.2, we have $\|(\mathbf{P}_{\text{end}} - \mathbf{P}_{\text{start}})/n^{(\ell)}\| \le O_p(n^{-\delta}) = o_p(\lambda^*)$. Therefore, both $\mathbf{P}_{\text{start}}$ and $\mathbf{P}_{\text{end}}$ are at most $O_p(\lambda^*)$. Consequently,

$$\max_{i=1,\dots,n^{(\ell)}-1}\left\|\mathbf{S}_i^{(\ell)}\right\| = O_p(n).$$

Then, the conditions in Proposition A.2, namely $\|\mathbf{S}_i^{(\ell)}\| < \lambda^* = O_p(n^\beta)$ cannot be satisfied for all $i = 1,\dots,n^{(\ell)}-1$.

Next, consider regime ℓ^* that consists of one true change point. In what follows, we show that regime ℓ^* must be either type L or type R. Since $\|\mathbf{S}_i^{(\ell^*)}\| < \lambda^*$ (Proposition A.2) for all $i = 1,\dots,n^{(\ell^*)}-1$, from Lemma B.3,

$$\max_{i=1,\dots,n^{(\ell^*)}-1}\left\|\mathbf{m}_i^{(\ell^*)}\right\| \le \lambda^* + \max\left\{\|\mathbf{P}_{\text{start}}\|, \|\mathbf{P}_{\text{end}}\|\right\} + \max_{i=1,\dots,n^{(\ell^*)}-1}\left\|\mathbf{y}_i^{(\ell^*)}\right\|.$$

Here, $\|\mathbf{y}_i^{(\ell^*)}\| \le O_p(\sqrt{n^{(\ell^*)}}) = o_p(\lambda^*)$ and is ignorable. Note that the quantity

$$\max_{i=1,\dots,n^{(\ell^*)}-1}\left\|\mathbf{m}_i^{(\ell^*)}\right\|$$

has the same order as $\min\{n_1^{\ell^*}, n_2^{\ell^*}\}$. Then, $\min\{n_1^{\ell^*}, n_2^{\ell^*}\} = o_p(\lambda^*)$. On the other hand, from condition MUB1–MUB2 (MBR1–MBR2), we have $\kappa > \beta$. Then, $n^{(\ell^*)} \gg O_p(\lambda^*)$. Consequently, regime ℓ^* must be either Type R or Type L.

Last, similar arguments as above show that for a regime with two change points, both the left portion and the right portion must not exceed some $O_p(\lambda^*)$ quantity while the middle portion is $O_p(n)$. Therefore, it is type M. □

Lemma B.5. *For $k = 0$, we have*

(i) *For the LASSO penalty, if $\lambda = O(n^{1/2})$, $P_0 \to C$ for some $C \in (0, 1)$.*
(ii) *For LASSO and SCAD penalties, if $n^{-1/2}\lambda \to \infty$, $P_0 \to 1$.*
(iii) *For the unbounded (bridge) penalty, if $\lambda > 0$, $P_0 \to 1$.*
(iv) *For the modified unbounded (modified bridge) penalty, if $n^{-1/2}\lambda^* \to \infty$, $P_0 \to 1$.*

Proof. The notations in Proposition A.1 are used. The proof is based on the following results (see, for example, [7]),

$$\frac{1}{\sqrt{n}}\mathbf{S}^{(1)}_{[sn]} \longrightarrow -(\mathbf{Y}_s - s\mathbf{Y}_1), \qquad 0 < s < 1,$$

where $\mathbf{Y}_s$ is a p-dimensional Brownian motion with covariance matrix

$$\mathbf{H}^{(1)}_{\boldsymbol{\theta\theta}} = -E\nabla^2_{\boldsymbol{\theta\theta}} \log f\big(X_1; \boldsymbol{\theta}^{(1)}_0, \boldsymbol{\phi}_0\big).$$

For the LASSO penalty with $\lambda = O_p(\sqrt{n})$, according to Proposition A.1,

$$P_0 = P\Big(\max_i \|\mathbf{S}^{(1)}_i\| < \mathcal{P}'_\lambda(0_+)\Big) \longrightarrow P\Big(\max_s \|\mathbf{Y}_s - s\mathbf{Y}_1\| < n^{-1/2}\mathcal{P}'_\lambda(0+)\Big).$$

If $\lambda = O_p(\sqrt{n})$ is chosen, P_0 lies in (0, 1). For the LASSO and SCAD penalties with $n^{-1/2}\lambda \to \infty$, and the modified unbounded and modified bridge penalties with $n^{-1/2}\lambda^* \to \infty$, the penalty term becomes $n^{-1/2}\mathcal{P}'_\lambda(0+) \to \infty$, and therefore, $P_0 \to 1$. For the bridge and the unbounded penalty with $\lambda > 0$, we have $n^{-1/2}\mathcal{P}'_\lambda(0+) = \infty$, and thus, $P_0 \to 1$. □

Lemma B.6. *For $k > 0$, we have*

(i) *For the LASSO penalty, if $\lambda = O(n^\alpha)$ and $1/2 \le \alpha < 1$, $P_k \to 0$.*
(ii) *For the SCAD penalty, if $n^{-1/2}\lambda \to \infty$ and $n^{-1}\lambda \to 0$, $P_k \to 1$.*
(iii) *For the unbounded (Bridge) penalty, if $\lambda > 0$ and $\lambda/n \to 0$, $P_k \to 1$.*
(iv) *For the modified unbounded (Bridge) penalty, if $\lambda > 0$, $n^{-1/2}\lambda^* \to \infty$, $n^{-1}\lambda^* \to 0$, and $\lambda \ll \lambda^*$, then $P_k \to 1$.*

Proof. Let Ω be the event that condition (i) in Proposition A.2 holds. From Proposition A.2, it suffices to establish the asymptotic properties of the followings $t^{(\ell)} = [nq^{(\ell)}]$, $\ell = 1, 2, \dots, k$: (a) $P(\Omega)$, (b) $P(\max_i \|\mathbf{S}^{(\ell)}_i\| < \mathcal{P}'_\lambda(0_+))$, for $\ell = 1, 2, \dots, k+1$. Here, Ω is defined in Proposition A.2.

