High-dimensional changepoint estimation via sparse projection

Tengyao Wang* and Richard J. Samworth†
University of Cambridge

(June 22, 2016)

Abstract

Changepoints are a very common feature of Big Data that arrive in the form of a data stream. In this paper, we study high-dimensional time series in which, at certain time points, the mean structure changes in a sparse subset of the coordinates. The challenge is to borrow strength across the coordinates in order to detect smaller changes than could be observed in any individual component series. We propose a two-stage procedure called `inspect` for estimation of the changepoints: first, we argue that a good projection direction can be obtained as the leading left singular vector of the matrix that solves a convex optimisation problem derived from the CUSUM transformation of the time series. We then apply an existing univariate changepoint detection algorithm to the projected series. Our theory provides strong guarantees on both the number of estimated changepoints and the rates of convergence of their locations, and our numerical studies validate its highly competitive empirical performance for a wide range of data generating mechanisms.

1 Introduction

One of the most commonly-encountered issues with Big Data is heterogeneity. When collecting vast quantities of data, it is usually unrealistic to expect that stylised, traditional statistical models of independent and identically distributed observations can adequately capture the complexity of the underlying data generating mechanism. Departures from such models may take many forms, including missing data, correlated errors and data combined from multiple sources, to mention just a few.

When data are collected over time, heterogeneity often manifests itself through non-stationarity, where the data generating mechanism varies with time. Perhaps the simplest form of non-stationarity assumes that population changes occur at a relatively small number

*Research supported by a Benefactors' Scholarship from St John's College, Cambridge.
†Research supported by an EPSRC Fellowship, an EPSRC Programme Grant and a Philip Leverhulme Prize.

of discrete time points. If correctly estimated, these 'changepoints' can be used to partition the original data set into shorter segments, which can then be analysed using methods designed for stationary time series. Moreover, the locations of these changepoints are often themselves of great practical interest.

In this paper, we study high-dimensional time series that may have changepoints; moreover, we consider in particular settings where at a changepoint, the mean structure changes in a sparse subset of the coordinates. Despite their simplicity, such models are of great interest in a wide variety of applications. For instance, in the case of stock price data, it may well be the case that stocks in related industry sectors experience virtually simultaneous 'shocks' (Chen and Gupta, 1997). In internet security monitoring, a sudden change in traffic at multiple routers may be an indication of a distributed denial of service attack (Peng, Leckie and Ramamohanarao, 2004). In functional Magnetic Resonance Imaging (fMRI) studies, a rapid change in blood oxygen level dependent (BOLD) constrast in a subset of voxels may suggest neurological activity of interest (Aston and Kirch, 2012).

Our main contribution is to propose a new method for estimating the number and locations of the changepoints in such high-dimensional time series, a challenging task in the absence of knowledge of the coordinates that undergo a change. In brief, we first seek a good projection direction, which should ideally be closely aligned with the vector of mean changes. We can then apply an existing univariate changepoint detection algorithm to the projected series. For this reason, we call our algorithm `inspect`, short for informative sparse projection for estimation of changepoints.

In more detail, in the single changepoint case, our first observation is that at the population level, the vector of mean changes is the leading left singular vector of the matrix obtained as the cumulative sum (CUSUM) transformation of the mean matrix of the time series. This motivates us to begin by applying the CUSUM transformation to the time series. Unfortunately, computing the k-sparse leading left singular vector of a matrix is a combinatorial optimisation problem, but nevertheless, we are able to formulate an appropriate convex relaxation of the problem, from which we derive our projection direction. At the second stage of our algorithm, we compute the vector of CUSUM statistics for the projected series, identifying a changepoint if the maximum absolute value of this vector is sufficiently large. For the case of multiple changepoints, we combine our single changepoint algorithm with the method of Wild Binary Segmentation (WBS) (Fryzlewicz, 2014) to identify changepoints recursively.

A brief illustration of the `inspect` algorithm in action is given in Figure 1. Here, we simulated a 2000×1000 data matrix having independent normal columns with identity covariance and with three changepoints in the mean structure at locations $500, 1000$ and 1500. Changes occur in 40 coordinates, where consecutive changepoints overlap in half of their coordinates, and the squared ℓ_2 norms of the vectors of mean changes were 0.4, 0.9 and 1.6 respectively. The top-left panel shows the original data matrix and the top-right shows its CUSUM transformation, while the bottom-left panel shows overlays for the three detected changepoints of the univariate CUSUM statistics after projection. Finally, the bottom-right panel displays the largest absolute values of the projected CUSUM statistics obtained by running the wild binary segementation algorithm to completion (in practice, we would apply a termination criterion instead, but this is still helpful for illustrative purposes). We see that

the three detected changepoints are very close to their true locations, and it is only for these three locations that we obtain a sufficiently large CUSUM statistic to declare a changepoint.

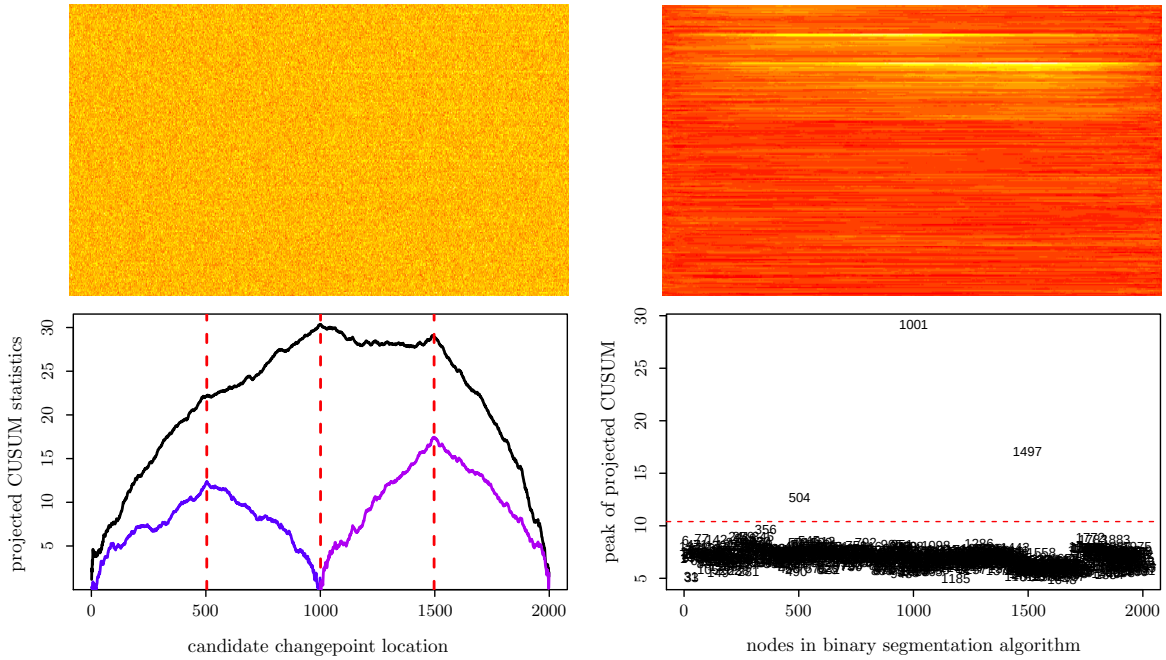

Figure 1: Example of `inspect` algorithm in action. Top-left: visualisation of the data matrix. Top-right: Its CUSUM transformation. Bottom-left: overlay of the projected CUSUM statistics for the three changepoints detected. Bottom-right: visualisation of thresholding; the three detected changepoints are above the threshold (dotted red line) whereas the remaining numbers are the test statistics obtained if we run the wild binary segmentation to completion without applying a termination criterion.

Our theoretical development proceeds first by controlling the angle between the estimated projection direction and the optimal direction, which is given by the normalised vector of mean changes. Under appropriate conditions, this enables us to provide finite-sample bounds which guarantee that with high probability we both recover the correct number of changepoints, and estimate their locations to within a specified accuracy. Our extensive numerical studies indicate that the algorithm performs extremely well in a wide variety of settings.

The study of changepoint problems dates at least back to Page (1955), and has since found applications in many different areas, including genetics (Olshen et al., 2004), disease outbreak watch (Sparks, Keighley and Muscatello, 2010) and aerospace engineering (Henry, Simani and Patton, 2010), in addition to those already mentioned. There is a vast and rapidly growing literature on different methods for changepoint detection and localisation, especially in the univariate problem. Surveys of various methods can be found in Csörgö and Horváth (1997) and Horváth and Rice (2014). In the case of univariate changepoint estimation, state-of-the-art methods include Pruned Exact Linear Time method (PELT)

(Killick, Fearnhead and Eckley, 2012), Wild Binary Segmentation (WBS) (Fryzlewicz, 2014) and Simultaneous Multiscale Changepoint Estimator (SMUCE) (Frick, Munk and Sieling, 2014).

Some of the univariate changepoint methodologies have been extended to multivariate settings. Examples include Horváth, Kokoszka and Steinebach (1999), Ombao, Von Sachs and Guo (2005), Aue et al. (2009) and Kirch, Mushal and Ombao (2014). However, there are fewer available tools for high-dimensional changepoint problems, where both the dimension p and the length n of the data stream may be large, and where we may allow a sparsity assumption on the coordinates of change. Bai (2010) investigates the performance of the least squares estimator of a single changepoint in the high-dimensional setting. Zhang et al. (2010), Horváth and Hušková (2012) and Enikeeva and Harchaoui (2014) consider estimators based on ℓ_2 aggregations of CUSUM statistics in all coordinates, but without using any sparsity assumptions. Enikeeva and Harchaoui (2014) also consider a scan statistic that takes sparsity into account. Jirak (2015) considers an ℓ_∞ aggregation of the CUSUM statistics that works well for sparse changepoints. Cho and Fryzlewicz (2015) propose Sparse Binary Segmentation (SBS), which also takes sparsity into account and can be viewed as a hard-thresholding of the CUSUM matrix followed by an ℓ_1 aggregation. Cho (2016) proposes a double-CUSUM algorithm that performs a CUSUM transformation along the location axis on the columwise-sorted CUSUM matrix. In a slightly different setting, Lavielle and Teyssiere (2006), Aue et al. (2009), Bücher et al. (2014), Preuß et al. (2015) and Cribben and Yu (2015) deal with changes in cross-covariance. Aston and Kirch (2014) considered the asymptotic efficiency of detecting a single changepoint in a high-dimensional setting, and the oracle projection-based estimator under cross-sectional dependence structure.

The outline of the rest of the paper is as follows. In Section 2, we give a formal description of the problem and the class of data generating mechanisms under which our theoretical results hold. Our methodological development in the single changepoint setting is presented in Section 3, and includes theoretical guarantees on both the projection direction and location of the estimated changepoint. Section 4 extends these ideas to the case of multiple changepoints with the aid of Wild Binary Segmentation, and our numerical studies are given in Section 5.

We conclude this section by introducing some notation used throughout the paper. For a vector $u = (u_1, \ldots, u_M)^\top \in \mathbb{R}^M$, a matrix $A = (A_{ij}) \in \mathbb{R}^{M \times N}$ and for $q \in [1, \infty)$, we write $\|u\|_q := \left(\sum_{i=1}^M |u_i|^q\right)^{1/q}$ and $\|A\|_q := \left(\sum_{i=1}^M \sum_{j=1}^N |A_{ij}|^q\right)^{1/q}$ for their (entrywise) ℓ_q-norms, as well as $\|u\|_\infty := \max_{i=1,\ldots,M} |u_i|$ and $\|A\|_\infty := \max_{i=1,\ldots,M, j=1,\ldots,n} |A_{ij}|$. We write $\|A\|_* := \sum_{i=1}^{\min(M,N)} \sigma_i(A)$ and $\|A\|_{\mathrm{op}} := \max_i \sigma_i(A)$ respectively for the nuclear norm and operator norm of matrix A, where $\sigma_1(A), \ldots, \sigma_{\min(M,N)}(A)$ are the singular values of A. We also write $\|u\|_0 := \sum_{i=1}^M \mathbb{1}_{\{u_i \neq 0\}}$. For $S \subseteq \{1, \ldots, M\}$ and $T \subseteq \{1, \ldots, N\}$, we write $u_S := (u_i : i \in S)^\top$ and write $M_{S,T}$ for the $|S| \times |T|$ submatrix of A obtained by extracting the rows and columns with indices in S and T respectively. For two matrices $A, B \in \mathbb{R}^{M \times N}$, we denote their trace inner product as $\langle A, B \rangle = \mathrm{tr}(A^\top B)$. For two non-zero vectors $u, v \in \mathbb{R}^p$, we write $\angle(u, v) := \cos^{-1}(\frac{\langle u, v \rangle}{\|u\|_2 \|v\|_2})$ for the acute angle bounded between them. We let $\mathbb{S}^{p-1} := \{x \in \mathbb{R}^p : \|x\|_2 = 1\}$ be the unit Euclidean sphere in $\mathbb{R}^p$, and let $\mathbb{S}^{p-1}(k) := \{x \in \mathbb{S}^{p-1} : \|x\|_0 \leq k\}$.

2 Problem description

Let $X_1, \ldots, X_n$ be independent p-dimensional random vectors sampled from

$$X_t \sim N_p(\mu_t, \sigma^2 I_p), \qquad 1 \leq t \leq n, \tag{1}$$

and combine the observations into a matrix $X = (X_1, \ldots, X_n) \in \mathbb{R}^{p \times n}$. We assume that the mean vectors follow a piecewise-constant structure with at most $\nu + 1$ segments. In other words, there exists ν *changepoints*

$$1 \leq z_1 < z_2 < \cdots < z_\nu \leq n - 1$$

such that

$$\mu_{z_i+1} = \cdots = \mu_{z_{i+1}} =: \mu^{(i)}, \qquad \forall\, 0 \leq i \leq \nu,$$

where we adopt the convention that $z_0 := 0$ and $z_{\nu+1} := n$. For $i = 1, \ldots, \nu$, write

$$\theta^{(i)} := \mu^{(i)} - \mu^{(i-1)}$$

for the difference in means between consecutive stationary segments. We assume that the changes in mean are sparse in the sense that there exists $k \in \{1, \ldots, p\}$ (typically k is much smaller than p) such that $\|\theta^{(i)}\|_0 \leq k$ for each $i = 1, \ldots, \nu$.

Our goal is to estimate the set of changepoints $\{z_1, \ldots, z_\nu\}$ in the high-dimensional regime, where p may be comparable to, or even larger than, the length n of the series. The signal strength of the estimation problem is determined by the magnitude of mean changes $\{\theta^{(i)} : 1 \leq i \leq \nu\}$ and the run lengths of stationary segments $\{z_{i+1} - z_i : 0 \leq i \leq \nu\}$, whereas the noise is related to the variance σ^2 and the dimensionality p of the observed data points. We let $\mathcal{P}(n, p, k, \nu, \vartheta, \tau, \sigma^2)$ denote the class of distributions of $X = (X_1, \ldots, X_n) \in \mathbb{R}^{p \times n}$ with independent columns drawn from (1), where the changepoint locations satisfy

$$n^{-1} \min\{z_{i+1} - z_i : 0 \leq i \leq \nu\} \geq \tau,$$

and the magnitudes of mean changes are such that

$$\|\theta^{(i)}\|_2^2 \geq k\vartheta^2, \qquad \forall\, 1 \leq i \leq \nu.$$

Suppose that an estimation procedure outputs $\hat{\nu}$ changepoints located at $1 \leq \hat{z}_1 < \cdots < \hat{z}_{\hat{\nu}} \leq n - 1$. Our finite-sample bounds will imply a rate of convergence for `inspect` in an asymptotic setting where $(p, k, \nu, \vartheta, \tau, \sigma^2) = (p_n, k_n, \nu_n, \vartheta_n, \tau_n, \sigma_n^2)$. In this context, we follow the convention in the literature (e.g. Venkatraman, 1992) and say that the procedure is consistent with rate of convergence ρ_n if

$$\sup_{P \in \mathcal{P}(n,p,k,\nu,\vartheta,\tau,\sigma^2)} \mathbb{P}_P\big\{\hat{\nu} = \nu \text{ and } |\hat{z}_i - z_i| \leq n\rho_n \text{ for all } 1 \leq i \leq \nu\big\} \to 1 \tag{2}$$

as $n \to \infty$. We remark that consistency as defined above is a rather strong notion, in the sense that it implies convergence in several other natural metrics. For example, if we let

$$d_{\mathrm{H}}(A, B) := \max\Big\{\sup_{a \in A} \inf_{b \in B} |a - b|, \sup_{b \in B} \inf_{a \in A} |a - b|\Big\}$$

denote the Hausdorff distance between non-empty sets A and B on $\mathbb{R}$, then (2) implies that with probability tending to 1,

$$\frac{1}{n} d_{\mathrm{H}}\big(\{\hat{z}_i : 1 \le i \le \hat{\nu}\}, \{z_i : 1 \le i \le \nu\}\big) \le \rho_n.$$

Similarly, denote the L_1-Wasserstein distance between probability measures P and Q on $\mathbb{R}$ by

$$d_{\mathrm{W}}(P, Q) := \inf_{(U,V)\sim(P,Q)} \mathbb{E}|U - V|$$

where the infimum is taken over all pairs of random variables U and V defined on the same probability space with $U \sim P$ and $V \sim Q$. Then (2) also implies that with probability tending to 1,

$$\frac{1}{n} d_{\mathrm{W}}\left(\frac{1}{\hat{\nu}} \sum_{i=1}^{\hat{\nu}} \delta_{\hat{z}_i}, \frac{1}{\nu} \sum_{i=1}^{\nu} \delta_{z_i}\right) \le \rho_n,$$

where δ_a denotes a Dirac point mass at a.

