On change-point estimation under Sobolev sparsity

Aurélie Fischer and Dominique Picard

LPSM (UMR CNRS 8001)
Université Paris Diderot
Case courrier 7012
75205 Paris Cedex 13, France
e-mail:
aurelie.fischer@univ-paris-diderot.fr; dominique.picard@univ-paris-diderot.fr

Abstract: In this paper, we consider the estimation of a change-point for possibly high-dimensional data in a Gaussian model, using a maximum likelihood method. We are interested in how dimension reduction can affect the performance of the method. We provide an estimator of the change-point that has a minimax rate of convergence, up to a logarithmic factor. The minimax rate is in fact composed of a fast rate —dimension-invariant— and a slow rate —increasing with the dimension. Moreover, it is proved that considering the case of sparse data, with a Sobolev regularity, there is a bound on the separation of the regimes above which there exists an optimal choice of dimension reduction, leading to the fast rate of estimation. We propose an adaptive dimension reduction procedure based on Lepski's method and show that the resulting estimator attains the fast rate of convergence. Our results are then illustrated by a simulation study. In particular, practical strategies are suggested to perform dimension reduction.

MSC 2010 subject classifications: 62H30, 62G05.
Keywords and phrases: Change-point estimation, maximum likelihood, minimax rate, dimension reduction, Lepski's method.

1. Introduction

1.1. The model

An important problem in the vast domain of statistical learning is the question of unsupervised classification of high-dimensional data. Many examples fall into this category such as the classification of curves or images. Here, we will address a framework where the change between classes occurs on a time scale, which casts the problem into the change-point estimation issue.

We consider, for the sake of simplicity, a change-point problem with exactly two classes: we assume that there exists a change-point τ: before $n\tau$, the observations are in a certain state, after $n\tau$, they are in another state. More precisely, we observe independent random vectors $Y_1, \ldots, Y_n$ with values in $\mathbb{R}^d$, such that

$$Y_i = \theta_i + \eta_i, \quad \eta_i \sim \mathcal{N}(0, \sigma^2 I_d) \text{ i.i.d.}, \quad 1 \le i \le n,$$
$$\forall i \le n\tau, \quad \theta_i = \theta^-,$$
$$\forall i > n\tau, \quad \theta_i = \theta^+.$$

0

In practice, such a model is obtained for instance in the monitoring of patients, where the variables Y_i are a bunch of d biological, chemical and/or clinical observations collected each ten minutes (for example) on a patient, and $n\tau$ reflects a time of change in the patient's condition.

Our aim is to estimate the change-point τ. Using the maximum likelihood approach, also known in this context as CUSUM method, we derive rates of convergence for the estimation of τ, under conditions specified below.

For high-dimensional data, from a computational point of view, there is an obvious need for dimension reduction when estimating τ. Without such a step, the segmentation algorithm might be unstable or even not work at all. Here, we will consider the dimension reduction problem from a theoretical point of view (as opposed to the algorithmic point of view). From a theoretical point of view, one might suspect that it should always be better to keep the whole data, to get the best precision on the estimation of the change-point. In fact, we show that this intuition is not correct. Addressing this dimension reduction problem can require sophisticated tools directly connected to smoothing questions in nonparametric estimation. Especially, as will be seen along the paper, sparsity assumptions, as well as smoothing adaptive methods, can be directly borrowed from the nonparametric statistical inference and fruitfully applied in this context.

To sum up, our goal is to answer the following questions:

(a) Without referring to the technical feasibility, is there a theoretical gain in reducing the dimension, which somehow reduces to the choice between two options : ignoring a part of the data, versus keeping all the data?

(b) If the data is high-dimensional but "sparse", is there a way to use this sparsity to get better results?

(c) If dimension reduction proves to be theoretically more efficient, how could it be performed? Do usual nonparametric smoothing methods work well in a change-point problem?

(d) Does on-line (signal by signal) dimension reduction perform as well as off-line (using a preprocessing involving all the signals)?

1.2. Literature review and related work

The change-point problem has a long history, going back at least to [49]. For an introduction to the domain, the reader may refer for instance to the monographs and articles by [53], [51], [46], [7], [10], [12], [17] or [27].

Change-point detection has many practical applications, ranging from genetics [47] or health [55] to aerospace industry [25]. Note that high-dimensional change-point problems may occur in a wide range of areas. This is the case for instance in the analysis of traffic network data ([43, 44]), in bioinformatics, when studying copy-number variation ([8, 60]), for studying functional magnetic resonance imaging (fMRI) ([2]), in astrostatistics ([9, 56, 45]) or in multimedia indexation ([22]). In these practical applications, the number of observations is

relatively small compared to their dimension, with the change-point possibly occurring only for a few components.

Various different change-point methods and settings have been developed in the literature, in particular in the univariate case. For instance, [33] introduced the pruned exact linear time method, [21] wild binary segmentation, and [20] a simultaneous multiscale change-point estimator, whereas change-point estimation based on resampling has been investigated in [19] and [1]. Some multivariate extensions are described among others in [28], [48], [3] and [34]. Dependent sequences are considered for instance in [23], [24], [4], [6] and [38]. Regarding the high-dimensional context, [31] considers several dependent change-point tests and studies the behavior of the maximum over all test statistics as both the sample size and the number of tests tend to infinity. [15] propose a sparse version of binary segmentation. [5], [26] and [14] are interested in the high-dimensional context of panel data. The high dimension problem is addressed through changes in cross-covariance in [39], [3], [11], [50], [16]. In [54], convex optimization is used to perform regularization for solving the high dimensional change-point problem. In [13, 52], graph-based approaches which are efficient in high dimension are designed. [30] proposes a method based on a statistics inspired by Hotelling's T^2 statistics. [18] consider high-dimensional change-point detection, from the testing point of view.