Result (a): Let $\boldsymbol{\vartheta}_0 = (\boldsymbol{\theta}^{(1)}_0, \dots, \boldsymbol{\theta}^{(k+1)}_0, \boldsymbol{\phi}_0)$ and $\hat{\boldsymbol{\vartheta}} = (\hat{\boldsymbol{\theta}}^{(1)}, \dots, \hat{\boldsymbol{\theta}}^{(k+1)}, \hat{\boldsymbol{\phi}})$. Let ∇ and $\mathbf{H} = -\nabla^2$ be the gradient and Hessian with respect to $\boldsymbol{\vartheta}$. Suppose that $M > 0$ and $\varpi < 0$ are two constants such that $n^\varpi \gg \max\{\lambda/n, n^{-1/2}\}$. Define $\mathcal{B}$ as the close ball centered at $\boldsymbol{\vartheta}_0$ with radius Mn^ϖ. Below, we show that with probability going to one there must be at least a local solution to $Q^*_\lambda(\boldsymbol{\vartheta})$ within $\mathcal{B}$. Then, $P(\Omega) \to 1$ follows immediately. The arguments are similar to that in [8]. Since the function Q^*_λ is continuous in the compact set $\mathcal{B}$, a minimum exists. What remains is to show that such minimum cannot be attained on the boundary of the ball $\partial\mathcal{B}$. Suppose by contradiction that the minimum fulfills $\|\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}\| = Mn^\varpi$. Then,

$$Q^*_\lambda(\hat{\boldsymbol{\vartheta}}) - Q^*_\lambda(\boldsymbol{\vartheta}_0) \approx -(\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0)^T \sum_{\ell=1}^{k+1} \sum_{i=nq^{(\ell-1)}}^{[nq^{(\ell)}]-1} \nabla \log f\big(X_i; \boldsymbol{\theta}^{(\ell)}_0, \boldsymbol{\phi}_0\big) + (\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0)^T \mathbf{H}(\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0)$$

$$+ \sum_{\ell=1}^{k} (\hat{\boldsymbol{\xi}}^{(\ell)} - \boldsymbol{\xi}^{(\ell)}_0)^T \mathcal{P}'_\lambda\big(\|\boldsymbol{\xi}^{(\ell)}_0\|\big) = Q_1 + Q_2 + Q_3.$$

Here, $Q_2 = O_p(n^{1+2\varpi})$ is always positive. Note that $Q_1 \ll O_p(n^{1/2+\varpi})$. Since n^{ϖ} is chosen such that $n^{\varpi} \geq n^{-1/2}$, $Q_1 \ll Q_2$ holds. For SCAD, $Q_3 = 0$ while for LASSO, bridge, unbounded, modified unbounded, and modified bridge, $Q_3 \leq O_p(\lambda n^{\varpi})$. It can be seen that $Q_3 \ll Q_2$ as $n^{\varpi} \gg \lambda/n$. That $\hat{\boldsymbol{\vartheta}}$ cannot be a minimum follows immediately from $Q_1 \ll Q_2$ and $Q_3 \ll Q_2$.

Results (b) is obvious for the bridge and unbounded penalties as $\mathcal{P}'_\lambda(0_+) = \infty$. The proofs for the LASSO, SCAD, modified unbounded, and modified bridge penalties are given below. The consistency result of (a) allows Taylor series approximation of Q^*_λ. Then,

$$\begin{aligned}
&\mathbf{H}^{(1)}_{\boldsymbol{\theta\theta}}\big[\hat{\boldsymbol{\theta}}^{(1)} - \boldsymbol{\theta}^{(1)}_0\big] + \mathbf{H}^{(1)}_{\boldsymbol{\phi\theta}}[\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0] \\
&\quad\approx \frac{1}{n^{(1)}}\Bigg\{-\mathcal{P}'_\lambda\big(\|\boldsymbol{\xi}^{(1)}_0\|\big)\cdot \mathbf{j}^{(1)}_0 + \sum_{i=1}^{[nq^{(1)}]-1} \nabla_{\boldsymbol{\theta}} \log f\big(X_i;\boldsymbol{\theta}^{(1)}_0,\boldsymbol{\phi}_0\big)\Bigg\}, \\
&\mathbf{H}^{(\ell)}_{\boldsymbol{\theta\theta}}\big[\hat{\boldsymbol{\theta}}^{(\ell)} - \boldsymbol{\theta}^{(\ell)}_0\big] + \mathbf{H}^{(\ell)}_{\boldsymbol{\phi\theta}}[\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0] \\
&\quad\approx \frac{1}{n^{(\ell)}}\Bigg\{-\big[\mathcal{P}'_\lambda\big(\|\boldsymbol{\xi}^{(\ell)}_0\|\big)\cdot \mathbf{j}^{(\ell)}_0 - \mathcal{P}'_\lambda\big(\|\boldsymbol{\xi}^{(\ell-1)}_0\|\big)\cdot \mathbf{j}^{(\ell-1)}_0\big] \qquad\qquad (B.4)\\
&\qquad + \sum_{i=[nq^{(\ell-1)}]}^{[nq^{(\ell)}]-1} \nabla_{\boldsymbol{\theta}} \log f\big(X_i;\boldsymbol{\theta}^{(1)}_0,\boldsymbol{\phi}_0\big)\Bigg\}, \qquad \ell = 2,3,\ldots,k, \\
&\mathbf{H}^{(k+1)}_{\boldsymbol{\theta\theta}}\big[\hat{\boldsymbol{\theta}}^{(k+1)} - \boldsymbol{\theta}^{(k+1)}_0\big] + \mathbf{H}^{(k+1)}_{\boldsymbol{\phi\theta}}[\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0] \\
&\quad\approx \frac{1}{n^{(k+1)}}\Bigg\{+\mathcal{P}'_\lambda\big(\|\boldsymbol{\xi}^{(k)}_0\|\big)\cdot \mathbf{j}^{(k)}_0 + \sum_{i=[nq^{(k)}]}^{n} \nabla_{\boldsymbol{\theta}} \log f\big(X_i;\boldsymbol{\theta}^{(k+1)}_0,\boldsymbol{\phi}_0\big)\Bigg\}.
\end{aligned}$$