3 Data-driven projection estimator for a single changepoint

We first consider the problem of estimating a single changepoint (i.e. $\nu = 1$) in a high-dimensional time series dataset $X \in \mathbb{R}^{p\times n}$. For simplicity, write $z := z_1$, $\theta := \theta^{(1)}$ and $\tau := n^{-1}\min\{z, n - z\}$. We seek to aggregate the rows of the data matrix X in an almost optimal way so as to maximise the signal-to-noise ratio, and then locate the changepoint using a one-dimensional procedure. For any $a \in \mathbb{S}^{p-1}$, $a^\top X$ is a one-dimensional time series with

$$a^\top X_t \sim N(a^\top \mu_t, \sigma^2).$$

Hence, the choice $a = \theta/\|\theta\|_2$ maximises the magnitude of the difference in means between the two segments. However, θ is typically unknown in practice, so we should seek a projection direction that is close to the oracle projection direction $v := \theta/\|\theta\|_2$. Our strategy is to perform sparse singular value decomposition on the CUSUM transformation of X. The method and limit theory of CUSUM statistics in the univariate case can be traced back to Darling and Erdős (1956). For $p \in \mathbb{N}$ and $n \ge 2$, we define the CUSUM transformation $\mathcal{T}_{p,n} : \mathbb{R}^{p\times n} \to \mathbb{R}^{p\times(n-1)}$ by

$$\begin{aligned}[\mathcal{T}_{p,n}(M)]_{j,t} :&= \sqrt{\frac{t(n-t)}{n}}\left(\frac{1}{n-t}\sum_{r=t+1}^{n} M_{j,r} - \frac{1}{t}\sum_{r=1}^{t} M_{j,r}\right) \\ &= \sqrt{\frac{n}{t(n-t)}}\left(\frac{t}{n}\sum_{r=1}^{n} M_{j,r} - \sum_{r=1}^{t} M_{j,r}\right). \qquad (3)\end{aligned}$$

In fact, to simplify the notation, we will write $\mathcal{T}$ for $\mathcal{T}_{p,n}$, since p and n can be inferred from the dimensions of the argument of $\mathcal{T}$. Note also that $\mathcal{T}$ reduces to computing the vector of

classical one-dimensional CUSUM statistics when $p = 1$. We write

$$X = \boldsymbol{\mu} + W,$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n) \in \mathbb{R}^{p \times n}$ and $W = (W_1, \ldots, W_n) \in \mathbb{R}^{p \times n}$ is a $p \times n$ matrix with independent $N_p(0, \sigma^2 I_p)$ columns. Let $T := \mathcal{T}(X)$, $A := \mathcal{T}(\boldsymbol{\mu})$ and $E := \mathcal{T}(W)$, so by the linearity of the CUSUM transformation we have the decomposition

$$T = A + E.$$

In the single changepoint case, the entries of the matrix A can be computed explicitly:

$$A_{j,t} = \begin{cases} \sqrt{\frac{t}{n(n-t)}}(n-z)\theta_j, & \text{if } t \leq z \\ \sqrt{\frac{n-t}{nt}} z\theta_j, & \text{if } t > z. \end{cases}$$

Hence we can write

$$A = \theta\gamma^\top, \tag{4}$$

where

$$\gamma := \frac{1}{\sqrt{n}}\left(\sqrt{\frac{1}{n-1}}(n-z), \sqrt{\frac{2}{n-2}}(n-z), \ldots, \sqrt{z(n-z)}, \sqrt{\frac{n-z-1}{z+1}}z, \ldots, \sqrt{\frac{1}{n-1}}z\right)^\top. \tag{5}$$

In particular, this implies that the oracle projection direction is the leading left singular vector of the rank 1 matrix A. In the ideal case where k is known, we could in principle let $\hat{v}_{\max,k}$ be a k-sparse leading left singular vector of T, defined by

$$\hat{v}_{\max,k} \in \operatorname*{argmax}_{\tilde{v} \in \mathbb{S}^{p-1}(k)} \|T^\top \tilde{v}\|_2, \tag{6}$$

and it can then be shown using a perturbation argument akin to the Davis–Kahan '$\sin\theta$' theorem (cf. Davis and Kahan (1970); Yu, Wang and Samworth (2015)) that $\hat{v}_{\max,k}$ is a consistent estimator of the oracle projection direction v under mild conditions (see Proposition 9 in Appendix B). However, the optimisation problem in (6) is non-convex and hard to implement. In fact, computing the k-sparse leading left singular vector of a matrix is known to be NP-hard (e.g. Tillmann and Pfetsch (2014)). The naive algorithm that scans through all possible k-subsets of the rows of T has running time exponential in k, which quickly becomes impractical to run for even moderate sizes of k.

A natural approach to remedy this computational issue is to work with a convex relaxation of the optimisation problem (6) instead. In fact, we can write

$$\begin{aligned} \max_{u \in \mathbb{S}^{p-1}(k)} \|u^\top T\|_2 &= \max_{u \in \mathbb{S}^{p-1}(k), w \in \mathbb{S}^{n-2}} u^\top T w \\ &= \max_{u \in \mathbb{S}^{p-1}, w \in \mathbb{S}^{n-2}, \|u\|_0 \leq k} \langle uw^\top, T \rangle = \max_{M \in \mathcal{M}} \langle M, T \rangle, \end{aligned} \tag{7}$$

where $\mathcal{M} := \{M \in \mathbb{R}^{p \times (n-1)} : \|M\|_* = 1, \operatorname{rank}(M) = 1, M \text{ has at most } k \text{ non-zero rows}\}$. The final expression in (7) has a convex (linear) objective function $M \mapsto \langle M, T \rangle$. The

requirement $\operatorname{rank}(M) = 1$ in the constraint set $\mathcal{M}$ is equivalent to $\|\sigma(M)\|_0 = 1$, where $\sigma(M) := (\sigma_1(M), \ldots, \sigma_{\min(p,n-1)}(M))^\top$ is the vector of singular values of M. This motivates us to absorb the rank constraint into the nuclear norm constraint, which we relax from an equality constraint to an inequality constraint in order to make it convex. Furthermore, we can relax the row sparsity constraint in the definition of $\mathcal{M}$ to an entrywise ℓ_1-norm penalty. The optimisation problem of finding

$$\hat{M} \in \operatorname*{argmax}_{M \in \mathcal{S}_1} \big\{ \langle T, M \rangle - \lambda \|M\|_1 \big\}, \tag{8}$$

where $\mathcal{S}_1 := \{M \in \mathbb{R}^{p \times (n-1)} : \|M\|_* \leq 1\}$ and $\lambda > 0$ is a tuning parameter to be chosen later, is therefore a convex relaxation of (6). The convex problem (8) may be solved using the alternating direction method of multipliers algorithm (ADMM, see Gabay and Mercier (1976); Boyd et al. (2011)) as in Algorithm 1. More specifically, the optimisation problem in (8) is equivalent to maximising $\langle T, Y \rangle - \lambda \|Z\|_1 - \mathbb{I}_{\mathcal{S}_1}(Y)$ subject to $Y = Z$, where $\mathbb{I}_{\mathcal{S}_1}$ is the function that is 0 on $\mathcal{S}_1$ and ∞ on $\mathcal{S}_1^c$. Its augmented Lagrangian is given by

$$L(Y, Z, R) := \langle T, Y \rangle - \mathbb{I}_{\mathcal{S}_1}(Y) - \lambda \|Z\|_1 - \langle R, Y - Z \rangle - \frac{1}{2} \|Y - Z\|_2^2,$$

with the Lagrange multiplier R being the dual variable. Each iteration of the main loop in Algorithm 1 first performs a primal update by maximising marginally with respect to Y and Z, then followed by a dual gradient update of R with constant step size. The function $\Pi_{\mathcal{S}_1}(\cdot)$ in Algorithm 1 denotes projection onto the convex set $\mathcal{S}_1$ with respect to the Frobenius norm distance. If $A = UDV^\top$ is the singular value decomposition of $A \in \mathbb{R}^{p \times (n-1)}$ with $\operatorname{rank}(A) = r$, where D is a diagonal matrix with diagonal entries $d_1, \ldots, d_r$, then $\Pi_{\mathcal{S}_1}(A) = U\tilde{D}V^\top$, where $\tilde{D}$ is a diagonal matrix with entries $\tilde{d}_1, \ldots, \tilde{d}_r$ such that $(\tilde{d}_1, \ldots, \tilde{d}_r)^\top$ is the Euclidean projection of the vector $(d_1, \ldots, d_r)^\top$ onto the standard $(r-1)$-simplex

$$\Delta^{r-1} := \left\{ (x_1, \ldots, x_r)^\top \in \mathbb{R}^r : \sum_{\ell=1}^{r} x_\ell = 1 \text{ and } x_\ell \geq 0 \text{ for all } \ell \right\}.$$

For an efficient algorithm for such simplicial projection, see Chen and Ye (2011). The **soft** function in Algorithm 1 denotes an entrywise soft-thresholding operator defined by $\big(\mathbf{soft}(A, \lambda)\big)_{ij} = \operatorname{sgn}(A_{ij}) \max\{|A_{ij}| - \lambda, 0\}$ for any $\lambda \geq 0$ and matrix $A = (A_{ij})$.

We remark that one may be interested to further relax (8) by replacing $\mathcal{S}_1$ with the larger set $\mathcal{S}_2 := \{M \in \mathbb{R}^{p \times (n-1)} : \|M\|_2 \leq 1\}$. We see from Lemma 10 in Appendix B that the smoothness of $\mathcal{S}_2$ results in a simple dual formulation, which implies that

$$\tilde{M} := \frac{\mathbf{soft}(T, \lambda)}{\|\mathbf{soft}(T, \lambda)\|_2} = \operatorname*{argmax}_{M \in \mathcal{S}_2} \big\{ \langle T, M \rangle - \lambda \|M\|_1 \big\} \tag{9}$$

is the unique optimiser of the primal problem. The soft-thresholding operation is significantly faster than the ADMM algorithm in Algorithm 1. Hence by enlarging $\mathcal{S}_1$ to $\mathcal{S}_2$, we can significantly speed up the running time of the algorithm in exchange for some loss in statistical efficiency caused by the further relaxation of the constraint set. See Section 5 for further discussion.

Algorithm 1: Pseudo-code for an ADMM algorithm that computes the solution to the optimisation problem (8).

Input: $T \in \mathbb{R}^{p\times(n-1)}$, $\lambda > 0$.
Set: $Y = Z = R = \mathbf{0} \in \mathbb{R}^{p\times(n-1)}$
repeat
  $Y \leftarrow \Pi_{\mathcal{S}_1}(Z - R + T)$
  $Z \leftarrow \mathbf{soft}(Y + R, \lambda)$
  $R \leftarrow R + (Y - Z)$
until $Y - Z$ *converges to* 0
$\hat{M} \leftarrow Y$
Output: $\hat{M}$

Let $\hat{v}$ be the leading left singular vector of

$$\hat{M} \in \operatorname*{argmax}_{M\in\mathcal{S}} \big\{ \langle T, M\rangle - \lambda \|M\|_1 \big\}, \tag{10}$$

for either $\mathcal{S} = \mathcal{S}_1$ or $\mathcal{S} = \mathcal{S}_2$. In Proposition 1 below, we provide an error bound on $\hat{v}$ as an estimator of the oracle projection direction v. It relies on a generalisation of the curvature lemma in Vu et al. (2013, Lemma 3.1), presented as Lemma 4 in Appendix B.

Proposition 1. *Suppose that $\hat{M}$ satisfies (10) for either $\mathcal{S} = \mathcal{S}_1$ or $\mathcal{S} = \mathcal{S}_2$. Let $\hat{v} \in \operatorname{argmax}_{\tilde{v}\in\mathbb{S}^{p-1}} \|\hat{M}^\top \tilde{v}\|_2$ be the leading left singular vector of $\hat{M}$. If $n \geq 6$ and if we choose $\lambda \geq 2\sigma\sqrt{\log(p\log n)}$, then*

$$\sup_{P\in\mathcal{P}(n,p,k,1,\vartheta,\tau,\sigma^2)} \mathbb{P}_P\left(\sin\angle(\hat{v}, v) > \frac{32\lambda}{\tau\vartheta\sqrt{n}} \right) \leq \frac{4}{(p\log n)^{1/2}}.$$

As an illustration, consider $\lambda = 2\sigma\sqrt{\log(p\log n)}$ and the asymptotic regime where $\log p = O(\log n)$, $\vartheta \asymp n^{-a}$ and $\tau \asymp n^{-b}$ for some $a \in \mathbb{R}$ and $b \geq 0$. Then Proposition 1 implies that as long as $a + b < 1/2$, we have $\angle(\hat{v}, v) \xrightarrow{\mathrm{p}} 0$.

Algorithm 2: Pseudo-code for a single high-dimensional changepoint estimation algorithm.

Input: $X \in \mathbb{R}^{p\times n}$, $\lambda > 0$.
Step 1: Perform the CUSUM transformation $T \leftarrow \mathcal{T}(X)$
Step 2: Use Algorithm 1 or (9) (with inputs T, λ in either case) to solve for an optimiser $\hat{M}$ of (10) for $\mathcal{S} = \mathcal{S}_1$ or $\mathcal{S}_2$
Step 3: Find $\hat{v} \in \operatorname{argmax}_{\tilde{v}\in\mathbb{S}^{p-1}} \|\hat{M}^\top \tilde{v}\|_2$.
Step 4: Let $\hat{z} \in \operatorname{argmax}_{1\leq t\leq n-1} |\hat{v}^\top T_t|$, where T_t is the tth column of T, and set $\bar{T}_{\max} \leftarrow |\hat{v}^\top T_{\hat{z}}|$
Output: $\hat{z}$, $\bar{T}_{\max}$

After obtaining a reasonable estimator $\hat{v}$ of the oracle projection direction, the natural next step is to project the data matrix X along the direction $\hat{v}$, and apply an existing one-dimensional changepoint localisation method on the projected data. In this work, we apply a one-dimensional CUSUM transformation to the projected time series and estimate the changepoint by the location of the maximum of the CUSUM vector. Our overall procedure for locating a single changepoint in a high-dimensional time series is given in Algorithm 2. In our description of this algorithm, the noise level σ is assumed to be known. If σ is unknown, we can estimate it robustly using, e.g., the median absolute deviation of the marginal one-dimensional time series (Hampel, 1974). Note that for convenience of later reference, we have required Algorithm 2 to output both the estimated changepoint location $\hat{z}$ and the associated maximum absolute post-projection one-dimensional CUSUM statistic $\bar{T}_{\max}$.

From a theoretical point of view, the fact that $\hat{v}$ is estimated using the entire dataset X makes it difficult to analyse the post-projection noise structure. For this reason, in the analysis below, we work with a slight variant of Algorithm 2. We assume for convenience that $n = 2n_1$ is even, and define $X^{(1)}, X^{(2)} \in \mathbb{R}^{p\times n_1}$ by

$$X^{(1)}_{j,t} := X_{j,2t-1} \quad \text{and} \quad X^{(2)}_{j,t} := X_{j,2t} \quad \text{for } 1 \le j \le p, 1 \le t \le n_1. \tag{11}$$

We then use $X^{(1)}$ to estimate the oracle projection direction and use $X^{(2)}$ to estimate the changepoint location after projection (see Algorithm 3). However, in our experience, $\hat{v}$ is almost independent of T and we recommend using Algorithm 2 in practice to exploit the full signal strength in the data.

Algorithm 3: Pseudo-code for a sample-splitting variant of Algorithm 2.

Input: $X \in \mathbb{R}^{p\times n}$, $\lambda > 0$.

Step 1: Perform the CUSUM transformation $T^{(1)} \leftarrow \mathcal{T}(X^{(1)})$ and $T^{(2)} \leftarrow \mathcal{T}(X^{(2)})$.

Step 2: Use Algorithm 1 or (9) (with inputs $T^{(1)}$, λ in either case) to solve for $\hat{M}^{(1)} \in \operatorname{argmax}_{M\in\mathcal{S}}\big\{\langle T^{(1)}, M\rangle - \lambda\|M\|_1\big\}$ with $\mathcal{S} = \{M \in \mathbb{R}^{p\times(n_1-1)} : \|M\|_* \le 1\}$ or $\{M \in \mathbb{R}^{p\times(n_1-1)} : \|M\|_2 \le 1\}$.

Step 3: Find $\hat{v}^{(1)} \in \operatorname{argmax}_{\tilde{v}\in\mathbb{S}^{p-1}} \|(\hat{M}^{(1)})^\top \tilde{v}\|_2$.

Step 4: Let $\hat{z} \in 2\operatorname{argmax}_{1\le t\le n_1-1}\big|(\hat{v}^{(1)})^\top T^{(2)}_t\big|$, where $T^{(2)}_t$ is the tth column of $T^{(2)}$, and set $\bar{T}_{\max} \leftarrow \big|(\hat{v}^{(1)})^\top T^{(2)}_{\hat{z}/2}\big|$.

Output: $\hat{z}, \bar{T}_{\max}$

We summarise the overall estimation performance of Algorithm 3 in the following theorem.

Theorem 2. *Suppose $\sigma > 0$ is known. Let $\hat{z}$ be the output of Algorithm 3 with input X and $\lambda := 2\sigma\sqrt{\log(p\log n)}$. If $n \ge 6$ is even and*

$$\frac{\sigma}{\vartheta\tau}\sqrt{\frac{\log(p\log n)}{n}} \le \frac{\sqrt{3}}{128}, \tag{12}$$

then

$$\sup_{P\in\mathcal{P}(n,p,k,1,\vartheta,\tau,\sigma^2)} \mathbb{P}_P\left(\frac{1}{n}|\hat{z}-z| > \frac{32\sigma}{\vartheta\sqrt{k\tau}}\sqrt{\frac{\log n}{n}}\right) \leq \frac{4}{\{p\log(n/2)\}^{1/2}} + \frac{2}{n}.$$

Again, to illustrate, suppose we are in the asymptotic regime where $\log p = O(\log n)$, $\vartheta \asymp n^{-a}$, $\tau \asymp n^{-b}$ and $k \asymp n^{c}$ for some $a \in \mathbb{R}$ and $b \in [0,1]$ and $c \geq 0$. If $a + b < 1/2$, then Theorem 2 implies that the output $\hat{z}$ of Algorithm 3 is a consistent estimator of the true changepoint z with rate of convergence $\rho_n = o(n^{-\frac{1-2a-b+c}{2}+\delta})$ for any $\delta > 0$.