Using Taylor expansion and equation (B.4),

$$\begin{aligned}
\frac{1}{\sqrt{n}}\mathbf{S}^{(1)}_i &= -\frac{1}{\sqrt{n}}\sum_{j=1}^{i} \nabla_{\boldsymbol{\theta}} \log f\big(X_j;\hat{\boldsymbol{\theta}}^{(1)},\hat{\boldsymbol{\phi}}\big) \\
&\approx -\frac{1}{\sqrt{n}}\sum_{j=1}^{i} \nabla_{\boldsymbol{\theta}} \log f\big(X_j;\boldsymbol{\theta}^{(1)}_0,\boldsymbol{\phi}_0\big) - \frac{1}{\sqrt{n}}\Bigg(\sum_{j=1}^{i} \nabla^2_{\boldsymbol{\theta\theta}} \log f\big(X_j;\boldsymbol{\theta}^{(1)}_0,\boldsymbol{\phi}_0\big)\Bigg)\big(\hat{\boldsymbol{\theta}}^{(1)} - \boldsymbol{\theta}^{(1)}_0\big) \\
&\quad - \frac{1}{\sqrt{n}}\Bigg(\sum_{j=1}^{i} \nabla^2_{\boldsymbol{\theta\phi}} \log f\big(X_j;\boldsymbol{\theta}^{(1)}_0,\boldsymbol{\phi}_0\big)\Bigg)(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0) \\
&\approx -\frac{1}{\sqrt{n}}\sum_{j=1}^{i} \nabla_{\boldsymbol{\theta}} \log f\big(X_j;\boldsymbol{\theta}^{(1)}_0,\boldsymbol{\phi}_0\big) + \frac{i}{\sqrt{n}}\mathbf{H}^{(1)}_{\boldsymbol{\theta\theta}}\big(\hat{\boldsymbol{\theta}}^{(1)} - \boldsymbol{\theta}^{(1)}_0\big) + \frac{i}{\sqrt{n}}\mathbf{H}^{(1)}_{\boldsymbol{\theta\phi}}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)
\end{aligned}$$

$$\approx -\frac{1}{\sqrt{n}}\sum_{j=1}^{i}\nabla_{\boldsymbol{\theta}}\log f\big(X_j;\boldsymbol{\theta}_0^{(1)},\boldsymbol{\phi}_0\big)+\frac{i}{n^{(1)}\sqrt{n}}\sum_{j=1}^{t^{(1)}}\nabla_{\boldsymbol{\theta}}\log f\big(X_j;\boldsymbol{\theta}_0^{(1)},\boldsymbol{\phi}_0\big)$$
$$-\frac{i}{n^{(1)}\sqrt{n}}\mathcal{P}'_\lambda\big(\|\boldsymbol{\xi}_0^{(1)}\|\big)\mathbf{j}_0^{(1)}.$$

Let $s=i/n$. For $0<s<q^{(1)}$, this stochastic process converges weakly to a drifted Brownian bridge

$$W_s=-\mathbf{Y}_s+\frac{s}{q^{(1)}}\mathbf{Y}_{q^{(1)}}-\frac{s}{q^{(1)}}\Big[\lim_{n\to\infty}n^{-1/2}\mathcal{P}'_\lambda\big(\|\boldsymbol{\xi}_0^{(1)}\|\big)\mathbf{j}_0^{(1)}\Big].$$

Here, $\mathbf{Y}_s$ is the Brownian motion with volatility $\mathbf{H}_{\boldsymbol{\theta}\boldsymbol{\theta}}$. Stochastic processes $\mathbf{S}_i^{(2)},\dots,\mathbf{S}_i^{(k+1)}$ can be handled in the same manner. Then, for $s=(t^{(k)}+i)/n$, such that $q^{(k')}<s<1$, $n^{-1/2}\mathbf{S}_i^{(k+1)}$ converges to

$$W_s=-\mathbf{Y}_s+\frac{1-s}{1-q^{(k)}}\mathbf{Y}_{q^{(k)}}+\frac{s-q^{(k)}}{1-q^{(k)}}\mathbf{Y}_1-\Big[\lim_{n\to\infty}n^{-1/2}\mathcal{P}'_\lambda\big(\|\boldsymbol{\xi}_0^{(k)}\|\big)\Big]\mathbf{j}_0^{(1)}$$
$$+\frac{s-q^{(k)}}{1-q^{(k)}}\big[n^{-1/2}\mathcal{P}'_\lambda\big(\|\boldsymbol{\xi}_0^{(k)}\|\big)\mathbf{j}_0^{(k)}\big].$$

For $\ell=2,3,\dots,k$, $s=(t^{(\ell)}+i)/n$, such that $q^{(\ell-1)}<s<q^{(\ell)}$, $n^{-1/2}\mathbf{S}_i^{(\ell)}$ converges to

$$W_s=-\mathbf{Y}_s+\frac{q^{\ell}-s}{q^{(\ell)}-q^{(\ell-1)}}\mathbf{Y}_{q^{(\ell-1)}}+\frac{s-q^{\ell-1}}{q^{(\ell)}-q^{(\ell-1)}}\mathbf{Y}_{q^{(\ell)}}-\Big[\lim_{n\to\infty}n^{-1/2}\mathcal{P}'_\lambda\big(\|\boldsymbol{\xi}_0^{(\ell-1)}\|\big)\Big]\mathbf{j}_0^{(\ell-1)}$$
$$-\frac{s-q^{(\ell-1)}}{q^{(\ell)}-q^{(\ell-1)}}\Big[\lim_{n\to\infty}n^{-1/2}\mathcal{P}'_\lambda\big(\|\boldsymbol{\xi}_0^{(\ell)}\|\big)\mathbf{j}_0^{(\ell)}-\lim_{n\to\infty}n^{-1/2}\mathcal{P}'_\lambda\big(\|\boldsymbol{\xi}_0^{(\ell-1)}\|\big)\mathbf{j}_0^{(\ell-1)}\Big].$$

It can be seen that at both $s=0$ and $s=1$, $W_s=0$ and at $s=q^{(\ell)}$, $\ell=1,2,\dots,k$, $\|W_s\|=\lim_{n\to\infty}n^{-1/2}\mathcal{P}'_\lambda(\|\boldsymbol{\xi}_0^{(\ell)}\|)$.