4 Estimating multiple changepoints

Our algorithm for a single changepoint can be combined with the wild binary segmentation scheme of Fryzlewicz (2014) to sequentially locate multiple changepoints in high-dimensional time series. The principal idea behind a wild binary segmentation procedure is as follows. We first randomly sample a large number of pairs, $(s_1, e_1), \ldots, (s_Q, e_Q)$ uniformly from the set $\{(\ell, r) \in \mathbb{Z}^2 : 0 \leq \ell < r \leq n\}$, and then apply our single changepoint algorithm to $X^{[q]}$, for $1 \leq q \leq Q$, where $X^{[q]}$ is defined to be the submatrix of X obtained by extracting columns $\{s_q + 1, \ldots, e_q\}$ of X. For each $1 \leq q \leq Q$, the single changepoint algorithm (Algorithm 2 or 3) will estimate an optimal sparse projection direction $\hat{v}^{[q]}$, compute a candidate changepoint location $s_q + \hat{z}^{[q]}$ within the time window $[s_q + 1, e_q]$ and return a maximum absolute CUSUM statistic $\bar{T}^{[q]}_{\max}$ along the projection direction. We aggregate the q candidate changepoint locations by choosing one that maximises the largest maximum projected CUSUM statistic, $T^{[q]}_{\max}$, as our best candidate. If $T^{[q]}_{\max}$ is above a certain threshold value ξ, we admit the best candidate to the set $\hat{Z}$ of estimated changepoint locations and repeat the above procedure recursively on the sub-segments to the left and right of the estimated changepoint. Note that while recursing on a sub-segment, we only consider those time windows that are completely contained in the sub-segment. The precise algorithm is detailed in Algorithm 4.

Algorithm 4 requires three tuning parameters: a regularisation parameter λ, a Monte Carlo parameter Q for the number of random time windows and a thresholding parameter ξ that determines termination of recursive segmentation. Theorem 3 below provides choices for λ, Q and ξ that yield theoretical guarantees for consistent estimation of all changepoints as defined in (2).

We remark that if we apply Algorithm 2 or 3 on the entire dataset X instead of random time windows of X, and then iterate after segmentation, we arrive at a multiple changepoint algorithm based on the classical binary segmentation scheme. The main disadvantage of this classical binary segmentation procedure is its sensitivity to model misspecification. Algorithms 2 and 3 are designed to optimise the detection of a single changepoint. When we apply them to a time series containing more than one changepoint, as in the classical binary segmentation scheme, the signals from multiple changepoints may cancel each other out in two different ways that will lead to a loss of power. First, as Fryzlewicz (2014) points out in the one-dimensional setting, multiple changepoints may offset each other in CUSUM computation, resulting in a smaller peak of the CUSUM statistic that is more easily contaminated by the noise. Moreover, in a high-dimensional setting, different changepoints can undergo

Algorithm 4: Pseudo-code for multiple changepoint algorithm based on sparse singular vector projection and wild binary segmentation.

Input: $X \in \mathbb{R}^{p\times n}$, $\lambda > 0$, $\xi > 0$, $\beta > 0$, $Q \in \mathbb{N}$.

Step 1: Set $\hat{Z} \leftarrow \emptyset$. Draw Q pairs of integers $(s_1, e_1), \ldots, (s_Q, e_Q)$ uniformly at random from the set $\{(\ell, r) \in \mathbb{Z}^2 : 0 \le \ell < r \le n\}$.

Step 2: Run `wbs`$(0,\ n)$ where `wbs` is defined below.

Step 3: Let $\hat{\nu} \leftarrow |\hat{Z}|$ and sort elements of $\hat{Z}$ in increasing order to yield $\hat{z}_1 < \cdots < \hat{z}_{\hat{\nu}}$.

Output: $\hat{z}_1, \ldots, \hat{z}_{\hat{\nu}}$

Function `wbs`$(s,\ e)$

  Set $\mathcal{Q}_{s,e} \leftarrow \{q : s + n\beta \le s_q < e_q \le e - n\beta\}$

  for $q \in \mathcal{Q}_{s,e}$ **do**

    Run Algorithm 2 with $X^{[q]}$, λ as input, and let $\hat{z}^{[q]}, \bar{T}^{[q]}_{\max}$ be the output.

  end

  Find $q_0 \in \mathrm{argmax}_{q\in\mathcal{Q}_{s,e}} \bar{T}^{[q]}_{\max}$ and set $b \leftarrow s_{q_0} + \hat{z}^{[q_0]}$

  if $\bar{T}^{[q_0]}_{\max} > \xi$ **then**

    $\hat{Z} \leftarrow \hat{Z} \cup \{b\}$

    `wbs`$(s,\ b)$

    `wbs`$(b,\ e)$

  end

end

changes in different sets of (sparse) coordinates. This also attenuates the signal strength in the sense that the estimated oracle projection direction from Algorithm 1 is aligned to some linear combination of $\theta^{(1)}, \ldots, \theta^{(\nu)}$, but not necessarily well-aligned to any one particular $\theta^{(i)}$. The wild binary segmentation scheme addresses the model misspecification issue by examining sub-intervals of the entire time length. When the number of time windows Q is sufficiently large and τ is not too small, with high probability we have reasonably long time windows that contain each individual changepoint. Hence the single changepoint algorithm will perform well on these segments.

Just as in the case of single changepoint detection, it is easier to analyse the theoretical performance of a sample-splitting version of Algorithm 4. However, to avoid notational clutter, we will prove a theoretical result without sample splitting, but with the assumption that whenever Algorithm 2 is used within Algorithm 4, its second and third steps (i.e. the steps for estimating the oracle projection direction) will be carried out based on an independent copy X' of X. We refer to such a variant of the algorithm with an access to an independent sample X' as Algorithm 4′. Theorem 3 below, which proves theoretical guarantees of Algorithm 4′, can then be readily adapted to work for a sample-splitting version of Algorithm 4, where we replace n by $n/2$ where necessary.

Theorem 3. *Suppose $\sigma > 0$ is known and $X, X' \overset{\text{iid}}{\sim} P \in \mathcal{P}(n, p, k, \nu, \vartheta, \tau, \sigma^2)$. Let $\hat{z}_1 < \cdots < \hat{z}_{\hat{\nu}}$ be the output of Algorithm 4′ with input X, X', $\lambda := 3\sigma\sqrt{\log(np)}$, $\xi := \lambda$, β and Q.*

Define $\rho = \rho_n := \frac{200\lambda}{\vartheta\sqrt{kn\tau^3}}$. *If* $n\tau \geq 14$, $2\rho < \beta < \frac{2}{9}\tau$ *and* $\rho\sqrt{k\tau} \leq 1$, *then*

$$\mathbb{P}_P\big\{\hat{\nu} = \nu \text{ and } |\hat{z}_i - z_i| \leq n\rho \text{ for all } 1 \leq i \leq \nu\big\} \geq 1 - \tau^{-1}e^{-\tau^2 Q/9} - 2n^{-3/2}p^{-5/2}.$$

To illustrate the conditions and conclusion of Theorem 3, we again consider the asymptotic setting where $\vartheta \asymp n^{-a}$, $\tau \asymp n^{-b}$, $k \asymp n^c$ and $\log p = O(\log n)$. In this case, the conditions of Theorem 3 hold for sufficiently large n if $a + b < 1/2$ and $2a + 5b - c < 1$. When these conditions are satisfied, Theorem 3 implies that Algorithm 4′ consistently estimates all changepoints with rate of convergence $\rho_n = o(n^{-\frac{1-2a-3b+c}{2}+\delta})$ for any $\delta > 0$.

5 Numerical studies

In this section, we examine the empirical performance of the `inspect` algorithm in a range of settings, and compare it with a variety of other recently-proposed methods. In both single- and multiple-changepoint scenarios, the implementation of `inspect` requires the choice of a regularisation parameter $\lambda > 0$ to be used in Algorithm 1 (which is called in Algorithms 2 and 4). In our experience, the theoretical choices $\lambda = \sigma\sqrt{2\log(p\log n)}$ and $\lambda = 3\sigma\sqrt{\log(np)}$ used in Theorems 2 and 3 produce consistent estimators as predicted by the theory, but are slightly conservative, and in practice we recommend the choice $\lambda = \sigma\sqrt{2^{-1}\log(p\log n)}$ in both cases. The noise level σ is estimated by concatenating the individual time series into a vector of length np and then computing the median absolute deviation using the scaling constant of 1.48 for the normal distribution (Hampel, 1974).

In Step 2 of Algorithm 2, we also have a choice between using $\mathcal{S} = \mathcal{S}_1$ and $\mathcal{S}_2$. The following numerical experiment demonstrates the difference in performance of the algorithm for these two choices. We took $n = 200$, $p = 100$, $k = 10$, with a single changepoint located at $z = 100$. Table 1 shows the angles between the oracle projection direction and estimated projection directions using both $\mathcal{S}_1$ and $\mathcal{S}_2$ as the signal level ϑ varies from 0.1 to 1. It can be seen that further relaxation from $\mathcal{S}_1$ to $\mathcal{S}_2$ incurs a relatively low cost in terms of the estimation quality of the projection direction, but it offers great improvement in running time due to the closed-form solution (cf. Lemma 10). Thus, even though the use of $\mathcal{S}_1$ remains a viable practical choice for offline data sets of moderate size, we use $\mathcal{S} = \mathcal{S}_2$ in the simulations that follow.

ϑ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$\angle(\hat{v}_{\mathcal{S}_1}, v)$	80.3	63.1	51.6	39.4	28.6	25.8	21.7	19.0	16.7	14.4
$\angle(\hat{v}_{\mathcal{S}_2}, v)$	79.5	63.9	52.9	40.6	30.2	27.3	23.4	20.4	18.0	15.6

Table 1: Angles (in degrees) between oracle projection direction v and estimated projection directions $\hat{v}_{\mathcal{S}_1}$ (using $\mathcal{S}_1$) and $\hat{v}_{\mathcal{S}_2}$ (using $\mathcal{S}_2$), for different choices of ϑ. Each reported value is averaged over 100 repetitions. Other simulation parameters: $n = 200$, $p = 100$, $k = 10$, $z = 100$, $\sigma^2 = 1$.

We compare the performance of the `inspect` algorithm with sparsified binary segmentation (`sbs`) (Cho and Fryzlewicz, 2015), the double CUSUM algorithm (`dc`) (Cho, 2016) and

a scan statistic-based algorithm (`scan`) derived from the work of Enikeeva and Harchaoui (2014). The latter statistic, rewritten in our notation, is

$$L_{\mathrm{scan}} := \max_{1 \le \tilde{z} \le n-1} \max_{1 \le \tilde{k} \le p} \frac{\sum_{1 \le j \le \tilde{k}} (T^2_{(j),\tilde{z}} - 1)}{\kappa \log\big\{ \binom{p}{\tilde{k}} np/\alpha \big\}}, \tag{13}$$

where $T^2_{(j),\tilde{z}}$ is the jth largest entry in absolute value in the $\tilde{z}$th column of $T := \mathcal{T}(X)$. We remark that Enikeeva and Harchaoui (2014) primarily concerns the use of L_{scan} to test for the existence of a changepoint. However, the scan statistic can be naturally modified into a changepoint location estimator by modifying the outermost max function in (13) to an argmax. It can then be extended a multiple changepoint estimation algorithm via a wild binary segmentation scheme in a similar way to our algorithm. Whenever tuning parameters are required in running these algorithms, we adopt the choices suggested by their authors in the relevant papers. In our simulations, all algorithms were run on the same data matrices, and the estimated changepoints over 100 repetitions were then aggregated and compared.

5.1 Single changepoint estimation

All four algorithms in our simulation study are top-down algorithms in the sense that their multiple changepoint procedure is built upon a single changepoint estimation submodule, which is used to locate recursively all changepoints via a (wild) binary segmentation scheme. It is therefore instructive first to compare their performance in the single changepoint estimation task. Our simulations were run for $n \in \{1000, 2000\}$, $p \in \{200, 500, 1000\}$, $k \in \{10, \lceil p^{1/2} \rceil, p\}$, $z = 0.4n$, $\sigma^2 = 1$ and $\vartheta \in \{1, 0.6, 0.4, 0.25, 0.18, 0.11, 0.07, 0.04\}$, with $\theta = (1, 2^{-1/2}, \ldots, k^{-1/2}, 0, \ldots, 0)^\top \in \mathbb{R}^p$. For definiteness, we let the n columns of X be independent, with the leftmost z columns drawn from $N_p(0, \sigma^2 I_p)$ and the remaining columns drawn from $N_p(\theta, \sigma^2 I_p)$. To avoid the influence of different threshold levels on the performance of the algorithms and to focus solely on their estimation precision, we assume that the existence of a single changepoint is known *a priori* and make all algorithms output their estimate of its location; estimation of the number of changepoints in a multiple-changepoint setting is studied in Section 5.2 below. In the interests of brevity, in Table 2, we report the root mean squared estimation error for only two values of ϑ, chosen to represent low and high signal-to-noise settings respectively. More precisely, for each choice of (n, p, k), the reported values of ϑ are the largest values in the set $\{1, 0.6, 0.4, 0.25, 0.18, 0.11, 0.07, 0.04\}$ such that the root mean squared error of at least one algorithm is less than $0.1n$ and $0.01n$ respectively. The omitted results are qualitatively similar. We also remark that the three choices for the parameter k correspond to constant/logarithmic sparsity, polynomial sparsity and non-sparse settings respectively. As a graphical illustration, Figure 2 displays density estimates of the estimated changepoint location by the different algorithms in two different settings taken from Table 2. One difficulty in presenting such estimates with kernel density estimators is the fact that different algorithms would require different choices of bandwidth, and these would need to be locally adaptive, due to the relatively sharp peaks. In order to avoid the choice of bandwidth skewing the visual representation, we therefore use the log-concave maximum likelihood estimators for each method (e.g. Dümbgen and Rufibach, 2009;

Cule, Samworth and Stewart, 2010), which is both locally adaptive and tuning-parameter free.

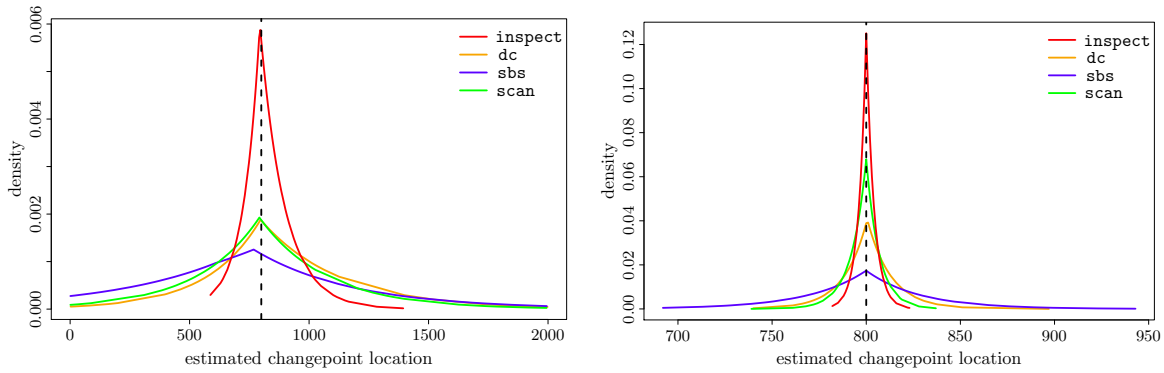

Figure 2: Estimated densities of location of changepoint estimates by `inspect`, `dc`, `sbs` and `scan`. Left panel: $(n, p, k, z, \vartheta, \sigma^2) = (2000, 1000, 32, 800, 0.07, 1)$; right panel: $(n, p, k, z, \vartheta, \sigma^2) = (2000, 1000, 32, 800, 0.18, 1)$.

It can be seen from Table 2 and Figure 2 that `inspect` has extremely competitive performance for the single changepoint estimation task, in both low and high signal-to-noise settings. In particular, despite the fact that it is designed for estimation of sparse changepoints, `inspect` performs relatively well even when $k = p$ (i.e. when the signal is highly non-sparse), especially when the signal strength is relatively large.

We now extend these ideas by investigating empirical performance under several other types of model misspecification. Recall that the noise matrix is $W = (W_{j,t}) := X - \boldsymbol{\mu}$ and we define $W_1, \ldots, W_n$ to be the column vectors of W. In models $\mathrm{M}_{\mathrm{unif}}$ and $\mathrm{M}_{\mathrm{exp}}$, we replace Gaussian noise by $W_{j,t} \overset{\text{iid}}{\sim} \mathrm{Unif}[-\sqrt{3}\sigma, \sqrt{3}\sigma]$ and $W_{j,t} \overset{\text{iid}}{\sim} \mathrm{Exp}(\sigma) - \sigma$ respectively. In model $\mathrm{M}_{\mathrm{cs,loc}}(\rho)$, we allow the noise to have a short-range cross-sectional dependence by sampling $W_1, \ldots, W_n \overset{\text{iid}}{\sim} N_p(0, \Sigma)$ for $\Sigma := (\rho^{|j-j'|})_{j,j'}$. In model $\mathrm{M}_{\mathrm{cs}}(\rho)$, we extend this to global cross-sectional dependence by sampling $W_1, \ldots, W_n \overset{\text{iid}}{\sim} N_p(0, \Sigma)$ for $\Sigma := (1-\rho)I_p + \frac{\rho}{p}\mathbf{1}_{p\times p}$, where $\mathbf{1}_{p\times p}$ is a $p \times p$ all-one matrix. In model $\mathrm{M}_{\mathrm{temp}}(\rho)$, we consider an auto-regressive AR(1) temporal dependence in the noise by first sampling $W'_{j,t} \overset{\text{iid}}{\sim} N(0, \sigma^2)$ and then setting $W_{j,1} := W'_{j,1}$ and $W_{j,t} := \rho^{1/2} W_{j,t-1} + (1-\rho)^{1/2} W'_{j,t}$ for $2 \le t \le n$. We report the performance of the different algorithms in the parameter setting $n = 2000$, $p = 1000$, $k = 32$, $z = 800$, $\vartheta = 0.25$, $\sigma^2 = 1$ in Table 3. It can be seen that `inspect` is robust to both temporal and spatial dependence structures, as well as noise misspecification.