It is interesting to note that for the LASSO penalty, $\mathcal{P}'_\lambda(0_+)=\mathcal{P}'_\lambda(\|\boldsymbol{\xi}_0^{(\ell)}\|)$. On the contrary, this does not hold for all other penalties. Due to the roughness and unbounded variation of the Brownian motion, for the LASSO penalty, $P(\max_i\|\mathbf{S}_i^{(1)}\|<\mathcal{P}'_\lambda(0_+))\to 0$. Therefore, $P_k\to 0$.

For the modified unbounded and the modified bridge penalties with $n^{-1/2}\lambda^*\to\infty$ and $\lambda\ll\lambda^*$, $\|W_{q^{(\ell)}}\|=O_p(n^{-1/2}\lambda)\ll O_p(n^{-1/2}\lambda^*)$. Then, $P(\max_i\|\mathbf{S}_i^{(\ell)}\|<\mathcal{P}'_\lambda(0_+))\approx P(\max_s\|W_s\|<n^{-1/2}\lambda^*)\to 1$.

For the SCAD penalty, $\mathcal{P}'_\lambda(\|\boldsymbol{\xi}_0^{(\ell)}\|)=0$ for all $\ell=1,2,\dots,k$ with probability going to one. Then, $\|W_{q^{(\ell)}}\|=0$ and in each regime, W_s is Brownian bridge. Therefore, $P(\max_i\|\mathbf{S}_i^{(1)}\|<\mathcal{P}'_\lambda(0_+))\approx P(\max_s\|W_s\|<n^{-1/2}\lambda)\to 1$ if $n^{-1/2}\lambda\to\infty$. □

Lemma B.7. *Suppose that the modified unbounded penalty satisfies conditions MUB1–MUB3. Consider a restricted local solution $\hat{\boldsymbol{\theta}}$ in $\aleph$. Let $\ell_1,\ell_1+1,\ell_1+2,\dots,\ell_2$ be subsequent regimes* *in $\hat{\boldsymbol{\theta}}$. Suppose that $\ell_2 \neq \ell_1$, regime ℓ_1 is either Type N or R, and regime ℓ_2 is either Type N or L. Let ℓ^* be the regime with the shortest length among $\ell_1, \ell_1+1, \ell_1+2, \dots, \ell_2$. Then, the followings cannot be satisfied simultaneously*: (i) *all regimes other than ℓ_1 and ℓ_2 are Type N* (*see Convention* B.3), (ii) $\bar{\boldsymbol{\theta}}^{(\ell_1)} - \bar{\boldsymbol{\theta}}^{(\ell_1+1)} = o[(\lambda/n^{(\ell^*)})^{1/2}]$ *and* $\bar{\boldsymbol{\theta}}^{(\ell_2-1)} - \bar{\boldsymbol{\theta}}^{(\ell_2)} = o[(\lambda/n^{(\ell^*)})^{1/2}]$, (iii) $\ell_1 = 1$ *or* $\bar{\boldsymbol{\theta}}^{(\ell_1-1)} - \bar{\boldsymbol{\theta}}^{(\ell_1)} = O_p(1)$, *and* (iv) $\ell_2 = k'+1$ *or* $\bar{\boldsymbol{\theta}}^{(\ell_2)} - \bar{\boldsymbol{\theta}}^{(\ell_2+1)} = O_p(1)$.

Lemma B.8. *Suppose that the modified bridge penalty satisfies condition MBR*1*–MBR*4. *Consider a restricted local solution $\hat{\boldsymbol{\theta}}$ in $\aleph$. Let $\ell_1, \ell_1+1, \ell_1+2, \dots, \ell_2$ be subsequent regimes in $\hat{\boldsymbol{\theta}}$. Suppose that $\ell_2 \neq \ell_1$, regime ℓ_1 is either Type N or R, and regime ℓ_2 is either Type N or L. Let ℓ^* be the regime with the shortest length among $\ell_1, \ell_1+1, \ell_1+2, \dots, \ell_2$. Then, the followings cannot be satisfied simultaneously*: (i) *all regimes other than ℓ_1 and ℓ_2 are Type N* (*see Convention* B.3), (ii) $\bar{\boldsymbol{\theta}}^{(\ell_1)} - \bar{\boldsymbol{\theta}}^{(\ell_1+1)} = o[(\lambda/n^{(\ell^*)})^{1/(2-\gamma)}]$ *and* $\bar{\boldsymbol{\theta}}^{(\ell_2-1)} - \bar{\boldsymbol{\theta}}^{(\ell_2)} = o[(\lambda/n^{(\ell^*)})^{1/(2-\gamma)}]$, (iii) $\ell_1 = 1$ *or* $\bar{\boldsymbol{\theta}}^{(\ell_1-1)} - \bar{\boldsymbol{\theta}}^{(\ell_1)} = O_p(1)$, *and* (iv) $\ell_2 = k'+1$ *or* $\bar{\boldsymbol{\theta}}^{(\ell_2)} - \bar{\boldsymbol{\theta}}^{(\ell_2+1)} = O_p(1)$.

Proofs of Lemmas B.7 and B.8. Here, the proof is given for the cases where regime ℓ_1 is Type R and regime ℓ_2 is Type L. Other cases can be handled in a similar manner. Suppose that (i)-(iv) are all satisfied. Then, both $\bar{\mathbf{H}}^{(\ell)}$ and $\bar{\boldsymbol{\theta}}^{(\ell)}$, $\ell = \ell_1, \dots, \ell_2$ are all asymptotically the same. Let $\boldsymbol{\theta}_0^*$ be the true parameter value for regimes $\ell = \ell_1+1, \dots, \ell_2-1$ and the dominating portions of regime ℓ_1 and regime ℓ_2. Define $\mathbf{H}^*_{\boldsymbol{\theta}\boldsymbol{\theta}} = -E\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}} \log f(X; \boldsymbol{\theta}_0^*, \boldsymbol{\phi}_0)$. Then, from Lemma B.2, we have

$$
\begin{pmatrix}
\mathbf{H}^*_{\boldsymbol{\theta}\boldsymbol{\theta}} \hat{\boldsymbol{\xi}}^{(\ell_1)} \\
\mathbf{H}^*_{\boldsymbol{\theta}\boldsymbol{\theta}} \hat{\boldsymbol{\xi}}^{(\ell_1+1)} \\
\vdots \\
\mathbf{H}^*_{\boldsymbol{\theta}\boldsymbol{\theta}} \hat{\boldsymbol{\xi}}^{(\ell_2-1)}
\end{pmatrix}
\approx
\begin{pmatrix}
\frac{1}{n^{(\ell_1)}} L_{\boldsymbol{\theta}}^{(\ell_1)} - \frac{1}{n^{(\ell_1+1)}} L_{\boldsymbol{\theta}}^{(\ell_1+1)} \\
\frac{1}{n^{(\ell_1+1)}} L_{\boldsymbol{\theta}}^{(\ell_1+1)} - \frac{1}{n^{(\ell_1+2)}} L_{\boldsymbol{\theta}}^{(\ell_1+2)} \\
\vdots \\
\frac{1}{n^{(\ell_2-1)}} L_{\boldsymbol{\theta}}^{(\ell_2-1)} - \frac{1}{n^{(\ell_2)}} L_{\boldsymbol{\theta}}^{(\ell_2)}
\end{pmatrix}
+
\begin{pmatrix}
[n^{(\ell_1)}]^{-1} \mathcal{P}'_\lambda(\|\hat{\boldsymbol{\xi}}^{(\ell_1-1)}\|) \hat{\mathbf{j}}^{(\ell_1-1)} + \mathbf{H}^*_{\boldsymbol{\theta}\boldsymbol{\theta}}(\bar{\boldsymbol{\theta}}^{(\ell_1)} - \bar{\boldsymbol{\theta}}^{(\ell_1+1)}) \\
0 \\
\vdots \\
[n^{(\ell_2)}]^{-1} \mathcal{P}'_\lambda(\|\hat{\boldsymbol{\xi}}^{(\ell_2)}\|) \hat{\mathbf{j}}^{(\ell_2)} + \mathbf{H}^*_{\boldsymbol{\theta}\boldsymbol{\theta}}(\bar{\boldsymbol{\theta}}^{(\ell_2-1)} - \bar{\boldsymbol{\theta}}^{(\ell_2)})
\end{pmatrix}
\quad \text{(B.5)}
$$

$$
-\begin{pmatrix}
\frac{\mathbf{I}_{p\times p}}{n^{(\ell_1)}}+\frac{\mathbf{I}_{p\times p}}{n^{(\ell_1+1)}} & -\frac{\mathbf{I}_{p\times p}}{n^{(\ell_1+1)}} & & & \\
-\frac{\mathbf{I}_{p\times p}}{n^{(\ell_1+1)}} & \frac{\mathbf{I}_{p\times p}}{n^{(\ell_1+1)}}+\frac{\mathbf{I}_{p\times p}}{n^{(\ell_1+2)}} & -\frac{\mathbf{I}_{p\times p}}{n^{(\ell_1+2)}} & & \\
 & \ddots & \ddots & \ddots & \\
 & & -\frac{\mathbf{I}_{p\times p}}{n^{(\ell_2-1)}} & \frac{\mathbf{I}_{p\times p}}{n^{(\ell_2-1)}}+\frac{\mathbf{I}_{p\times p}}{n^{(\ell_2)}} &
\end{pmatrix}
\times\begin{pmatrix}
\mathcal{P}'_\lambda(\|\hat{\boldsymbol{\xi}}^{(\ell_1)}\|)\hat{\mathbf{j}}^{(\ell_1)} \\
\mathcal{P}'_\lambda(\|\hat{\boldsymbol{\xi}}^{(\ell_1+1)}\|)\hat{\mathbf{j}}^{(\ell_1+1)} \\
\vdots \\
\mathcal{P}'_\lambda(\|\hat{\boldsymbol{\xi}}^{(\ell_2-1)}\|)\hat{\mathbf{j}}^{(\ell_2-1)}
\end{pmatrix}.
$$

Let $b^{(\ell)}=(\hat{\mathbf{j}}^{(\ell)})^T\mathbf{H}_{\boldsymbol{\theta}\boldsymbol{\theta}}\hat{\mathbf{j}}^{(\ell)}$, $\ell=\ell_1,\dots,\ell_2-1$, and

$$
\begin{aligned}
A^{(\ell_1)}&=\left[\left(\frac{1}{n^{(\ell_1)}}+\frac{1}{n^{(\ell_1+1)}}\right)-\frac{1}{n^{(\ell_1+1)}}\hat{\mathbf{j}}^{(\ell_1+1)}\cdot\hat{\mathbf{j}}^{(\ell_1)}\right],\\
A^{(\ell_1+1)}&=\left[-\frac{1}{n^{(\ell_1+1)}}\hat{\mathbf{j}}^{(\ell_1)}\cdot\hat{\mathbf{j}}^{(\ell_1+1)}+\left(\frac{1}{n^{(\ell_1+1)}}+\frac{1}{n^{(\ell_1+2)}}\right)-\frac{1}{n^{(\ell_1+2)}}\hat{\mathbf{j}}^{(\ell_1+2)}\cdot\hat{\mathbf{j}}^{(\ell_1+1)}\right],\\
\vdots&=\vdots\\
A^{(\ell_2-2)}&=\left[-\frac{1}{n^{(\ell_2-2)}}\hat{\mathbf{j}}^{(\ell_2-3)}\cdot\hat{\mathbf{j}}^{(\ell_2-2)}+\left(\frac{1}{n^{(\ell_2-2)}}+\frac{1}{n^{(\ell_2-1)}}\right)-\frac{1}{n^{(\ell_2-1)}}\hat{\mathbf{j}}^{(\ell_2-1)}\cdot\hat{\mathbf{j}}^{(\ell_2-2)}\right],\\
A^{(\ell_2-1)}&=\left[-\frac{1}{n^{(\ell_2-1)}}\hat{\mathbf{j}}^{(\ell_2-2)}\cdot\hat{\mathbf{j}}^{(\ell_2-1)}+\left(\frac{1}{n^{(\ell_2-1)}}+\frac{1}{n^{(\ell_2)}}\right)\right].
\end{aligned}
$$