5.2 Multiple changepoint estimation

The use of the 'burn-off' parameter β in Algorithm 4 was mainly to facilitate our theoretical analysis. In our simulations, we found that taking $\beta = 0$ rarely resulted in the changepoint being estimated more than once, and we therefore recommend setting $\beta = 0$ in practice, unless prior knowledge of the distribution of the changepoints suggests otherwise. To choose

n	p	k	z	ϑ	inspect	dc	sbs	scan
1000	200	10	400	0.18	**32.3**	82.2	99.6	46.2
1000	200	14	400	0.11	**97.2**	274.5	215.7	218.1
1000	200	200	400	0.04	**65.5**	262.3	180.1	156.4
1000	500	10	400	0.18	**48.2**	125.7	181.4	106.1
1000	500	22	400	0.11	**86.9**	240.5	235.5	190.3
1000	500	500	400	0.04	24.5	106.4	96.8	**22.5**
1000	1000	10	400	0.18	**48.6**	118.6	185.4	149.4
1000	1000	32	400	0.11	**58.7**	143.9	171.4	151.3
1000	1000	1000	400	0.04	**10.1**	28.1	42.7	15.1
2000	200	10	800	0.11	**126.3**	327.5	293.9	221.1
2000	200	14	800	0.11	**88.1**	213.7	155.2	121.0
2000	200	200	800	0.04	**57.6**	221.3	155.1	60.9
2000	500	10	800	0.11	**169.9**	348.1	456.0	305.5
2000	500	22	800	0.07	**195.2**	578.4	511.8	535.9
2000	500	500	800	0.04	**21.3**	45.0	62.4	27.0
2000	1000	10	800	0.11	**131.5**	416.4	460.5	397.7
2000	1000	32	800	0.07	**138.4**	441.0	448.6	401.6
2000	1000	1000	800	0.04	**6.7**	30.8	33.7	13.8
1000	200	10	400	0.4	**4.1**	8.4	16.9	5.1
1000	200	14	400	0.25	**7.4**	16.7	31.6	9.4
1000	200	200	400	0.11	**4.4**	15.8	12.7	6.2
1000	500	10	400	0.4	**2.9**	8.9	27.5	4.5
1000	500	22	400	0.25	**4.7**	13.0	20.0	7.2
1000	500	500	400	0.07	**3.7**	13.0	22.0	8.4
1000	1000	10	400	0.4	**3.1**	10.8	30.0	6.1
1000	1000	32	400	0.25	**3.0**	12.0	20.3	6.8
1000	1000	1000	400	0.07	**1.9**	10.1	12.0	4.1
2000	200	10	800	0.25	**7.8**	23.9	45.2	11.6
2000	200	14	800	0.18	**12.1**	44.2	47.7	20.5
2000	200	200	800	0.07	**7.6**	41.6	33.1	17.7
2000	500	10	800	0.25	14.3	28.6	54.4	**14.1**
2000	500	22	800	0.18	**14.5**	33.7	35.4	15.8
2000	500	500	800	0.07	**4.8**	16.2	17.6	9.2
2000	1000	10	800	0.25	**10.5**	29.7	68.3	16.1
2000	1000	32	800	0.18	**6.8**	19.3	39.8	12.6
2000	1000	1000	800	0.07	**1.4**	7.7	13.1	4.5

Table 2: Root mean squared error for `inspect`, `dc`, `sbs` and `scan` in single changepoint estimation. The smallest root mean squared error is given in bold.

Model	n	p	k	z	ϑ	`inspect`	`dc`	`sbs`	`scan`
$\mathrm{M}_{\mathrm{unif}}$	2000	1000	32	800	0.25	**3.0**	13.8	17.6	3.8
$\mathrm{M}_{\mathrm{exp}}$	2000	1000	32	800	0.25	**2.8**	11.9	47.7	5.5
$\mathrm{M}_{\mathrm{cs,loc}}(0.2)$	2000	1000	32	800	0.25	**3.4**	8.4	17.5	6.8
$\mathrm{M}_{\mathrm{cs,loc}}(0.5)$	2000	1000	32	800	0.25	**5.6**	10.8	23.7	8.4
$\mathrm{M}_{\mathrm{cs}}(0.5)$	2000	1000	32	800	0.25	**1.5**	7.5	14.2	3.5
$\mathrm{M}_{\mathrm{cs}}(0.9)$	2000	1000	32	800	0.25	**2.5**	6.5	10.2	2.9
$\mathrm{M}_{\mathrm{temp}}(0.1)$	2000	1000	32	800	0.25	**4.0**	16.9	96.2	10.1
$\mathrm{M}_{\mathrm{temp}}(0.3)$	2000	1000	32	800	0.25	**14.5**	24.9	226.4	14.7

Table 3: Root mean squared error for `inspect`, `dc`, `sbs` and `scan` in single changepoint estimation, under different forms of model misspecification.

ξ in the multiple changepoint estimation simulation studies, for each (n, p), we first applied `inspect` to 1000 data sets drawn from the null model with no changepoint, and took ξ to be the largest value of $\bar{T}_{\max}$ from Algorithm 2. We also set $Q = 1000$.

We consider the simulation setting where $n = 2000$, $p = 200$, $k = 40$, $\sigma^2 = 1$ and $z = (500, 1000, 1500)$. Define $\vartheta^{(i)} := \|\theta^{(i)}\|_2 / \|\theta^{(i)}\|_0^{1/2}$ to be the signal strength at the ith changepoint. We set $(\vartheta^{(1)}, \vartheta^{(2)}, \vartheta^{(3)}) = \vartheta(1, 1.5, 2)$ and let ϑ vary to see the performance of the algorithms at different signal strengths. We also considered different levels of overlap between the coordinates in which the three changes in mean structure occur: in the *complete overlap* case, changes occur in the same k coordinates at each changepoint; in the *half overlap* case, the changes occur in coordinates $\frac{i-1}{2}k+1, \ldots, \frac{i+1}{2}k$ for $i = 1, 2, 3$; in the *no overlap* case, the changes occur in disjoint sets of coordinates. Table 4 summarises the results. We report both the frequency counts of the number of changepoints detected over 100 runs and two quality measures of the location of changepoints. In particular, since changepoint estimation can be viewed as a special case of classification, the quality of the estimated changepoints can be measured by the Adjusted Rand Index (ARI) of the estimated segmentation against the truth (Rand, 1971; Hubert and Arabie, 1985). We report both the ARI and the percentage of runs for which a particular method attains the largest ARI among the four. Figure 3 gives a pictorial representation of the results for one particular collection of parameter settings. Again, we find that the performance of `inspect` is very encouraging on all performance measures.

A Appendix: Proofs of main results

Proof of Proposition 1. We note that the matrix A as defined in Section 3 has rank 1, and its only non-zero singular value is $\|\theta\|_2\|\gamma\|_2$. By Proposition 6 in Appendix B, on the event $\Omega_1 := \{\|E\|_\infty \leq \lambda\}$, we have

$$\sin \angle(\hat{v}, v) \leq \frac{8\lambda\sqrt{kn}}{\|\theta\|_2\|\gamma\|_2}.$$

$(\vartheta^{(1)}, \vartheta^{(2)}, \vartheta^{(3)})$	method	$\hat{\nu}$						ARI	% best
		0	1	2	3	4	5		
$(0.10, 0.15, 0.20)$	`inspect`	0	0	8	65	27	0	0.90	41
	`dc`	0	0	37	61	2	0	0.84	19
	`sbs`	0	0	3	62	30	5	0.88	18
	`scan`	0	0	63	35	2	0	0.80	22
$(0.08, 0.12, 0.16)$	`inspect`	0	0	39	50	11	0	0.78	41
	`dc`	0	1	74	24	1	0	0.73	25
	`sbs`	0	0	34	48	15	3	0.75	18
	`scan`	0	1	95	4	0	0	0.70	18
$(0.06, 0.09, 0.12)$	`inspect`	0	6	61	28	5	0	0.66	40
	`dc`	0	26	72	2	0	0	0.56	18
	`sbs`	0	9	65	27	4	0	0.63	25
	`scan`	0	11	88	1	0	0	0.68	19
$(0.10, 0.15, 0.20)$	`inspect`	0	0	10	73	14	3	0.91	46
	`dc`	0	0	23	63	13	1	0.86	14
	`sbs`	0	0	6	69	22	3	0.85	24
	`scan`	0	0	65	33	2	0	0.79	16
$(0.08, 0.12, 0.16)$	`inspect`	0	0	23	50	22	5	0.82	52
	`dc`	0	0	47	40	12	1	0.76	22
	`sbs`	0	0	30	48	14	8	0.77	20
	`scan`	0	0	94	6	0	0	0.71	7
$(0.06, 0.09, 0.12)$	`inspect`	0	0	48	42	10	0	0.77	55
	`dc`	0	7	66	23	4	0	0.69	18
	`sbs`	0	0	58	36	6	0	0.70	14
	`scan`	0	11	88	1	0	0	0.68	26
$(0.10, 0.15, 0.20)$	`inspect`	0	0	10	74	15	1	0.92	56
	`dc`	0	0	37	57	6	0	0.81	12
	`sbs`	0	0	2	68	28	2	0.86	18
	`scan`	0	0	63	35	2	0	0.78	17
$(0.08, 0.12, 0.16)$	`inspect`	0	0	38	54	8	0	0.81	54
	`dc`	0	0	73	26	1	0	0.69	12
	`sbs`	0	0	26	60	14	0	0.76	26
	`scan`	0	1	89	10	0	0	0.70	9
$(0.06, 0.09, 0.12)$	`inspect`	0	1	66	31	2	0	0.71	52
	`dc`	0	12	78	10	0	0	0.62	17
	`sbs`	0	1	60	30	8	1	0.66	24
	`scan`	0	21	77	2	0	0	0.61	14

Table 4: Multiple changepoint simulation results. The top, middle and bottom blocks refer to the complete, half and no overlap settings respectively. Other simulation parameters: $n = 2000$, $p = 200$, $k = 40$, $z = (500, 1000, 1500)$ and $\sigma^2 = 1$.

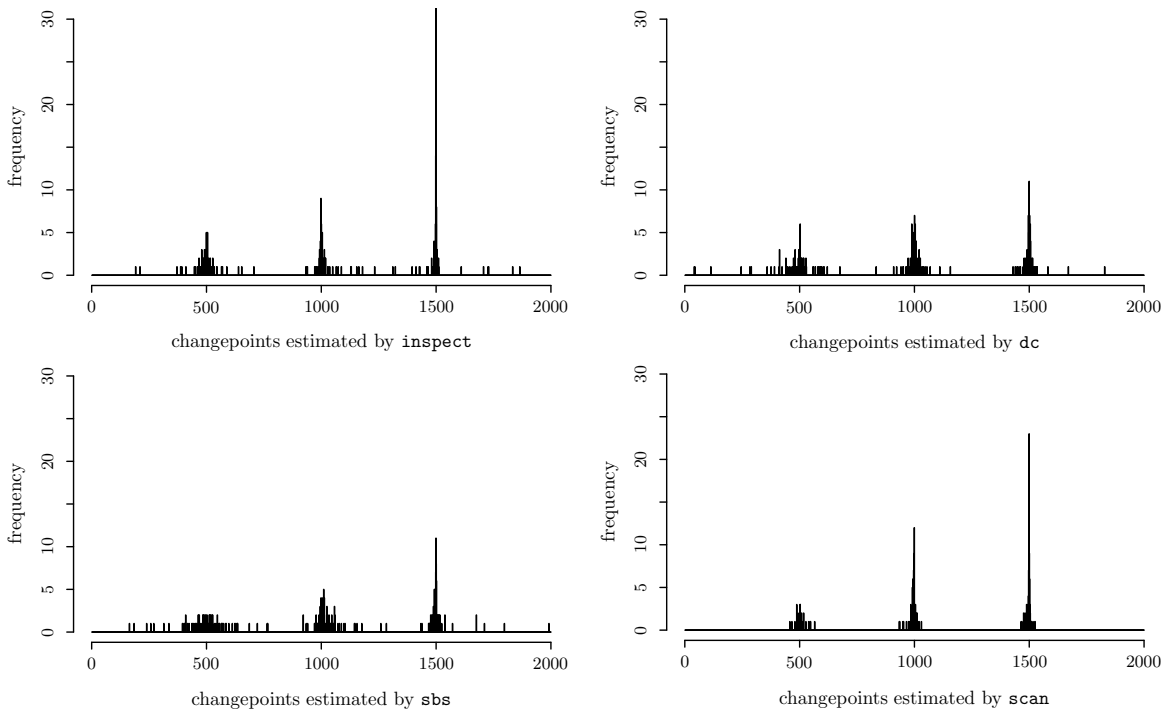

Figure 3: Histograms of estimated changepoint locations by `inspect` (top-left), `dc` (top-right), `sbs` (bottom-left) and `scan` (bottom-right) in the half overlap case. Parameter settings: $n = 2000$, $p = 200$, $k = 40$, $z = (500, 1000, 1500)$, $(\vartheta^{(1)}, \vartheta^{(2)}, \vartheta^{(3)}) = (0.10, 0.15, 0.20)$, $\sigma^2 = 1$.

By definition, $\|\theta\|_2 \geq \sqrt{k}\vartheta$, and by Lemma 8 in Appendix B, $\|\gamma\|_2 \geq \frac{1}{4}n\tau$. Thus, $\sin\angle(\hat{v}, v) \leq \frac{32\lambda}{\vartheta\tau\sqrt{n}}$ on Ω_1. It remains to verify that $\mathbb{P}(\Omega_1^c) \leq 4(p\log n)^{-1/2}$ for $n \geq 6$. By Lemma 5,

$$\begin{aligned}\mathbb{P}\big(\|E\|_\infty \geq 2\sigma\sqrt{\log(p\log n)}\big) &\leq 2\sqrt{\frac{2}{\pi}}p\lceil\log n\rceil\sqrt{\log(p\log n)}\Big\{1 + \frac{1}{\log(p\log n)}\Big\}(p\log n)^{-2} \\ &\leq 6(p\log n)^{-1}\sqrt{\log(p\log n)} \leq 4(p\log n)^{-1/2}, \end{aligned} \tag{14}$$

as desired. □

Proof of Theorem 2. Recall the definition of $X^{(2)}$ in (11) and the definition $T^{(2)} := \mathcal{T}(X^{(2)})$. Define similarly $\boldsymbol{\mu}^{(2)} = (\mu_1^{(2)}, \ldots, \mu_{n_1}^{(2)}) \in \mathbb{R}^{p\times n_1}$ and a random matrix $W^{(2)} = (W_1^{(2)}, \ldots, W_{n_1}^{(2)})$ taking values in $\mathbb{R}^{p\times n_1}$ by $\mu_t^{(2)} := \mu_{2t}$ and $W_t^{(2)} = W_{2t}$; now let $A^{(2)} := \mathcal{T}(\boldsymbol{\mu}^{(2)})$ and $E^{(2)} := \mathcal{T}(W^{(2)})$. Furthermore, we write $\bar{X} := (\hat{v}^{(1)})^\top X^{(2)}$, $\bar{\mu} := (\hat{v}^{(1)})^\top \boldsymbol{\mu}^{(2)}$, $\bar{W} := (\hat{v}^{(1)})^\top W^{(2)}$, $\bar{T} := (\hat{v}^{(1)})^\top T^{(2)}$, $\bar{A} := (\hat{v}^{(1)})^\top A^{(2)}$ and $\bar{E} := (\hat{v}^{(1)})^\top E^{(2)}$ for the one-dimensional projected images (as row vectors) of the corresponding p-dimensional quantities. We note that $\bar{T} = \mathcal{T}(\bar{X})$, $\bar{A} = \mathcal{T}(\bar{\mu})$ and $\bar{E} = \mathcal{T}(\bar{W})$.

Now, conditional on $\hat{v}^{(1)}$, the random variables $\bar{X}_1, \ldots, \bar{X}_{n_1}$ are independent, with

$$\bar{X}_t \mid \hat{v}^{(1)} \sim N(\bar{\mu}_t, \sigma^2),$$

and the row vector $\bar{\mu}$ undergoes a single change at $z^{(2)} := z/2$ with magnitude of change

$$\bar{\theta} := \bar{\mu}_{z^{(2)}+1} - \bar{\mu}_{z^{(2)}} = (\hat{v}^{(1)})^\top\theta.$$

Finally, let $\hat{z}^{(2)} \in \operatorname{argmax}_{1\leq t\leq n_1-1}|\bar{T}_t|$, so we may assume the first component of the output of the algorithm is $\hat{z} = 2\hat{z}^{(2)}$. Consider the set

$$\Upsilon := \{\tilde{v} \in \mathbb{S}^{p-1} : \angle(\tilde{v}, v) \leq \pi/3\}.$$

By condition (12) in the statement of the theorem and Proposition 1,

$$\mathbb{P}(\hat{v}^{(1)} \in \Upsilon) \geq 1 - 4(p\log n_1)^{-1/2}. \tag{15}$$

Note that $\hat{v}^{(1)}$ and $W^{(2)}$ are independent, so $\bar{W}$ has independent $N(0, \sigma^2)$ entries. Hence, by Lemma 5 in Appendix B,

$$\mathbb{P}\big(\|\bar{E}\|_\infty \geq 2\sigma\sqrt{\log n_1}\big) \leq \sqrt{\frac{2}{\pi}}\lceil\log n_1\rceil\left(2\sqrt{\log n_1} + \frac{1}{\sqrt{\log n_1}}\right)n_1^{-2} \leq n_1^{-1}. \tag{16}$$

Since $\bar{T} = \bar{A} + \bar{E}$, and since $(\bar{A}_t)_t$ and $(\bar{T}_t)_t$ are respectively maximised at $t = z^{(2)}$ and $t = \hat{z}^{(2)}$, we have on the event $\Omega_0 := \big\{\hat{v}^{(1)} \in \Upsilon, \|\bar{E}\|_\infty \leq 2\sigma\sqrt{\log n_1}\big\}$ that

$$\begin{aligned}\bar{A}_{z^{(2)}} - \bar{A}_{\hat{z}^{(2)}} &= (\bar{A}_{z^{(2)}} - \bar{T}_{z^{(2)}}) + (\bar{T}_{z^{(2)}} - \bar{T}_{\hat{z}^{(2)}}) + (\bar{T}_{\hat{z}^{(2)}} - \bar{A}_{\hat{z}^{(2)}}) \\ &\leq |\bar{A}_{z^{(2)}} - \bar{T}_{z^{(2)}}| + |\bar{T}_{\hat{z}^{(2)}} - \bar{A}_{\hat{z}^{(2)}}| \leq 4\sigma\sqrt{\log n_1}.\end{aligned}$$

The row vector $\bar{A}$ has the following explicit form

$$\bar{A}_t = \begin{cases} \sqrt{\frac{t}{n_1(n_1-t)}}(n_1 - z^{(2)})\bar{\theta}, & \text{if } t \le z^{(2)} \\ \sqrt{\frac{n_1-t}{n_1 t}} z^{(2)}\bar{\theta}, & \text{if } t > z^{(2)}. \end{cases}$$

Hence, by Lemma 11, on the event Ω_0,

$$\hat{z}^{(2)} \in \big[z^{(2)} - 2\Delta z^{(2)}, z^{(2)} + 2\Delta(n_1 - z^{(2)})\big],$$

where $\Delta := 4\sigma\bar{\theta}^{-1}\sqrt{\frac{n_1 \log n_1}{z^{(2)}(n_1 - z^{(2)})}}$. Since $\hat{z} = 2\hat{z}^{(2)}$ and $z = 2z^{(2)}$, we have that on Ω_0,

$$\frac{1}{n}|\hat{z} - z| \le 2\Delta \le \frac{8\sqrt{2}\sigma}{\bar{\theta}}\sqrt{\frac{n\log n}{z(n-z)}} \le \frac{16\sigma}{\bar{\theta}\sqrt{\tau}}\sqrt{\frac{\log n}{n}}. \tag{17}$$

On the event Ω_0, we have $\bar{\theta} \ge \sqrt{k}\vartheta/2$. We deduce from (15), (16) and (17), we find that

$$\mathbb{P}\left\{\frac{1}{n}|\hat{z} - z| > \frac{32\sigma}{\vartheta\sqrt{k\tau}}\sqrt{\frac{\log n}{n}}\right\} \le 4\left\{p\log\left(\frac{n}{2}\right)\right\}^{-1/2} + \frac{2}{n},$$

as desired. □

Proof of Theorem 3. For $i \in \{0, 1, \ldots, \nu\}$, we define $J_i := \big[z_i + \lceil\frac{z_{i+1}-z_i}{3}\rceil, z_{i+1} - \lceil\frac{z_{i+1}-z_i}{3}\rceil\big]$ and

$$\Omega_1 := \bigcap_{i=1}^{\nu}\bigcup_{q=1}^{Q}\{s_q \in J_{i-1}, e_q \in J_i\}.$$

By a union bound, we have

$$\begin{aligned}\mathbb{P}(\Omega_1^c) &\le \nu\left(1 - \frac{(z_i - z_{i-1} - 2\lceil\frac{z_i - z_{i-1}}{3}\rceil)(z_{i+1} - z_i - 2\lceil\frac{z_{i+1}-z_i}{3}\rceil)}{n(n+1)/2}\right)^Q \\ &\le \nu\left(1 - \frac{(z_i - z_{i-1})(z_{i+1} - z_i)}{9n^2}\right)^Q \le \tau^{-1}(1 - \tau^2/9)^Q \le \tau^{-1}e^{-\tau^2 Q/9},\end{aligned}$$

where the second inequality uses the fact that $n\tau \ge 14$. For any matrix $M \in \mathbb{R}^{p\times n}$ and $1 \le \ell \le r \le n$, we write $M^{[\ell,r]}$ for the submatrix obtained by extracting columns $\{\ell, \ell+1, \ldots, r\}$ of M. Also define $\boldsymbol{\mu}' := \mathbb{E}X' = \boldsymbol{\mu}$ and $W' := X' - \boldsymbol{\mu}'$. Let $\hat{v}^{[\ell,r]}$ be a leading left singular vector of a maximiser of

$$M \mapsto \langle \mathcal{T}(X'^{[\ell,r]}), M\rangle - \lambda\|M\|_1,$$

for $M \in \mathcal{S}$, where $\mathcal{S} = \mathcal{S}_1$ or $\mathcal{S}_2$. For definiteness, we assume both the maximiser and its leading left singular vector of it are chosen to be the lexicographically smallest possibilities. For $q = 1, \ldots, Q$, we also write $M^{[q]}$ for $M^{[s_q+1, e_q]}$ and $\hat{v}^{[q]}$ for $\hat{v}^{[s_q+1,e_q]}$. Define events

$$\begin{aligned}\Omega_2 &:= \bigcap_{1\le\ell<r\le n}\{\|\mathcal{T}(W'^{[\ell,r]})\|_\infty \le \lambda\}, \\ \Omega_3 &:= \bigcap_{1\le\ell<r\le n}\{\|(\hat{v}^{[\ell,r]})^\top\mathcal{T}(W^{[\ell,r]})\|_\infty \le \lambda\}.\end{aligned}$$

By Lemma 5,

$$\mathbb{P}(\Omega_2^c) \le \binom{n}{2}\sqrt{\frac{2}{\pi}}p\lceil\log n\rceil\left(3\sqrt{\log(np)}+\frac{2}{3\sqrt{\log(np)}}\right)(np)^{-9/2} \le n^{-3/2}p^{-5/2}.$$

Also, since $\hat{v}^{[\ell,r]}$ and X are independent, $(\hat{v}^{[\ell,r]})^\top\mathcal{T}(W)$ has the same distribution as $\mathcal{T}(G)$, where G is a row vector of length $r-\ell+1$ with independent $N(0,\sigma^2)$ entries. So by Lemma 5 again, for sufficiently large n,

$$\mathbb{P}(\Omega_3^c) \le \binom{n}{2}\mathbb{P}\big\{\|\mathcal{T}(G)\|_\infty > \lambda\big\} \le n^{-3/2}p^{-5/2}.$$

We claim that the desired event $\Omega^* := \{\hat{\nu}=\nu \text{ and } |\hat{z}_i - z_i| \le n\rho \text{ for all } 1\le i\le\nu\}$ occurs if the following two statements hold every time the function `wbs` is called in Algorithm 4′:

(i) There exist unique $i_1, i_2 \in \{0,1,\dots,\nu+1\}$ such that $|s-z_{i_1}| \le n\rho$ and $|e-z_{i_2}| \le n\rho$, where (s,e) is the pair of arguments of the `wbs` function call.
(ii) $\bar{T}^{[q_0]}_{\max} > \xi$ if and only if $i_2 - i_1 \ge 2$, where i_1 and i_2 are the indices defined in (i).

To see this, observe that the set of all arguments used in the calls of the function `wbs` is $\hat{Z}\cup\{0,n\}$, so (i) ensures that

$$\max_{\hat{z}\in\hat{Z}\cup\{0,n\}}\min_{i\in\{0,1,\dots,\nu+1\}}|\hat{z}-z_i| \le n\rho.$$

If $|\hat{z}-z_i| \le n\rho$, we say $\hat{z}$ is 'identified' to z_i. Moreover, each candidate changepoint b identified by the function call `wbs`(s,e) in Algorithm 4′ satisfies $\min\{b-s, e-b\} \ge n\beta > 2n\rho$. It follows that different elements of $\hat{Z}\cup\{0,n\}$ cannot be identified to the same z_i, so no element of $\hat{Z}$ is identified to z_0 or $z_{\nu+1}$, and the second part of the event Ω^* holds. It remains to show that each element of $\{z_1,\dots,z_\nu\}$ is identified by some element of $\hat{Z}$. To see this, note that if z_i is not identified, we can let (s^*,e^*) be the shortest interval such that $s^*+1 \le z_i \le e^*$ and such that (s^*,e^*) are a pair of arguments called by the `wbs` function in Algorithm 4′. By (i), the two endpoints s^* and e^* are identified to z_{i_1} and z_{i_2} respectively, say, for some $i_1 \le i-1$ and $i_2 \ge i+1$. But then by (ii) a new point b will be added to $\hat{Z}$ and the recursion continues on the pairs (s^*,b) and (b,e^*), contradicting the minimality of the pair (s^*,e^*).

We now prove by induction on the depth of the recursion that on $\Omega_1\cap\Omega_2\cap\Omega_3$, statements (i) and (ii) hold every time `wbs` is called in Algorithm 4′. The first time `wbs` is called, $s=0$ and $e=n$, so (i) is satisfied with the unique choice $i_1=0$ and $i_2=\nu+1$. This proves the base case. Now suppose `wbs` is called with the pair (s,e) satisfying (i), yielding indices $i_1,i_2\in\{0,1,\dots,\nu+1\}$ with $|s-z_{i_1}|\le n\rho$, $|e-z_{i_2}|\le n\rho$. To complete the inductive step, we need to show that (ii) also holds, and if a new changepoint b is detected, then (i) holds for the pairs of arguments (s,b) and (b,e). We have two cases.

Case 1: $i_2 - i_1 = 1$. In this case, $(s+n\beta, e-n\beta]$ contains no changepoint. Since $\xi=\lambda$, on Ω_3 we always have

$$\bar{T}^{[q_0]}_{\max} = \max_{q\in\mathcal{Q}_{s,e}}\|(\hat{v}^{[q]})^\top\mathcal{T}(X^{[q]})\|_\infty \le \xi,$$

so (ii) is satisfied with no additional changepoint detected.

Case 2: $i_2 - i_1 \geq 2$. On the event Ω_1, for any $i^* \in \{i_1+1, \ldots, i_2-1\}$, there exists $q^* \in \{1, \ldots, Q\}$ such that $s_{q^*} \in J_{i^*-1}$ and $e_{q^*} \in J_{i^*}$. Moreover, since $\min\{s_{q^*} - s, e - e_{q^*}\} \geq \lceil n\tau/3 \rceil - n\rho > n\beta$ by the condition on β in the theorem, we have $q^* \in \mathcal{Q}_{s,e}$. Since there is precisely one changepoint within the segment $(s_{q^*}, e_{q^*}]$, the matrix $\mathcal{T}(\boldsymbol{\mu}'^{[q^*]})$ has rank 1; cf. (4). On Ω_2, we have $\|\mathcal{T}(W'^{[q^*]})\|_\infty \leq \lambda$. Thus, by Proposition 6 and Lemma 8 in Appendix B,

$$\sin \angle\big(\hat{v}^{[q^*]}, \theta^{(i^*)}/\|\theta^{(i^*)}\|_2\big) \leq \frac{8\lambda\sqrt{k(e_{q^*} - s_{q^*})}}{\|\theta^{(i^*)}\|_2 n\tau/12} \leq \frac{96\lambda}{\vartheta\tau\sqrt{n}} = \frac{96}{200}\rho\sqrt{k\tau} \leq \frac{12}{25}$$

under the conditions of the theorem. Therefore, recalling the definition of q_0 in Algorithm 4′, and on the event Ω_3,

$$\begin{aligned}
\bar{T}^{[q_0]}_{\max} \geq \bar{T}^{[q^*]}_{\max} &= \|(\hat{v}^{[q^*]})^\top \mathcal{T}(X^{[q^*]})\|_\infty \geq \|(\hat{v}^{[q^*]})^\top \mathcal{T}(\boldsymbol{\mu}^{[q^*]})\|_\infty - \|(\hat{v}^{[q^*]})^\top \mathcal{T}(W^{[q^*]})\|_\infty \\
&\geq \big|(\hat{v}^{[q^*]})^\top \theta^{(i^*)}\big| \sqrt{\frac{(z_{i^*} - s_{q^*})(e_{q^*} - z_{i^*})}{e_{q^*} - s_{q^*}}} - \lambda \\
&\geq \sqrt{1 - (12/25)^2}\|\theta^{(i^*)}\|_2 \sqrt{\frac{n\tau}{6}} - \lambda > 0.358\sqrt{n\tau}\|\theta^{(i^*)}\|_2 - \lambda > \left(\frac{71.6}{\rho\tau} - 1\right)\lambda > \xi. \quad (18)
\end{aligned}$$

Thus (ii) is satisfied with a new changepoint $b := s_{q_0} + \hat{z}^{[q_0]}$ detected. It remains to check that (i) holds for the pairs of arguments (s, b) and (b, e), for which it suffices to show that $\min_{1 \leq i \leq \nu} |b - z_i| \leq n\rho$. To this end, we study the behaviour of univariate CUSUM statistics of the projected series $(\hat{v}^{[q_0]})^\top X^{[q_0]}$. To simplify notation, we define $\bar{X} := (\hat{v}^{[q_0]})^\top X^{[q_0]}$, $\bar{\mu} := (\hat{v}^{[q_0]})^\top \boldsymbol{\mu}^{[q_0]}$, $\bar{W} := (\hat{v}^{[q_0]})^\top W^{[q_0]}$, $\bar{T} := \mathcal{T}(\bar{X})$, $\bar{A} := \mathcal{T}(\bar{\mu})$ and $\bar{E} := \mathcal{T}(\bar{W})$. The row vector $\bar{\mu} \in \mathbb{R}^{e_{q_0} - s_{q_0}}$ is piecewise constant with changepoints at $z_{i_1+1} - s_{q_0}, \ldots, z_{i_2-1} - s_{q_0}$. Recall that $\hat{z}^{[q_0]} \in \operatorname{argmax}_{1 \leq t \leq e_{q_0} - s_{q_0} - 1} |\bar{T}_t|$. We may assume that $\bar{T}_{\hat{z}^{[q_0]}} > 0$ (the case $\bar{T}_{\hat{z}^{[q_0]}} < 0$ can be handled similarly). On Ω_3,

$$\bar{A}_{\hat{z}^{[q_0]}} = \bar{T}_{\hat{z}^{[q_0]}} - \bar{E}_{\hat{z}^{[q_0]}} \geq \bar{T}^{[q_0]}_{\max} - \lambda \geq \left(\frac{71.6}{\rho\tau} - 2\right)\lambda > 0, \quad (19)$$

and in particular, there is at least one changepoint in $(s_{q_0}, e_{q_0}]$. We may assume that $\hat{z}^{[q_0]}$ is not equal to $z_i - s_{q_0}$ for any $i_1 + 1 \leq i \leq i_2 - 1$, since otherwise $\min_{1 \leq i \leq \nu} |b - z_i| = 0$ and we are done. By Lemma 12 and after possibly reflecting the time direction, we may also assume that there is at least one changepoint to the left of $\hat{z}^{[q_0]}$, and that if $z_{i_0} - s_{q_0}$ is the changepoint immediately left of $\hat{z}^{[q_0]}$, then the series $\{\bar{A}_t : z_{i_0} - s_{q_0} \leq t \leq \hat{z}^{[q_0]}\}$ is positive and strictly decreasing. It therefore follows by (19), and (18) with i_0 in place of i^*, that on Ω_3,

$$\bar{A}_{z_{i_0} - s_{q_0}} \geq \bar{A}_{\hat{z}^{[q_0]}} \geq \bar{T}^{[q_0]}_{\max} - \lambda \geq 0.358\sqrt{n\tau}\|\theta^{(i_0)}\|_2 - 2\lambda > \left(0.358 - \frac{\rho\tau}{100}\right)\sqrt{n\tau}\|\theta^{(i_0)}\|_2 > \frac{71.4\lambda}{\rho\tau}, \quad (20)$$

where we use $9\rho < \tau \leq 1/2$ in the final inequality. On the other hand, on the event Ω_3, by the maximality of $\bar{T}_{\hat{z}^{[q_0]}}$, we have that

$$\bar{A}_{\hat{z}^{[q_0]}} \geq \bar{T}_{\hat{z}^{[q_0]}} - \lambda \geq \bar{T}_{z_{i_0} - s_{q_0}} - \lambda \geq \bar{A}_{z_{i_0} - s_{q_0}} - 2\lambda \geq \left(1 - \frac{\rho\tau}{35.7}\right)\bar{A}_{z_{i_0} - s_{q_0}}. \quad (21)$$

Our strategy here is to characterise the rate of decay of the series $\{\bar{A}_t : z_{i_0} - s_{q_0} \leq t \leq \hat{z}^{[q_0]}\}$ from its left endpoint, so that we can conclude from (21) that $\hat{z}^{[q_0]}$ is close to $z_{i_0} - s_{q_0}$. This is achieved by considering the following three cases: (a) there is no changepoint to the right of $\hat{z}^{[q_0]}$, i.e. $z_{i_0+1} \geq e_{q_0}$; (b) $z_{i_0+1} \leq e_{q_0} - 1$ and $\bar{A}_{z_{i_0}} \geq \bar{A}_{z_{i_0+1}}$; (c) $z_{i_0+1} \leq e_{q_0} - 1$ and $\bar{A}_{z_{i_0}} < \bar{A}_{z_{i_0+1}}$.

In case (a), we apply Lemma 11 with $e_{q_0} - s_{q_0}$ and $z_{i_0} - s_{q_0}$ taking the roles of n and z in the lemma respectively, while noting that

$$\left|(\hat{v}^{[q_0]})^\top \theta^{(i_0)}\right|^{-1} \bar{A}_t = \sqrt{\frac{e_{q_0} - s_{q_0} - t}{(e_{q_0} - s_{q_0})t}}(z_{i_0} - s_{q_0}) \qquad \forall\ z_{i_0} - s_{q_0} \leq t \leq e_{q_0} - s_{q_0}$$

takes the same form as the function f in the corresponding range in Lemma 11. Thus, we conclude from (21) and Lemma 11 that

$$\hat{z}^{[q_0]} - (z_{i_0} - s_{q_0}) \leq 2(e_{q_0} - z_{i_0})\frac{\rho\tau}{35.7} \leq n\rho,$$

as desired.