Multiplying both sides of equation (B.5) by $((\hat{\mathbf{j}}^{(\ell_1)})^T,(\hat{\mathbf{j}}^{(\ell_1+1)})^T,\dots,(\hat{\mathbf{j}}^{\ell_2-1})^T)$, we have

$$
\begin{aligned}
&\big(b^{(\ell_1)}\|\hat{\boldsymbol{\xi}}^{(\ell_1)}\|+A^{(\ell_1)}\mathcal{P}'_\lambda(\|\hat{\boldsymbol{\xi}}^{(\ell_1)}\|)\big)+\cdots+\big(b^{(\ell_2-1)}\|\hat{\boldsymbol{\xi}}^{(\ell_2-1)}\|+A^{(\ell_2-1)}\mathcal{P}'_\lambda(\|\hat{\boldsymbol{\xi}}^{(\ell_2-1)}\|)\big)\\
&\quad\approx-\frac{1}{n^{(\ell_1)}}\big(\hat{\mathbf{j}}^{(\ell_1)}\big)^TL_{\boldsymbol{\theta}}^{(\ell_1)}-\frac{1}{n^{(\ell_1+1)}}\big(\hat{\mathbf{j}}^{(\ell_1+1)}-\hat{\mathbf{j}}^{(\ell_1)}\big)^TL_{\boldsymbol{\theta}}^{(\ell_1+1)}-\cdots\\
&\qquad-\frac{1}{n^{(\ell_2-1)}}\big(\hat{\mathbf{j}}^{(\ell_2-1)}-\hat{\mathbf{j}}^{(\ell_2-2)}\big)^TL_{\boldsymbol{\theta}}^{(\ell_2-1)}+\frac{1}{n^{(\ell_2)}}\big(\hat{\mathbf{j}}^{(\ell_2)}\big)^TL_{\boldsymbol{\theta}}^{(\ell_2)}\\
&\qquad+\frac{1}{n^{(\ell_1)}}\mathcal{P}'_\lambda(\|\hat{\boldsymbol{\xi}}^{(\ell_1-1)}\|)\hat{\mathbf{j}}^{(\ell_1-1)}\cdot\hat{\mathbf{j}}^{(\ell_1)}+\frac{1}{n^{(\ell_2)}}\mathcal{P}'_\lambda(\|\hat{\boldsymbol{\xi}}^{(\ell_2)}\|)\hat{\mathbf{j}}^{(\ell_2)}\cdot\hat{\mathbf{j}}^{(\ell_2-1)}\\
&\qquad+\big(\hat{\mathbf{j}}^{(\ell_1)}\big)^T\mathbf{H}^*_{\boldsymbol{\theta}\boldsymbol{\theta}}\big(\bar{\boldsymbol{\theta}}^{(\ell_1)}-\bar{\boldsymbol{\theta}}^{(\ell_1+1)}\big)+\big(\hat{\mathbf{j}}^{(\ell_2-1)}\big)^T\mathbf{H}^*_{\boldsymbol{\theta}\boldsymbol{\theta}}\big(\bar{\boldsymbol{\theta}}^{(\ell_2-1)}-\bar{\boldsymbol{\theta}}^{(\ell_2)}\big)\\
&\quad=I^{(\ell_1)}+\cdots+I^{(\ell_2)}+T_1+T_2+U_1+U_2.
\end{aligned}
\tag{B.6}
$$

In what follows, we show that the left-hand side must dominate the right-hand side. Then, this is a contradiction as the equality sign is impossible. A lower bound for the left-hand side is given, then the terms $I^{(\ell_1)}, \dots, I^{(\ell_2)}, T_1, T_2, U_1, U_2$ are compared to such lower bound.

Modified unbounded penalty: Consider the approximation in Convention B.4. To give a lower bound for the left-hand side of equation (B.6), the following two results are used. (a) Let A and b be two positive constants and x be a positive variable. The minimum value of $bx + Ax^{-1}$ is $2A^{1/2}b^{1/2}$. (b) For $A_1, A_2 > 0$, the following inequality holds:

$$(A_1 + A_2)^{1/2} > \big[A_1^{1/2} + A_2^{1/2}\big]/2.$$

After using the above two results, we see that the left-hand side of equation (B.6) is bounded below by

$$\begin{aligned}&\big(n^\alpha \tau^{-1}(\tau-2)\big)^{1/2}\big\{\big(b^{(\ell_1)}/n^{(\ell_1)}\big)^{1/2} + 2\big(b^{(\ell_1+1)}\big(1-\hat{\mathbf{j}}^{(\ell_1+1)}\cdot\hat{\mathbf{j}}^{(\ell_1)}\big)/n^{(\ell_1+1)}\big)^{1/2}\\ &\quad + \dots + 2\big(b^{(\ell_2-2)}\big(1-\hat{\mathbf{j}}^{(\ell_2-2)}\cdot\hat{\mathbf{j}}^{(\ell_2-1)}\big)/n^{(\ell_2-2)}\big)^{1/2} + \big(b^{(\ell_2-1)}/n^{(\ell_2-1)}\big)^{1/2}\big\}.\end{aligned} \tag{B.7}$$

Terms T_1 and T_2: From conditions (i) and (iii), we have

$$\|\hat{\boldsymbol{\xi}}^{(\ell_1-1)}\| \geq \|\bar{\boldsymbol{\theta}}^{(\ell_1-1)} - \bar{\boldsymbol{\theta}}^{(\ell_1)}\| - \|\hat{\boldsymbol{\theta}}^{(\ell_1-1)} - \bar{\boldsymbol{\theta}}^{(\ell_1-1)}\| - \|\hat{\boldsymbol{\theta}}^{(\ell_1)} - \bar{\boldsymbol{\theta}}^{(\ell_1)}\| = O_p(1).$$

Here, Lemma B.1 is used. T_1 is dominated by the first term in (B.7), that is an $n^{(\ell_1)}$ relating term in (B.7) $(n^\alpha/n^{(\ell_1)})^{1/2}$ if ignoring the $O(1)$ multiple. Therefore, T_1 can be ignored. Similarly, T_2 is also ignorable.