For case (b), we define $\tilde{\mu} := \frac{1}{e_{q_0} - s_{q_0}} \sum_{t=1}^{e_{q_0} - s_{q_0}} \bar{\mu}_t$ to be the overall average of the $\bar{\mu}$ series, and let

$$\tilde{\mu}_{\mathrm{L}} := \frac{1}{z_{i_0} - s_{q_0}} \sum_{t=1}^{z_{i_0} - s_{q_0}} \bar{\mu}_t - \tilde{\mu}, \quad \tilde{\mu}_{\mathrm{M}} := \bar{\mu}_{z_{i_0}+1-s_{q_0}} - \tilde{\mu} \text{ and } \tilde{\mu}_{\mathrm{R}} := \frac{1}{e_{q_0} - z_{i_0+1}} \sum_{t=z_{i_0+1}-s_{q_0}+1}^{e_{q_0}-s_{q_0}} \bar{\mu}_t - \tilde{\mu}$$

be the centred averages of the $\bar{\mu}$ series on the segments $(0, z_{i_0} - s_{q_0}]$, $(z_{i_0} - s_{q_0}, z_{i_0+1} - s_{q_0}]$ and $(z_{i_0+1} - s_{q_0}, e_{q_0} - s_{q_0}]$ respectively. Using (3), we have that for $z_{i_0} - s_{q_0} \leq t \leq z_{i_0+1} - s_{q_0}$,

$$\bar{A}_t = [\mathcal{T}(\bar{\mu})]_t = \sqrt{\frac{e_{q_0} - s_{q_0}}{t(e_{q_0} - s_{q_0} - t)}}\Big\{(z_{i_0} - s_{q_0})(-\tilde{\mu}_{\mathrm{L}}) + (t - z_{i_0} + s_{q_0})(-\tilde{\mu}_{\mathrm{M}})\Big\}. \tag{22}$$

We claim that $z_{i_0} - s_{q_0} \geq n\tau/15$. For, if not, then in particular, $z_{i_0-1} < s_{q_0}$ and $\tilde{\mu}_{\mathrm{L}} = \bar{\mu}_{z_{i_0}-s_{q_0}} - \tilde{\mu}$. By (22) and the fact that $\bar{A}_{z_{i_0}-s_{q_0}} > 0$, we have $\tilde{\mu}_{\mathrm{L}} < 0$. It follows from (20) that

$$\begin{aligned}
0.357\sqrt{n\tau}(\tilde{\mu}_{\mathrm{M}} - \tilde{\mu}_{\mathrm{L}}) &\leq \left(0.358 - \frac{\rho\tau}{100}\right)\sqrt{n\tau}|(\hat{v}^{[q_0]})^\top \theta^{(i_0)}| \leq \bar{A}_{z_{i_0}-s_{q_0}} \\
&= \sqrt{\frac{(e_{q_0} - s_{q_0})(z_{i_0} - s_{q_0})}{e_{q_0} - z_{i_0}}}(-\tilde{\mu}_{\mathrm{L}}) \\
&\leq \sqrt{\frac{n\tau + z_{i_0} - s_{q_0}}{n\tau}}\sqrt{z_{i_0} - s_{q_0}}(-\tilde{\mu}_{\mathrm{L}}) \leq \frac{4\sqrt{n\tau}}{15}(-\tilde{\mu}_{\mathrm{L}}),
\end{aligned}$$

which can be rearranged to give $-\tilde{\mu}_{\mathrm{M}} > 0.25(-\tilde{\mu}_{\mathrm{L}})$. Consequently,

$$\begin{aligned}
\bar{A}_{z_{i_0+1}-s_{q_0}} &= \sqrt{\frac{e_{q_0}-s_{q_0}}{(z_{i_0+1}-s_{q_0})(e_{q_0}-z_{i_0+1})}}\Big\{(-\tilde{\mu}_{\mathrm{L}})(z_{i_0}-s_{q_0})+(-\tilde{\mu}_{\mathrm{M}})(z_{i_0+1}-z_{i_0})\Big\} \\
&> \sqrt{\frac{e_{q_0}-s_{q_0}}{(z_{i_0+1}-s_{q_0})(e_{q_0}-z_{i_0+1})}}\Big\{(-\tilde{\mu}_{\mathrm{L}})(z_{i_0}-s_{q_0})+0.25(-\tilde{\mu}_{\mathrm{L}})(z_{i_0+1}-z_{i_0})\Big\} \\
&\geq 0.25\sqrt{\frac{(e_{q_0}-s_{q_0})(z_{i_0+1}-s_{q_0})}{e_{q_0}-z_{i_0+1}}}(-\tilde{\mu}_{\mathrm{L}}) \\
&\geq 0.25\bar{A}_{z_{i_0}-s_{q_0}}\sqrt{\frac{z_{i_0+1}-s_{q_0}}{z_{i_0}-s_{q_0}}} \geq \bar{A}_{z_{i_0}-s_{q_0}},
\end{aligned}$$

contradicting the assumption of case (b). Hence we have established the claim. We can then apply Lemma 13 with $\bar{A}_t$, $e_{q_0}-s_{q_0}$, $z_{i_0}-s_{q_0}$, $z_{i_0+1}-s_{q_0}$, $-\tilde{\mu}_{\mathrm{L}}$, $-\tilde{\mu}_{\mathrm{M}}$ and $\tau/15$ taking the roles of $g(t)$, n, z, z', μ_0, μ_1 and τ in the lemma respectively. By (21) and this lemma, we conclude that

$$\hat{z}^{[q_0]} - (z_{i_0}-s_{q_0}) \leq \frac{\rho\tau\bar{A}_{z_{i_0}-s_{q_0}}/35.7}{0.52\bar{A}_{z_{i_0}-s_{q_0}}n^{-1}\tau/15} \leq n\rho.$$

For case (c), by Lemma 12, the series $(\bar{A}_t : z_{i_0}-s_{q_0} \leq t \leq z_{i_0+1}-s_{q_0})$ must be strictly decreasing, then strictly increasing, while staying positive throughout. Define $\zeta := \max\{t \in [z_{i_0}-s_{q_0}, z_{i_0+1}-s_{q_0}] : \bar{A}_t \leq \bar{A}_{z_{i_0+1}-s_{q_0}}-2\lambda\}$. Using a very similar argument to that in case (b), we find that $e_{q_0}-z_{i_0+1} \geq n\tau/15$, and therefore by Lemma 13 again, $z_{i_0+1}-s_{q_0}-(\zeta+1) \leq n\rho$. Now on Ω_3, we have $\bar{A}_{z_{i_0}-s_{q_0}} > \bar{A}_{\hat{z}^{[q_0]}} > \bar{A}_{z_{i_0+1}-s_{q_0}}-2\lambda \geq \bar{A}_\zeta$ and $\zeta-(z_{i_0}-s_{q_0}) \geq n\tau-n\rho-1$. So we can apply the same argument as in case (b) with ζ taking the role of z_{i_0+1} and $\tau-\rho-1/n$ in place of τ, and obtain that

$$\hat{z}^{[q_0]} - (z_{i_0}-s_{q_0}) \leq \frac{\rho\tau\bar{A}_{z_{i_0}-s_{q_0}}/35.7}{0.52\bar{A}_{z_{i_0}-s_{q_0}}n^{-1}(\tau-\rho-1/n)/15} \leq \frac{15/0.52}{35.7(1-1/9-1/14)}n\rho \leq n\rho,$$

as desired. □

B Appendix: Auxiliary results

The following result is a generalisation of the curvature lemma of Vu et al. (2013, Lemma 3.1).

Lemma 4. *Let $v \in \mathbb{S}^{p-1}$ and $u \in \mathbb{S}^{n-1}$ be the leading left and right singular vectors of $A \in \mathbb{R}^{p\times n}$ respectively. Suppose that the first and second largest singular values of A are separated by $\delta > 0$. Let $M \in \mathbb{R}^{p\times n}$. If either of the following two conditions holds,*

(a) $\mathrm{rank}(A) = 1$ *and* $\|M\|_2 \leq 1$,
(b) $\|M\|_* \leq 1$,

then

$$\|vu^\top - M\|_2^2 \leq \frac{2}{\delta}\langle A, vu^\top - M\rangle.$$

Remark: We note that if $v \in \mathbb{S}^{p-1}$ and $u \in \mathbb{S}^{n-1}$ are the leading left and right singular vectors respectively of $A \in \mathbb{R}^{p\times n}$, then since the matrix operator norm and the nuclear norm are dual norms with respect to the trace inner product, we have that

$$\langle A, vu^\top \rangle = v^\top A u = \|A\|_{\mathrm{op}} = \sup_{M \in \mathcal{S}_1} \langle A, M \rangle.$$

Thus, Lemma 4 provides a lower bound on the curvature of the function $M \mapsto \langle A, M\rangle$ as M moves away from the maximiser of the function in $\mathcal{S}_1$.

Proof. Let $A = VDU^\top$ be the singular value decomposition of A, where $V \in \mathbb{R}^{p\times p}$ and $U \in \mathbb{R}^{n\times n}$ are orthogonal matrices with column vectors $v_1 = v, v_2, \ldots, v_p$ and $u_1 = u, u_2, \ldots, u_n$ respectively, and $D \in \mathbb{R}^{p\times n}$ is a rectangular diagonal matrix with nonnegative entries along its main diagonal. The diagonal entries $\sigma_i := D_{ii}$ are the singular values of A, and we may assume without loss of generality that $\sigma_1 \geq \cdots \geq \sigma_r > 0$ are all the positive singular values, for some $r \leq \min\{n, p\}$.

Let $\tilde{M} := V^\top M U$ and denote $e_1^{[d]} := (1, 0, \ldots, 0)^\top \in \mathbb{R}^d$. Then by unitary invariance of the Frobenius norm, we have

$$\|v_1 u_1^\top - M\|_2^2 = \|e_1^{[p]}(e_1^{[n]})^\top - \tilde{M}\|_2^2 = \|\tilde{M}\|_2^2 + 1 - 2\tilde{M}_{11}. \tag{23}$$

On the other hand,

$$\langle A, v_1 u_1^\top - M\rangle = \langle D, e_1^{[p]}(e_1^{[n]})^\top - \tilde{M}\rangle = \sigma_1 - \sum_{i=1}^r \sigma_i \tilde{M}_{ii} \geq \sigma_1(1 - \tilde{M}_{11}) - \sigma_2 \sum_{i=2}^r |\tilde{M}_{ii}|. \tag{24}$$

If condition (a) holds, then $\sigma_2 = 0$ and $\delta = \sigma_1$, so by (23) and (24), we have

$$\|v_1 u_1^\top - M\|_2^2 \leq 2(1 - \tilde{M}_{11}) = \frac{2}{\delta}\langle A, v_1 u_1^\top - M\rangle,$$

as desired.

On the other hand, if condition (b) holds, then by the characterisation of the nuclear norm in Lemma 7 as well as its unitary invariance, we have

$$\sum_{i=1}^r |\tilde{M}_{ii}| = \sup_{\substack{U \in \mathbb{R}^{p\times n} \text{ diagonal} \\ U_{ii} \in \{\pm 1\}\ \forall i}} \langle U, \tilde{M}\rangle \leq \|\tilde{M}\|_* = \|M\|_* \leq 1. \tag{25}$$

On the other hand, if $\|M\|_* \leq 1$, then $\sigma_i \leq 1$ for all i, so

$$\|M\|_2 = \left(\sum_{i=1}^r \sigma_i^2\right)^{1/2} \leq \left(\sum_{i=1}^r \sigma_i\right)^{1/2} \leq 1. \tag{26}$$

Using (23), (24), (25) and (26), we therefore have

$$\begin{aligned}\langle A, v_1 u_1^\top - M\rangle &\geq \sigma_1(1 - \tilde{M}_{11}) - \sigma_2 \sum_{i=2}^r |\tilde{M}_{ii}| \geq (\sigma_1 - \sigma_2)(1 - \tilde{M}_{11}) \\ &\geq \frac{\delta}{2}(\|\tilde{M}\|_2^2 + 1 - 2\tilde{M}_{11}) = \frac{\delta}{2}\|v_1 u_1^\top - M\|_2^2,\end{aligned}$$

as desired. □

Lemma 5. *Let $W \in \mathbb{R}^{p\times n}$ have independent $N(0,\sigma^2)$ entries and let $E := \mathcal{T}(W)$. Then for $u > 0$, we have*

$$\mathbb{P}\big(\|E\|_\infty \geq u\sigma\big) \leq \sqrt{\frac{2}{\pi}} p\lceil \log n\rceil (u + 2/u) e^{-u^2/2}.$$

Proof. Let B be a standard Brownian bridge on $[0,1]$. Then for every $j \in \{1,\dots,p\}$,

$$(E_{j,1},\dots,E_{j,(n-1)}) \stackrel{d}{=} \left(\frac{\sigma B(t)}{\sqrt{t(1-t)}}\right)_{t=\frac{1}{n},\dots,\frac{n-1}{n}}.$$

Thus,

$$\mathbb{P}\big\{\|E\|_\infty \geq \sigma u\big\} \leq p\mathbb{P}\left\{\sup_{t\in[1/n,1-1/n]} \frac{B(t)}{\sqrt{t(1-t)}} \geq u\right\}.$$

Let $t = t(s) := e^{2s}/(e^{2s}+1)$ and define the process X by $X(s) := \{t(s)(1-t(s))\}^{-1/2} B(t(s))$. Recall that the Ornstein–Uhlenbeck process is the centred continuous Gaussian process $\{U(s) : s \in \mathbb{R}\}$ having covariance function $\mathrm{Cov}(U(s_1),U(s_2)) = e^{-|s_1-s_2|}$. We compute that

$$\begin{aligned}\mathrm{Cov}\big(X(s_1),X(s_2)\big) &= \mathrm{Cov}\left(\frac{B\big(e^{2s_1}/(e^{2s_1}+1)\big)}{\sqrt{e^{2s_1}/(e^{2s_1}+1)^2}}, \frac{B\big(e^{2s_2}/(e^{2s_2}+1)\big)}{\sqrt{e^{2s_2}/(e^{2s_2}+1)^2}}\right)\\ &= \left(\frac{e^{s_1}}{e^{2s_1}+1}\frac{e^{s_2}}{e^{2s_2}+1}\right)^{-1} \frac{e^{2\min(s_1,s_2)}}{e^{2\min(s_1,s_2)}+1}\frac{1}{e^{2\max(s_1,s_2)}+1} = e^{-|s_1-s_2|}.\end{aligned}$$

Thus, X is the Ornstein–Uhlenbeck process and we have

$$\begin{aligned}\mathbb{P}\left\{\sup_{t\in[1/n,1-1/n]}\left|\frac{B(t)}{\sqrt{t(1-t)}}\right| \geq u\right\} &= \mathbb{P}\left\{\sup_{s\in[0,\log(n-1)]}|X(s)| \geq u\right\}\\ &\leq \lceil\log n\rceil \mathbb{P}\left\{\sup_{s\in[0,1]}|X(s)| \geq u\right\},\end{aligned}$$

where the inequality follows from the stationarity of the Ornstein–Uhlenbeck process and a union bound. Let $Y = \{Y(t) : t \in \mathbb{R}\}$ be a centred continuous Gaussian process with covariance function $\mathrm{Cov}(Y(s),Y(t)) = \max(1-|s-t|,0)$. Since $\mathbb{E}X(t)^2 = \mathbb{E}Y(t)^2 = 1$ for all t and $\mathrm{Cov}(X(s),X(t)) \geq \mathrm{Cov}(Y(s),Y(t))$, by Slepian's inequality (Slepian, 1962), $\sup_{s\in[0,1]}|Y(s)|$ stochastically dominates $\sup_{s\in[0,1]}|X(s)|$. Hence it suffices to establish the required bound with Y in place of X. The process Y, known as the Slepian process, has excursion probabilities given by closed-form expressions (Slepian, 1961; Shepp, 1971): for $x < u$,

$$\mathbb{P}\left\{\sup_{s\in[0,1]} Y(s) \geq u \;\middle|\; Y(0) = x\right\} = 1 - \Phi(u) + \frac{\phi(u)}{\phi(x)}\Phi(x),$$

where ϕ and Φ are respectively the density and distribution functions of the standard normal

distribution. Hence for $u > 0$ we can write

$$\begin{aligned}
\mathbb{P}\Big\{\sup_{s\in[0,1]}|Y(s)| \geq u\Big\} &= \int_{-\infty}^{\infty} \mathbb{P}\Big\{\sup_{s\in[0,1]}|Y(s)| \geq u \;\Big|\; Y(0) = x\Big\}\phi(x)\,dx \\
&\leq \mathbb{P}(|Y(0)| \geq u) + 2\int_{-u}^{u} \mathbb{P}\Big\{\sup_{s\in[0,1]} Y(s) \geq u \;\Big|\; Y(0) = x\Big\}\phi(x)\,dx \\
&= 2\Phi(-u) + 2\int_{-u}^{u}\big\{\phi(x)\Phi(-u) + \phi(u)\Phi(x)\big\}\,dx \\
&= 2u\phi(u) + 4\Phi(-u)\{1 - \Phi(-u)\} \\
&\leq 2(u + 2u^{-1})\phi(u),
\end{aligned}$$

as desired. □

Proposition 6. *Suppose the first and second largest singular values of $A \in \mathbb{R}^{p\times n}$ are separated by $\delta > 0$. Let unit vectors $v \in \mathbb{S}^{p-1}(k)$ and $u \in \mathbb{S}^{n-1}(\ell)$ be left and right leading singular vectors of A respectively. Let $T \in \mathbb{R}^{p\times n}$ satisfy $\|T - A\|_\infty \leq \lambda$ for some $\lambda > 0$, and let $\mathcal{S}$ be a subset of $p \times n$ real matrices containing $vu^\top$. Suppose one of the following two conditions holds:*

(a) $\operatorname{rank}(A) = 1$ *and* $\mathcal{S} \subseteq \{M \in \mathbb{R}^{p\times n} : \|M\|_2 \leq 1\}$
(b) $\mathcal{S} \subseteq \{M \in \mathbb{R}^{p\times n} : \|M\|_* \leq 1\}$.