Terms U_1 and U_2: Condition (ii) guarantees that U_1 is dominated by the $n^{(\ell_1)}$ relating term in (B.7), which is $(n^\alpha/n^{(\ell_1)})^{1/2}$. We can similarly show for U_2.

$I^{(\ell_1)}, \dots, I^{(\ell_2)}$: Let $\varepsilon > 0$ be arbitrarily small. Note that the $n^{(\ell_1)}$ relating term in (B.7) has the same order as $(n^\alpha/n^{(\ell_1)})^{1/2}$. It is positive and dominates the term $I^{(\ell_1)}$ because according to condition R2, $I^{(\ell_1)} = o_p(n^\varepsilon/\sqrt{n^{(\ell_1)}})$ and such bound does not depend on the locations of the change points of $\hat{\boldsymbol{\theta}}$. For the $n^{(\ell_1+1)}$ relating terms, it has the same order as

$$\left(\frac{n^\alpha}{n^{(\ell_1+1)}}\right)^{1/2}\big[1-\hat{\mathbf{j}}^{(\ell_1)}\cdot\hat{\mathbf{j}}^{(\ell_1+1)}\big]^{1/2}.$$

This term can be compared to $I^{(\ell_1+1)}$. Note that $1-\hat{\mathbf{j}}^{(\ell_1)}\cdot\hat{\mathbf{j}}^{(\ell_1+1)} = \|\hat{\mathbf{j}}^{(\ell_1)} - \hat{\mathbf{j}}^{(\ell_1+1)}\|^2/2$. Then, $I^{(\ell_1+1)}$ is dominated by bound (B.7).

We see that all terms $I^{(\ell_1)}, \dots, I^{(\ell_2)}, T_1, T_2, U_1$, and U_2 are dominated by some corresponding terms in bound (B.7). Therefore, the equality sign of equation (B.6) is impossible with probability going to one.

Modified bridge penalty: Using similar arguments as in the unbounded penalty cases, a lower bound for the left-hand side of equation (B.6) is

$$\begin{aligned}
&(n^\alpha\gamma)^{1/(2-\gamma)}\big[(1-\gamma)^{1/(2-\gamma)}+(1-\gamma)^{-(1-\gamma)/(2-\gamma)}\big]\\
&\quad\times\Big\{\big[b^{(\ell_1)}\big]^{(1-\gamma)/(2-\gamma)}\big(1/n^{(\ell_1)}\big)^{1/(2-\gamma)}\\
&\quad+2\big[b^{(\ell_1+1)}\big]^{(1-\gamma)/(2-\gamma)}\big[\big(1-\hat{\mathbf{j}}^{(\ell_1+1)}\cdot\hat{\mathbf{j}}^{(\ell_1)}\big)/n^{(\ell_1+1)}\big]^{1/(2-\gamma)}+\cdots\\
&\quad+2\big[b^{(\ell_2-2)}\big]^{(1-\gamma)/(2-\gamma)}\big[\big(1-\hat{\mathbf{j}}^{(\ell_2-2)}\cdot\hat{\mathbf{j}}^{(\ell_2-1)}\big)/n^{(\ell_2-2)}\big]^{1/(2-\gamma)}\\
&\quad+\big[b^{(\ell_2-1)}\big]^{(1-\gamma)/(2-\gamma)}\big(1/n^{(\ell_2-1)}\big)^{1/(2-\gamma)}\Big\}\\
&=V^{(\ell_1)}+\cdots+V^{(\ell_2-1)}.
\end{aligned} \tag{B.8}$$

Terms T_1, T_2, U_1, and U_2 can be handled in a similar manner as in the proof of Lemma B.7.

Terms $I^{(\ell_1)},\dots,I^{(\ell_2)}$: The $n^{(\ell_1)}$ relating term in (B.8), that is, $V^{(\ell_1)}$, has the same order as $(n^\alpha/n^{(\ell_1)})^{1/(2-\gamma)}$ and is positive. Let $\varepsilon>0$ be arbitrarily small. Under condition MBR1 and R2, it dominates term $I^{(\ell_1)}$ because $I^{(\ell_1)}=o_p(n^\varepsilon/\sqrt{n^{(\ell_1)}})$ and such bound does not depend on the locations of the change points of $\hat{\boldsymbol{\theta}}$. For the $n^{(\ell_1+1)}$ relating terms, it has the same order as

$$\left(\frac{n^\alpha}{n^{(\ell_1+1)}}\right)^{1/(2-\gamma)}\|\hat{\mathbf{j}}^{(\ell_1)}-\hat{\mathbf{j}}^{(\ell_1+1)}\|^{2/(2-\gamma)}.$$

In what follows, we show that the above mentioned quantity V dominates $I^{(\ell_1+1)}$.

For univariate cases, $\hat{\mathbf{j}}$ only takes two possible values: $+1$ or -1. If $\hat{\mathbf{j}}^{(\ell_1)}=\hat{\mathbf{j}}^{(\ell_1+1)}$, both V and $I^{(\ell+1)}$ are zero. If $\hat{\mathbf{j}}^{(\ell_1)}\neq\hat{\mathbf{j}}^{(\ell_1+1)}$, quantity V, being positive, dominates $I^{(\ell+1)}$ under condition MBR1.

For $\dim(\Theta)>1$, condition MBR4 is required. Consider two types of regimes (excluding regime ℓ_1 and regime ℓ_2). Type I: $\|\hat{\mathbf{j}}^{(\ell-1)}-\hat{\mathbf{j}}^{(\ell)}\|^2>n^{(\ell)}n^{-2\alpha/\gamma}$. Type II: $\|\hat{\mathbf{j}}^{(\ell-1)}-\hat{\mathbf{j}}^{(\ell)}\|^2\le n^{(\ell)}n^{-2\alpha/\gamma}$. For type I regimes, V, being positive, dominates $I^{(\ell+1)}$ under condition MBR1. For type II regimes, $\|I^{(\ell+1)}\|$ is bounded by $n^{-\alpha/\gamma}\|L_{\boldsymbol{\theta}}^{(\ell)}/\sqrt{n^{(\ell)}}\|=O_p(n^{-\alpha/\gamma})$. From MBR2, the maximum number of regimes is $n^{1-\kappa}$. As a result, the standard deviation of the sum of $I^{(\ell)}$ for ℓ belonging to Type II regimes is only $n^{-\alpha/\gamma+(1-\kappa)/2}$. Term $V^{(\ell_1)}$, where $\hat{\mathbf{j}}$ does not appear, is $O[(\lambda/n^{(\ell_1)})^{1/(2-\gamma)}]$ is bounded below by $O_p(n^{-(1-\alpha)/(2-\gamma)})$. Condition MBR4 guarantees that $n^{-(1-\alpha)/(2-\gamma)}$ is always dominated by $V^{(\ell_1)}$. □

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2011-0030810), the 2013 Chonnam National University Research Program grant (No. 2013-2299) and the Brain Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (2014M3C7A1062896).

References

[1] Bai, J., Lumsdaine, R.L. and Stock, J.H. (1998). Testing for and dating common breaks in multivariate time series. *Rev. Econ. Stud.* **65** 395–432. MR1635358

[2] Bai, J. and Perron, P. (1998). Testing for and estimation of multiple structural changes. *Econometrica* **66** 817–858.

[3] Beaulieu, C., Chen, J. and Sarmiento, L. (2012). Change-point analysis as a tool to detect abrupt climate variations. *Philos. Trans. R. Soc. A* **370** 1228–1249.

[4] Boysen, L., Kempe, A., Liebscher, V., Munk, A. and Wittich, O. (2009). Consistencies and rates of convergence of jump-penalized least squares estimators. *Ann. Statist.* **37** 157–183. MR2488348

[5] Boysen, L., Liebscher, V., Munk, A. and Wittich, O. (2007). Scale space consistency of piecewise constant least squares estimators—another look at the regressogram. In *Asymptotics: Particles, Processes and Inverse Problems. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **55** 65–84. Beachwood, OH: IMS. MR2459931

[6] Cleynen, A., Rigaill, G. and Koskas, M. (2013). Segmentor3IsBack: A Fast Segmentation Algorithm. R package version 1.8. Available at http://CRAN.R-project.org/package=Segmentor3IsBack.

[7] Csörgő, M. and Horváth, L. (1997). *Limit Theorems in Change-Point Analysis. Wiley Series in Probability and Statistics*. Wiley: Chichester. MR2743035

[8] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

[9] Frank, I. and Friedman, J. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics* **35** 109–148.

[10] Frick, K., Munk, A. and Sieling, H. (2014). Multiscale change point inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 495–580. MR3210728

[11] Fu, W.J. (1998). Penalized regressions: The bridge versus the lasso. *J. Comput. Graph. Statist.* **7** 397–416. MR1646710

[12] Gillet, O., Essid, S. and Richard, G. (2007). On the correlation of automatic audio and visual segmentation of music video. *IEEE Trans. Circuits Syst. Video Technol.* **17** 347–355.

[13] Golub, G.H. and Van Loan, C.F. (1996). *Matrix Computations*, 3rd ed. *Johns Hopkins Studies in the Mathematical Sciences*. Baltimore, MD: Johns Hopkins Univ. Press. MR1417720

[14] Harchaoui, Z. and Levy-Leduc, C. (2008). Catching change-points with lasso. *Advances in Neural Information Processing Systems (NIPS)*, 617–624.

[15] Hotz, T. and Sieling, H. (2015). stepR: Fitting Step-Functions. R package version 1.0-3. Available at http://CRAN.R-project.org/package=stepR.

[16] Huang, T., Wu, B., Lizardi, P. and Zhao, H. (2005). Detection of DNA copy number alterations using penalized least squares regression. *Bioinformatics* **21** 3811–3817.

[17] Hunter, D.R. and Li, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33** 1617–1642. MR2166557

[18] Jackson, B. *et al.* (2005). An algorithm for optional partitioning of data on an interval. *IEEE Signal Process. Lett.* **12** 105–108.

[19] Killick, R. and Eckley, I.A. (2014). changepoint: An R package for changepoint analysis. *J. Stat. Softw.* **58** 1–19.

[20] Killick, R., Fearnhead, P. and Eckley, I.A. (2012). Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.* **107** 1590–1598. MR3036418

[21] Lai, T.L. and Xing, H. (2013). *Change-Point Models of Asset Returns and Their Volatilities* (C.F. Lee, ed.). *Handbook of Financial Econometrics and Risk*. New York: Springer.

[22] Lavielle, M. and Moulines, E. (2000). Least-squares estimation of an unknown number of shifts in a time series. *J. Time Series Anal.* **21** 33–59. MR1766173

[23] Lebarbier, E. (2005). Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing* **85** 717–736.
[24] Lee, Y. and Oh, H.-S. (2014). A new sparse variable selection via random-effect model. *J. Multivariate Anal*. **125** 89–99. MR3163832
[25] Li, H., Munk, A. and Sieling, H. (2016). FDR-control in multiscale change-point segmentation. *Electron. J. Stat*. **10** 918–959. MR3486421
[26] Li, H. and Sieling, H. (2015). FDRSeg: FDR-Control in Multiscale Change-point Segmentation. R package version 1.0-1.
[27] Perron, P. (2005). Dealing with structural breaks. In *Palgrave Handbook of Econometrics*, *Vol*. 1: *Econometric Theory* (K. Patterson and T.C. Mills, eds.). Basingstoke, UK: Palgrave Macmillan.
[28] Rigaill, G. (2010). Pruned dynamic programming for optimal multiple change-point detection. Available at arXiv.org:1004.0887v1.
[29] Rinaldo, A. (2009). Properties and refinements of the fused lasso. *Ann. Statist*. **37** 2922–2952. MR2541451
[30] Snijders, A.M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A.K., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J.P., Gray, J.W., Jain, A.N., Pinkel, D. and Albertson, D.G. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat. Genet*. **29** 263–264.
[31] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242
[32] Tibshirani, R. and Wang, P. (2008). Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics* **9** 18–29.
[33] Yao, Y.-C. (1987). Approximating the distribution of the maximum likelihood estimate of the change-point in a sequence of independent random variables. *Ann. Statist*. **15** 1321–1328. MR0902262
[34] Zhang, N.R. and Siegmund, D.O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* **63** 22–32, 309. MR2345571
[35] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol*. **67** 301–320. MR2137327

Received January 2015 and revised March 2016