Then for any

$$\hat{M} \in \operatorname*{argmax}_{M\in\mathcal{S}}\big\{\langle T, M\rangle - \lambda\|M\|_1\big\},$$

we have

$$\|vu^\top - \hat{M}\|_2 \leq \frac{4\lambda\sqrt{k\ell}}{\delta}.$$

Furthermore, if $\hat{v}$ and $\hat{u}$ are leading left and right singular vectors of $\hat{M}$ respectively, then

$$\max\{\sin\angle(\hat{v}, v), \sin\angle(\hat{u}, u)\} \leq \frac{8\lambda\sqrt{k\ell}}{\delta}. \tag{27}$$

Proof. Using Lemma 4, we have

$$\begin{aligned}
\|vu^\top - \hat{M}\|_2^2 &\leq \frac{2}{\delta}\langle A, vu^\top - \hat{M}\rangle \\
&= \frac{2}{\delta}\big(\langle T, vu^\top - \hat{M}\rangle + \langle A - T, vu^\top - \hat{M}\rangle\big).
\end{aligned} \tag{28}$$

Since $\hat{M}$ is a maximiser of the objective function $M \mapsto \langle T, M\rangle - \lambda\|M\|_1$ over the set $\mathcal{S}$, and since $vu^\top \in \mathcal{S}$, we have the basic inequality

$$\langle T, vu^\top - \hat{M}\rangle \leq \lambda(\|vu^\top\|_1 - \|\hat{M}\|_1). \tag{29}$$

Denote $S_v := \{j : 1 \le j \le p, v_j \neq 0\}$ and $S_u := \{t : 1 \le t \le n, u_t \neq 0\}$. From (28) and (29) and the fact that $\|T - A\|_\infty \le \lambda$, we have

$$\begin{aligned}\|vu^\top - \hat{M}\|_2^2 &\le \frac{2}{\delta}\big(\lambda\|vu^\top\|_1 - \lambda\|\hat{M}\|_1 + \lambda\|vu^\top - \hat{M}\|_1\big)\\ &= \frac{2\lambda}{\delta}\big(\|v_{S_v}u_{S_u}^\top\|_1 - \|\hat{M}_{S_vS_u}\|_1 + \|v_{S_v}u_{S_u}^\top - \hat{M}_{S_vS_u}\|_1\big)\\ &\le \frac{4\lambda}{\delta}\|v_{S_v}u_{S_u}^\top - \hat{M}_{S_vS_u}\|_1 \le \frac{4\lambda\sqrt{k\ell}}{\delta}\|vu^\top - \hat{M}\|_2.\end{aligned}$$

Dividing through by $\|vu^\top - \hat{M}\|_2$, we have the first desired result.

Now, by definition of the operator norm, we have

$$\begin{aligned}\|vu^\top - \hat{M}\|_2^2 &= 1 + \|\hat{M}\|_2^2 - 2v^\top\hat{M}u\\ &\ge 1 + \|\hat{M}\|_2^2 - 2\|\hat{M}\|_{\mathrm{op}} = 1 + \|\hat{M}\|_2^2 - 2\hat{v}^\top\hat{M}\hat{u} = \|\hat{v}\hat{u}^\top - \hat{M}\|_2^2.\end{aligned}$$

Thus,

$$\|vu^\top - \hat{v}\hat{u}^\top\|_2 \le \|vu^\top - \hat{M}\|_2 + \|\hat{v}\hat{u}^\top - \hat{M}\|_2 \le 2\|vu^\top - \hat{M}\|_2 \le \frac{8\lambda\sqrt{k\ell}}{\delta}. \tag{30}$$

We claim that

$$\max\big\{\sin^2\angle(\hat{u}, u), \sin^2\angle(\hat{v}, v)\big\} \le \|vu^\top - \hat{v}\hat{u}^\top\|_2^2. \tag{31}$$

Let $v_0 := (v + \hat{v})/2$ and $\Delta := v - v_0$. Then

$$\begin{aligned}\|vu^\top - \hat{v}\hat{u}^\top\|_2^2 &= \|(v_0 + \Delta)u^\top - (v_0 - \Delta)\hat{u}^\top\|_2^2 = \|v_0(u - \hat{u})^\top\|_2^2 + \|\Delta(u + \hat{u})^\top\|_2^2\\ &= \|v_0\|_2^2\|u - \hat{u}\|_2^2 + \|\Delta\|_2^2\|u + \hat{u}\|_2^2\\ &\ge (\|v_0\|_2^2 + \|\Delta\|_2^2)\min(\|u - \hat{u}\|_2^2, \|u + \hat{u}\|_2^2)\\ &\ge \big\{1 - (\hat{u}^\top u)^2\big\} = \sin^2\angle(\hat{u}, u),\end{aligned}$$

where the penultimate step uses the fact that $\|v_0\|_2^2 + \|\Delta\|_2^2 = 1$. A similar inequality holds for $\sin^2\angle(\hat{v}, v)$, which establishes the desired claim (31). Inequality (27) now follows from (30) and (31). □

The lemma below gives a characterisation of the nuclear norm of a real matrix.

Lemma 7. *For $n, p \ge 1$, let $\mathcal{V}_n$ and $\mathcal{V}_p$ be respectively the sets of $n \times \min(n, p)$ and $p \times \min(n, p)$ real matrices having orthonormal columns. Let $A \in \mathbb{R}^{p\times n}$. Then*

$$\|A\|_* = \sup_{V\in\mathcal{V}_p, U\in\mathcal{V}_n} \langle VU^\top, A\rangle$$

Proof. Suppose we have the singular value decomposition $A = \tilde{V}D\tilde{U}^\top$ where $\tilde{V} \in \mathcal{V}_p$, $\tilde{U} \in \mathcal{V}_n$ and where $D = (D_{ij}) \in \mathbb{R}^{\min(n,p)\times\min(n,p)}$ is a diagonal matrix with decreasing non-negative

diagonal entries. Write V_j for the jth column of V and similarly U_j for the jth column of U. Then

$$\begin{aligned}\sup_{V\in\mathcal{V}_p,U\in\mathcal{V}_n}\langle VU^\top, A\rangle &= \sup_{V\in\mathcal{V}_p,U\in\mathcal{V}_n}\langle VU^\top, \tilde{V}D\tilde{U}^\top\rangle = \sup_{V\in\mathcal{V}_p,U\in\mathcal{V}_n}\langle VU^\top, D\rangle \\ &= \sup_{V\in\mathcal{V}_p,U\in\mathcal{V}_n}\sum_{j=1}^{\min(n,p)} D_{jj}V_j^\top U_j = \sum_{j=1}^{\min(n,p)} D_{jj} = \|A\|_*,\end{aligned}$$

as desired. □

Lemma 8. *Let $\gamma \in \mathbb{R}^{n-1}$ be defined as in (5) for some $n \geq 6$ and $2 \leq z \leq n-2$. Let $\tilde{z} := \min(z, n-z)$. Then*

$$\begin{aligned}\frac{1}{4}\tilde{z} \leq \|\gamma\|_2 &\leq \sqrt{\log(en/2)}\,\tilde{z} \\ \frac{1}{2}\sqrt{n}\tilde{z} \leq \|\gamma\|_1 &\leq 2.1\sqrt{n}\tilde{z}.\end{aligned}$$

Proof. Since the norms of γ are invariant under substitution $z \mapsto n-z$, we may assume without loss of generality that $z \leq n-z$. Hence $\tilde{z} = z$. We have that

$$\begin{aligned}\|\gamma\|_2^2 &= \frac{1}{n}\biggl\{\sum_{t=1}^{z}\frac{t(n-z)^2}{n-t} + \sum_{t=z+1}^{n-1}\frac{(n-t)z^2}{t}\biggr\} \\ &= n^2\biggl\{\sum_{t=1}^{z}\frac{(t/n)(1-z/n)^2}{(1-t/n)}\cdot\frac{1}{n} + \sum_{t=z+1}^{n-1}\frac{(1-t/n)(z/n)^2}{t/n}\cdot\frac{1}{n}\biggr\},\end{aligned}$$

where the expression inside the bracket can be interpreted as a Riemann sum approximation to an integral. We therefore find that

$$n^2\biggl\{I_1 - \frac{(z/n)(1-z/n)}{n}\biggr\} \leq \|\gamma\|_2^2 \leq n^2\biggl\{I_1 + \frac{(z/n)(1-z/n)}{n}\biggr\},$$

where

$$\begin{aligned}I_1 &:= (1-z/n)^2\int_0^{z/n}\frac{r}{1-r}\,dr + (z/n)^2\int_{z/n}^1\frac{1-r}{r}\,dr \\ &= (1-z/n)^2\bigl\{-\log(1-z/n) - z/n\bigr\} + (z/n)^2\bigl\{-\log(z/n) - (1-z/n)\bigr\}.\end{aligned}$$

Since $-\log(1-x) \geq x + x^2/2$ for $0 \leq x < 1$, we have

$$I_1 \geq (z/n)^2(1-z/n)^2.$$

When $n \geq 6$ and $2 \leq z \leq n/2$, we find $\frac{(z/n)(1-z/n)}{n} \leq 3I_1/4$. Hence,

$$\|\gamma\|_2 \geq \frac{1}{2}n(z/n)(1-z/n) \geq \frac{1}{4}z.$$

On the other hand, under the assumption that $z \le n/2$, we have

$$-\log(1-z/n) - z/n \le (z/n)^2.$$

Hence

$$\|\gamma\|_2^2 \le n^2\big\{(1-z/n)^2(z/n)^2 + (z/n)^2\log(n/2)\big\} \le z^2\log(en/2),$$

as required.

For the ℓ_1 norm, we similarly write $\|\gamma\|_1$ as a Riemann sum:

$$\begin{aligned}\|\gamma\|_1 &= \frac{1}{\sqrt{n}}\bigg\{\sum_{t=1}^{z}\sqrt{\frac{t}{n-t}}(n-z) + \sum_{t=z+1}^{n-1}\sqrt{\frac{(n-t)}{t}}z\bigg\}\\ &= n^{3/2}\bigg\{\sum_{t=1}^{z}\sqrt{\frac{t/n}{1-t/n}}(1-z/n)\cdot\frac{1}{n} + \sum_{t=z+1}^{n-1}\frac{1-t/n}{t/n}(z/n)\cdot\frac{1}{n}\bigg\}.\end{aligned}$$

So

$$n^{3/2}\bigg\{I_2 - \frac{\sqrt{z/n(1-z/n)}}{n}\bigg\} \le \|\gamma\|_1 \le n^{3/2}\bigg\{I_2 + \frac{\sqrt{z/n(1-z/n)}}{n}\bigg\},$$

where

$$I_2 := (1-z/n)\int_0^{z/n}\sqrt{\frac{r}{1-r}}\,dr + (z/n)\int_{z/n}^1\sqrt{\frac{1-r}{r}}\,dr = (1-z/n)g(z/n) + (z/n)g(1-z/n),$$

where function $g(a) := \int_0^a \sqrt{r/(1-r)}\,dr = \arcsin(\sqrt{a}) - \sqrt{a(1-a)}$. We can check that $g(a)/a^{3/2}$ has positive first derivative throughout $(0,1)$, and $g(a)/a^{3/2} \searrow 2/3$ as $a \searrow 0$. This implies that $2a^{3/2}/3 \le g(a) \le \pi a^{3/2}/2$. Consequently,

$$\frac{2z}{3n}\bigg(1-\frac{z}{n}\bigg)\bigg(\sqrt{\frac{z}{n}} + \sqrt{1-\frac{z}{n}}\bigg) \le I_2 \le \frac{\pi}{2}\frac{z}{n}\bigg(1-\frac{z}{n}\bigg)\bigg(\sqrt{\frac{z}{n}} + \sqrt{1-\frac{z}{n}}\bigg)$$

Also, for $n \ge 6$ and $2 \le z \le n/2$,

$$\frac{\sqrt{z/n(1-z/n)}}{n} \le \frac{\sqrt{3}}{4+2\sqrt{2}}\frac{z}{n}\bigg(1-\frac{z}{n}\bigg)\bigg(\sqrt{\frac{z}{n}} + \sqrt{1-\frac{z}{n}}\bigg).$$

Therefore,

$$\|\gamma\|_1 \le (\pi/2 + \sqrt{3}/(4+2\sqrt{2}))\sqrt{n}z \sup_{0\le y\le 1/2}(1-y)(\sqrt{y} + \sqrt{1-y}) \le 2.1\sqrt{n}z,$$

and

$$\|\gamma\|_1 \ge (1 - \sqrt{3}/(4+2\sqrt{2}))\sqrt{n}z \inf_{0\le y\le 1/2}(1-y)(\sqrt{y} + \sqrt{1-y}) \ge \frac{1}{2}\sqrt{n}z,$$

□

Proposition 9. *Let $X \sim P \in \mathcal{P}(n, p, k, 1, \vartheta, \tau, \sigma^2)$, with the single changepoint located at z, say (so we may take $\tau = n^{-1}\min\{z, n-z\}$). Define A, E and T as in Section 3. Let $v \in \operatorname{argmax}_{\tilde{v}\in\mathbb{S}^{p-1}} \|A^\top \tilde{v}\|_2$ and $\hat{v} \in \operatorname{argmax}_{\tilde{v}\in\mathbb{S}^{p-1}(k)} \|T^\top \tilde{v}\|_2$. If $n \geq 6$, then with probability at least $1 - 4(p\log n)^{-1/2}$,*

$$\sin\angle(\hat{v}, v) \leq \frac{16\sqrt{2}\sigma}{\tau\vartheta}\sqrt{\frac{\log(p\log n)}{n}}.$$

Proof. From the definition in Section 3, $A = \theta\gamma^\top$, for some $\theta \in \mathbb{R}^p$ satisfying $\|\theta\|_0 \leq k$ and $\|\theta\|_2^2 \geq k\vartheta^2$ and γ defined by (5). Then we have $v = \theta/\|\theta\|_2$. Define also $u := \gamma/\|\gamma\|_2$ and $\hat{u} := T^\top\hat{v}/\|T^\top\hat{v}\|_2$. Then by definition of $\hat{v}$, we have

$$\langle \hat{v}\hat{u}^\top, T\rangle = \|T^\top\hat{v}\|_2 \geq v^\top T u = \langle vu^\top, T\rangle. \tag{32}$$

By Lemma 4 and (32), we obtain

$$\begin{aligned}\|vu^\top - \hat{v}\hat{u}^\top\|_2^2 &\leq \frac{2}{\|\theta\|_2\|\gamma\|_2}\langle A, vu^\top - \hat{v}\hat{u}^\top\rangle \\ &\leq \frac{2}{\|\theta\|_2\|\gamma\|_2}\langle A - T, vu^\top - \hat{v}\hat{u}^\top\rangle \leq \frac{2}{\|\theta\|_2\|\gamma\|_2}\|E\|_\infty\|vu^\top - \hat{v}\hat{u}^\top\|_1. \end{aligned}\tag{33}$$

Note that in fact $v \in \mathbb{S}^{p-1}(k)$, by definition of the matrix A. Moreover, $\hat{v} \in \mathbb{S}^{p-1}(k)$ too, so the matrix $vu^\top - \hat{v}\hat{u}^\top$ has at most $2k$ non-zero rows. Thus, by the Cauchy–Schwarz inequality,

$$\|vu^\top - \hat{v}\hat{u}^\top\|_1 \leq \sqrt{2kn}\|vu^\top - \hat{v}\hat{u}^\top\|_2.$$

By (31) in the proof of Proposition 6, and (33), we find that

$$\sin\angle(\hat{v}, v) \leq \|vu^\top - \hat{v}\hat{u}^\top\|_2 \leq \frac{2\sqrt{2}\|E\|_\infty\sqrt{kn}}{\|\theta\|_2\|\gamma\|_2} \leq \frac{8\sqrt{2}\|E\|_\infty}{\vartheta\tau\sqrt{n}},$$

where we have used Lemma 8 in the final inequality. The desired result follows from bounding $\|E\|_\infty$ with high probability as in (14). □

Lemma 10. *Let $T \in \mathbb{R}^{p\times(n-1)}$ and $\lambda > 0$. Then the following optimisation problem*

$$\max_{M\in\mathcal{S}_2}\big\{\langle T, M\rangle - \lambda\|M\|_1\big\}$$

has a unique solution given by

$$\tilde{M} = \frac{\mathbf{soft}(T, \lambda)}{\|\mathbf{soft}(T, \lambda)\|_2}. \tag{34}$$

Proof. Define $\phi(M, R) := \langle T - R, M\rangle$ and $\mathcal{R} := \{R \in \mathbb{R}^{p\times(n-1)} : \|R\|_\infty \leq \lambda\}$. Then the objective function in the lemma is given by

$$f(M) = \min_{R\in\mathcal{R}}\phi(M, R).$$

We also define

$$g(R) := \max_{M \in \mathcal{S}_2} \phi(M, R) = \|T - R\|_2.$$

Since $\mathcal{S}_2$ and $\mathcal{R}$ are compact, convex subsets of $\mathbb{R}^{p\times(n-1)}$ endowed with the trace inner product, and since ϕ is affine and continuous in both M and R, we can use the minimax equality theorem Fan (1953, Theorem 1) to obtain

$$\max_{M \in \mathcal{S}_2} f(M) = \max_{M \in \mathcal{S}_2} \min_{R \in \mathcal{R}} \phi(M, R) = \min_{R \in \mathcal{R}} \max_{M \in \mathcal{S}_2} \phi(M, R) = \min_{R \in \mathcal{R}} g(R).$$

We note that the dual function g has a unique minimum over $\mathcal{R}$ at $R^{(d)}$, say, where $R^{(d)}_{j,t} := \mathrm{sgn}(T_{j,t}) \min(\lambda, |T_{j,t}|)$. Let

$$M^{(d)} \in \operatorname*{argmax}_{M \in \mathcal{S}_2} \phi(M, R^{(d)}), \quad M^{(p)} \in \operatorname*{argmax}_{M \in \mathcal{S}_2} f(M) \quad \text{and} \quad R^{(p)} \in \operatorname*{argmin}_{R \in \mathcal{R}} \phi(M^{(p)}, R).$$

Then

$$\min_{R \in \mathcal{R}} g(R) = \langle T - R^{(d)}, M^{(d)} \rangle \geq \langle T - R^{(d)}, M^{(p)} \rangle \geq \langle T - R^{(p)}, M^{(p)} \rangle = \max_{M \in \mathcal{S}_2} f(M).$$

Since the two extreme ends of the chain of inequalities are equal, we necessarily have

$$R^{(d)} \in \operatorname*{argmin}_{R \in \mathcal{R}} \langle T - R, M^{(p)} \rangle,$$

and consequently,

$$M^{(p)} \in \operatorname*{argmax}_{M \in \mathcal{S}_2} \langle T - R^{(d)}, M \rangle.$$

The objective $M \mapsto \langle T - R^{(d)}, M \rangle = \langle \mathbf{soft}(T, \lambda), M \rangle$ has a unique maximiser at $\tilde{M}$ defined in (34). Thus, $M^{(p)}$ is unique and has the form given in the lemma. □

The following lemma is used to control the rate of decay of the univariate CUSUM statistic from its peak in the single changepoint setting.

Lemma 11. *For $n \in \mathbb{N}$ and $z \in \{0, 1, \ldots, n\}$, define $f : [0, n] \to \mathbb{R}$ by*

$$f(t) := \begin{cases} \sqrt{\frac{t}{n(n-t)}}(n - z), & \text{if } t \leq z \\ \sqrt{\frac{n-t}{nt}}z, & \text{if } t > z. \end{cases}$$

Then for $\Delta \leq 1$,

$$\{t : f(t) \geq f(z)(1 - \Delta)\} \subseteq [z - 2z\Delta, z + 2(n - z)\Delta].$$

Proof. We note first that $f(t)$ is maximised at $t = z$. If $t \leq z$ and $f(t) \geq f(z)(1 - \Delta)$, then

$$\frac{t}{n - t} \geq \frac{z}{n - z}(1 - \Delta)^2.$$

Equivalently,

$$t \geq \frac{nz(1-\Delta)^2}{n - z + z(1-\Delta)^2} \geq z(1-\Delta)^2 \geq z - 2z\Delta. \tag{35}$$

By symmetry, for $t > z$ with $f(t) \geq f(z)(1-\Delta)$, we have that

$$n - t \geq (n-z)(1-\Delta)^2 \geq n - z - 2(n-z)\Delta. \tag{36}$$

Combining (35) and (36), we have the desired result. □

Lemma 12. *Suppose that $0 = z_0 < z_1 < \cdots < z_\nu < z_{\nu+1} = n$ are integers and that $\mu \in \mathbb{R}^n$ satisfies $\mu_t = \mu_{t'}$ for all $z_i < t \leq t' \leq z_{i+1}$, $0 \leq i \leq \nu$. Define $A := \mathcal{T}(\mu) \in \mathbb{R}^{n-1}$, where we treat μ as a row vector. If the series $(A_t : z_i + 1 \leq t \leq z_{i+1})$ is not constantly zero, then one of the following is true:*

(a) $i = 0$ and $(A_t : z_i + 1 \leq t \leq z_{i+1})$ does not change sign and has strictly increasing absolute values,
(b) $i = \nu$ and $(A_t : z_i + 1 \leq t \leq z_{i+1})$ does not change sign and has strictly decreasing absolute values,
(c) $1 \leq i \leq \nu - 1$ and $(A_t : z_i + 1 \leq t \leq z_{i+1})$ is strictly monotonic,
(d) $1 \leq i \leq \nu - 1$ and $(A_t : z_i + 1 \leq t \leq z_{i+1})$ does not change sign and its absolute values are strictly decreasing then strictly increasing.

Proof. This follows from the proof of Venkatraman (1992, Lemma 2.2). □

The following lemma is used to control the rate of decay of the univariate CUSUM statistic of the the mean series away from its maximum absolute value in the case of two changepoints.

Lemma 13. *Let $1 \leq z < z' \leq n - 1$ be integers and $\mu_0, \mu_1 \in \mathbb{R}$. Define $g : [z, z'] \to \mathbb{R}$ by*

$$g(y) := \sqrt{\frac{n}{y(n-y)}}\{z\mu_0 + (y - z)\mu_1\}$$

Suppose that $\min\{z, z' - z\} \geq n\tau$ and

$$G := \max_{y \in [z, z']} |g(y)| = g(z). \tag{37}$$

Then

$$\sup_{y \in [z, z + 0.2n\tau]} g'(y) \leq -0.52Gn^{-1}\tau.$$

Proof. Define $r := z/n$, $r' := z'/n$, $B := r(\mu_0 - \mu_1)$ and $f(x) := n^{-1/2}g(nx)$ for $x \in [r, r']$. Then

$$f(x) = \frac{B + \mu_1 x}{\sqrt{x(1-x)}} \quad \text{and} \quad f'(x) = \frac{(\mu_1 + 2B)x - B}{2\{x(1-x)\}^{3/2}}.$$

Condition (37) is equivalent to

$$Gn^{-1/2} = \max_{x \in [r, r']} |f(x)| = f(r) = \frac{r\mu_0}{\sqrt{r(1-r)}}. \tag{38}$$

The desired result of the lemma is equivalent to

$$\sup_{x\in[r,r+0.2\tau]} f'(x) \le -0.52Gn^{-1/2}\tau.$$

We may assume without loss of generality that it is not the case that $\mu_0 = \mu_1 = 0$, because otherwise f is the zero function and $G = 0$, so the result holds. In that case, $G > 0$, so $\mu_0 > 0$, and we prove the above inequality by considering the following three cases.

Case 1: $B \le 0$. Then $\mu_1 \ge \mu_0$ and in fact $\mu_1 + 2B < 0$, because otherwise f' is non-negative on $[r, r']$, and if $f'(r) = 0$ (which is the only remaining possibility from (38)) then $B = 0$ and $\mu_1 = 0$, so $\mu_0 = 0$, a contradiction. Moreover, since $\mathrm{sgn}(f'(x)) = \mathrm{sgn}\big((\mu_1 + 2B)x - B\big)$, we deduce that $\frac{B}{\mu_1+2B} \le r \le 1$. In particular, $\mu_1 \le -B = r(\mu_1 - \mu_0) \le \mu_1 - \mu_0$ and hence $\mu_0 \le 0$, again a contradiction.

Case 2: $B > 0$ and $\mu_1 + 2B \le 0$. By (38) and the fact that $\mu_1 < 0$, so that $B > r\mu_0$, we have for $x \in [r, r + \tau]$ that

$$\begin{aligned} f'(x) &\le \frac{-B}{2\{x(1-x)\}^{3/2}} \\ &\le \frac{-B}{2\{r(1-r)\}^{1/2}} \inf_{x\in[r,r+\tau]} \frac{\{r(1-r)\}^{1/2}}{\{x(1-x)\}^{3/2}} \le -2Gn^{-1/2} \inf_{x\in[r,r+\tau]} \frac{r^{1/2}}{x^{1/2}} \le -\sqrt{2}Gn^{-1/2}. \end{aligned}$$

Here, we used the fact that $\min\{r, r' - r\} \ge \tau$ in the final bound.

Case 3: $B > 0$ and $\mu_1 + 2B > 0$, so that $\mu_0 > \mu_1$. In this case, considering $\mathrm{sgn}(f'(x))$ again yields $r \le \frac{B}{\mu_1+2B}$. We claim that

$$\frac{B}{\mu_1 + 2B} \ge r + 0.4\tau. \tag{39}$$

By the fundamental theorem of calculus,

$$\begin{aligned} f(r) - f\Big(\frac{B}{\mu_1+2B}\Big) &= \int_r^{\frac{B}{\mu_1+2B}} \frac{B - (\mu_1+2B)x}{2\{x(1-x)\}^{3/2}}\,dx \\ &= (\mu_1+2B)\Big(\frac{B}{\mu_1+2B} - r\Big)^2 \int_0^1 \frac{u}{2\{x(u)(1-x(u))\}^{3/2}}\,du, \end{aligned} \tag{40}$$

where we have used the substitution $x = x(u) := \frac{B}{\mu_1+2B} - (\frac{B}{\mu_1+2B} - r)u$ in the second step. Similarly,

$$\begin{aligned} f(r+\tau) - f\Big(\frac{B}{\mu_1+2B}\Big) &= \int_{\frac{B}{\mu_1+2B}}^{r+\tau} \frac{B - (\mu_1+2B)\tilde{x}}{2\{\tilde{x}(1-\tilde{x})\}^{3/2}}\,d\tilde{x} \\ &= (\mu_1+2B)\Big(r + \tau - \frac{B}{\mu_1+2B}\Big)^2 \int_0^1 \frac{u}{2\{\tilde{x}(u)(1-\tilde{x}(u))\}^{3/2}}\,du, \end{aligned} \tag{41}$$

using the substitution $\tilde{x} = \tilde{x}(u) := \frac{B}{\mu_1+2B} + (r+\tau-\frac{B}{\mu_1+2B})u$. For every $u \in [0,1]$, we have $x(u) \leq \tilde{x}(u) \leq (1+u)x(u)$. It follows that

$$\frac{\int_0^1 u\{\tilde{x}(u)(1-\tilde{x}(u))\}^{-3/2}\,du}{\int_0^1 u\{x(u)(1-x(u))\}^{-3/2}\,du} \geq \frac{\int_0^1 ux(u)^{-3/2}(1+u)^{-3/2}\,du}{\int_0^1 ux(u)^{-3/2}\,du} = \frac{1}{2^{1/2}}\left\{\frac{(\frac{B}{\mu_1+2B})^{1/2}+r^{1/2}}{(\frac{2B}{\mu_1+2B})^{1/2}+r^{1/2}}\right\}^2$$
$$\geq \frac{1}{2^{1/2}}\left\{\frac{(r+\tau)^{1/2}+r^{1/2}}{2^{1/2}(r+\tau)+r^{1/2}}\right\}^2 \geq 0.45. \tag{42}$$

Therefore, using (40), (41) and (42), together with the fact that $f(r) \geq f(r+\tau)$, we deduce that

$$\frac{B}{\mu_1+2B} - r \geq \frac{\tau}{1+0.45^{-1/2}} > 0.4\tau.$$

Hence (39) holds. For $x \in [r, r+0.2\tau]$, we have

$$f'(x) \leq \frac{-(\mu_1+2B)\big(\frac{B}{2(\mu_1+2B)}-\frac{r}{2}\big)}{2\{x(1-x)\}^{3/2}} \leq \frac{-0.4\tau(\mu_1+2B)}{\sqrt{1.2r(1-r)}}. \tag{43}$$

If $\mu_1 \geq 0$, then $r \leq \frac{B}{\mu_1+2B} \leq 1/2$ and

$$\mu_1 + 2B = 2r\mu_0 + (1-2r)\mu_1 \geq 2r\mu_0. \tag{44}$$

If $\mu_1 < 0$ and $r \geq 1/2$, then

$$\mu_1 + 2B = 2r\mu_0 + (2r-1)(-\mu_1) \geq 2r\mu_0. \tag{45}$$

Finally, if $\mu_1 < 0$ and $r < 1/2$, then, writing $a := 1-2r$ and $b := \frac{2B}{\mu_1+2B} - 1$, we have that $a+b \geq 0.8\tau$ and

$$(\mu_1+2B)\left(\frac{B}{\mu_1+2B}-r\right) = r(1-2r)\mu_0 - 2r(1-r)\mu_1 = ar\mu_0 + \frac{(1-a^2)B}{1+b^{-1}}$$
$$\geq \left(a + \frac{1-a^2}{1+(0.8\tau-a)^{-1}}\right)r\mu_0 \geq 0.57\tau r\mu_0. \tag{46}$$

It follows from (43), (44), (45), (46) and (38) that for $x \in [r, r+0.2\tau]$,

$$f'(x) \leq \frac{-0.57\tau r\mu_0}{\sqrt{1.2r(1-r)}} \leq -0.52Gn^{-1/2}\tau,$$

as desired. □

References

Aston, J. A. D. and Kirch, C. (2012) Evaluating stationarity via change-point alternatives with applications to fMRI data. *Ann. Appl. Stat.*, **6**, 1906–1948.

Aston, J. A. D. and Kirch, C. (2014) Change points in high dimensional settings. *arXiv preprint*, arxiv:1409.1771.

Aue, A., Hörmann, S., Horváth, L. and Reimherr, M. (2009) Break detection in the covariance structure of multivariate time series models. *Ann. Statist.* **37**, 4046–4087.

Bai, J. (2010) Common breaks in means and variances for panel data. *J. Econometrics*, **157**, 78–92.

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2011) Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, **3**, 1–122.

Bücher, A., Kojadinovic, I. Rohmer, T. and Seger, J. (2014) Detecting changes in cross-sectional dependence in multivariate time series. *J. Mult. Anal.*, **132**, 111–128.

Chen, J. and Gupta, A. K. (1997) Testing and locating variance changepoints with application to stock prices. *J. Amer. Statist. Assoc.*, **92**, 739–747.

Chen, Y. and Ye, X. (2011) Projection onto a simplex. *arXiv preprint*, arxiv:1101.6081.

Cho, H. (2016) Change-point detection in panel data via double CUSUM statistic. *Preprint*.

Cho, H. and Fryzlewicz, P. (2012) Multiscale and multilevel technique for consistent segmentation of nonstationary time series. *Statist. Sinica*, **22**, 207–229.

Cho, H. and Fryzlewicz, P. (2015) Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *J. R. Stat. Soc. Ser. B*, **77**, 475–507.

Cribben, I. and Yu, Y. (2015) Estimating whole brain dynamics using spectral clustering. *arXiv preprint*, arxiv:1509.03730.

Csörgö, M. and Horváth, L. (1997) *Limit Theorems in Change-Point Analysis*. John Wiley and Sons, New York.

Cule, M., Samworth, R. J. and Stewart, M. (2010) Maximum likelihood estimation of a multi-dimensional log-concave density. *J. R. Stat. Soc. Ser. B. (with discussion)*, **72**, 545–607.

Darling, D. A. and Erdős, P. (1956) A limit theorem for the maximum of normalised sums of independent random variables. *Duke Math. J.*, **23**, 143–155.

Davis, C. and Kahan, W. M. (1970) The rotation of eigenvectors by a pertubation. III. *SIAM J. Numer. Anal.*, **7**, 1–46.

Dümbgen, L. and Rufibach, K. (2009) Maximum likelihood estimation of a log-concave density and its distribution function: basic properties and uniform consistency. *Bernoulli*, **15**, 40–68.

Enikeeva, F. and Harchaoui, Z. (2014) High-dimensional change-point detection with sparse alternatives. *arXiv preprint*, arxiv:1312.1900v2.

Fan, K. (1953) Minimax theorems. *Proc. Natl. Acad. Sci.*, **39**, 42–47.

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

Fryzlewicz, P. (2014) Wild binary segmentation for multiple change-point detection. *Ann. Statist.*, **42**, 2243–2281.

Gabay, D. and Mercier, B. (1976) A dual algorithm for the solution of nonlinear variational problems via finite element approximations. *Comput. Math. Appl.*, **2**, 17–40.

Hampel, F. R. (1974) The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, **69**, 383–393.

Henry, D., Simani, S. and Patton, R. J. (2010) Fault detection and diagnosis for aeronautic and aerospace missions. In Edwards, C., Lombaerts, T., and Smaili H., eds, *Fault Tolerant Flight Control — A Benchmark Challenge*, pp. 91–128. Springer-Verlag, Berlin.

Horváth, L., Kokoszka, P. and Steinebach, J. (1999) Testing for changes in dependent observations with an application to temperature changes. *J. Multi. Anal.*, **68**, 96–199.

Horváth, L. and Rice, G. (2014) Extensions of some classical methods in change point analysis. *TEST*, **23**, 219–255.

Horváth, L. and Hušková, M. (2012) Change-point detection in panel data. *J. Time Ser. Anal.*, **33**, 631–648.

Hubert, L. and Arabie, P. (1985) Comparing partitions. *J. Classification*, **2**, 193–218.

Jirak, M. (2015) Uniform change point tests in high dimension. *Ann. Statist.*, **43**, 2451–2483.

Kirch, C., Mushal, B. and Ombao, H. (2015) Detection of changes in multivariate time series with applications to EEG data. *J. Amer. Statist. Assoc.*, **110**, 1197–1216.

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

Lavielle, M. and Teyssiere, G. (2006) Detection of multiple change-points in multivariate time series. *Lithuanian Mathematical Journal*, **46**, 287–306.

Olshen, A. B., Venkatraman, E. S., Lucito, R. and Wigler, M. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biometrika*, **5**, 557–572.

Ombao, H., Von Sachs, R. and Guo, W. (2005) SLEX analysis of multivariate nonstationary time series. *J. Amer. Statist. Assoc.*, **100**, 519–531.

Page, E. S. (1955) A test for a change in a parameter occurring at an unknown point. *Biometrika*, **42**, 523–527.

Peng, T., Leckie, C. and Ramamohanarao, K. (2004) Proactively detecting distributed denial ofservice attacks using source IP address monitoring. In Mitrou, N., Kontovasilis, K., Rouskas, G. N., Iliadis, I. and Merakos, L. eds, *Networking 2004*, pp. 771–782. Springer-Verlag, Berlin.

Preuß, P., Puchstein, R. and Dette, H. (2015) Detection of multiple structural breaks in multivariate time series. *J. Amer. Statist. Assoc., to appear.*

Rand, W. M. (1971) Objective criteria for the evaluation of clustering methods. *J. Amer. Statist. Assoc.*, **66**, 846–850.

Rosenfield, D., Zhou, E., Wilhelm, F. H., Conrad, A., Roth, W. T. and Meuret, A. E. (2010) Change point analysis for longitudinal physiological data: Detection of cardio-respiratory changes preceding panic attacks. *Biological Psychology*, **84**, 112–120.

Shepp, L. A. (1971) First passage time for a particular Gaussian process. *Ann. Math. Statist.*, **42**, 946–951.

Slepian, D. (1961) First passage time for a particular Gaussian process. *Ann. Math. Statist.*, **32**, 610–612.

Slepian, D. (1962) The one-sided barrier problem for Gaussian noise. *Bell System Techinical Journal*, **41**, 463–501.

Sparks, R., Keighley, T. and Muscatello, D. (2010) Early warning CUSUM plans for surveillance of negative binomial daily disease counts. *J. Appl. Stat.*, **37**, 1911–1930.

Tillmann, A. N. and Pfetsch M. E. (2014) The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Trans. Inform. Theory*, **60**, 1248–1259.

Venkatraman, E. S. (1992) Consistency results in multiple change-point problems. Doctoral dissertation, to the Department of Statistics, Stanford University.

Vu, V. Q., Cho, J., Lei, J. and Rohe, K. (2013) Fantope projection and selection: a near-optimal convex relaxation of sparse PCA. *Advances in Neural Information Processing Systems 26.*

Yu, Y., Wang, T. and Samworth, R. J. (2015) A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, **102**, 315–323.

Zhang, N. R., Siegmund, D. O., Ji, H. and Li, J. Z. (2010) Detecting simultaneous change-points in multiple sequences. *Biometrika*, **97**, 631–645.