In this paper, we will study the performance of the procedure from a minimax point of view, in a high-dimensional context. This approach provides an evaluation of the best expectable performances in a particular framework, and the aim is then to provide a procedure attaining these performances.

Minimax estimation is considered already in [36], in the Gaussian white noise model. High-dimensional change-point problems are also studied in [35], which proposes an asymptotically minimax estimator of the change-point location, when the Euclidean norm of the gap tends to infinity as the dimension d goes to infinity.

Our approach is deeply connected to this paper and can be considered as a continuation of this project. The main difference is that in [35], the authors do not question the dimension reduction problem and do not consider the same estimation method. Moreover, [35] make the assumption that the change-point only occurs after a known number of observations and before another known number of observations. This is a crucial difference since knowing that some observations are in the first or last state allows to provide an efficient estimation of this state. Without this assumption, we do not have this opportunity, which adds a difficulty to the problem.

Another related reference is the paper by [59], who proposed a two-stage procedure based on a projection followed by a univariate change point estimation algorithm applied to the projected data, providing rates of convergence for the estimator of the change-point location.

1.3. Outline of the paper

The paper is organized as follows. We begin Section 2 by introducing the change-point model. We also present the problem of dimension reduction and the maximum likelihood estimator of the change-point. We prove that, for a fixed dimension, up to a logarithmic term, the maximum likelihood method, has a minimax rate of convergence. Let us point out that we do not know whether this logarithmic term is necessary or not. Indeed, as explained earlier, in [35], "the edges are known" (and the estimation method uses this knowledge), meaning that the minimax rate is established in the case where the change-point cannot occur before or after a *known* proportion of the observations. Our method is agnostic to this knowledge, creating obvious additional difficulties. Moreover, we show that if the data is sparse, in a Sobolev sense, there exists an optimal dimension reduction, depending on the sparsity constants.

Of course, these constants are not known in practice. The aim of Section 3 is to provide a procedure which behaves as well as if the sparsity constants were known. To attain this optimal projection dimension in an adaptive way, we provide a method relying on the Lepski method. The Lepski method (see for instance [40, 41, 42] and Section 3.1 for more details) is one of the famous methods to obtain adaptivity in various functional estimation settings such as white noise model, regression, density estimation... In these models, minimax optimality is linked with the regularity assumptions imposed on the functions which are estimated. Adaptation methods provide ways to escape from this knowledge and still perform optimally. Note that the proposed method has the advantage of being performed off-line, before the main segmentation step.

Numerical experiments are provided in Section 4. Section 5 is devoted to the proofs.

2. Main result: minimax convergence rate

2.1. Change-point model and assumptions

Let $n \geq 3$. We observe n independent signals $Y_1, \ldots, Y_n$. We assume that each signal Y_i, $i = 1, \ldots, n$, is a d dimensional vector: for every i, $Y_i = (Y_{i,1}, \ldots, Y_{i,d})$ is a random vector with values in $\mathbb{R}^d$.

We suppose that there exist a change-point $0 < \tau < 1$ and two vectors θ^- and θ^+ of $\mathbb{R}^d$, such that the model is given by

$$Y_i = \theta_i + \eta_i, \quad \eta_i \sim \mathcal{N}(0, \sigma^2 I_d) \text{ i.i.d.}, \quad 1 \leq i \leq n, \tag{1}$$
$$\forall i \leq n\tau, \quad \theta_i = \theta^-,$$
$$\forall i > n\tau, \quad \theta_i = \theta^+.$$

Remark 1.
More than one change. Dealing with a finite and *known* number $N \geq 2$ of change-points would not change the theoretical results but would add unnecessary complexity to the proofs.

Connection to Functional Data Analysis : White Noise model. Theoretically, Model (1) is directly connected to the model in [35] where the observation is a sequence of white noise models:

$$X_i(u) = \mu_i(u) + \sigma Z_i(u), \ u \in [0,1], \ i = 1, \ldots, n,$$

where the Z_i's are i.i.d. Gaussian white noises. The goal is to locate a change occurring in the functions μ_i (i.e. $\forall i \leq n\tau$, $\mu_i(u) = \mu_-(u)$, and $\forall i > n\tau$, $\mu_i(u) = \mu_+(u)$, $\forall u \in [0,1]$). To connect this model to (1), we simply project the observations on an orthonormal basis $(\Phi_\ell)_{\ell \geq 1}$ of $\mathbb{L}^2([0,1])$, limiting the observation to the first d projections, with the following formula:

$$Y_i^\ell = \int_{[0,1]} X_i(u)\Phi_\ell(u)du = \theta_i^\ell + \eta_i^\ell, \ i = 1, \ldots, n, \ \ell = 1 \ldots, d,$$

where $\theta_i^\ell = \int_{[0,1]} \mu_i(u)\Phi_\ell(u)du$ and the variables η_i^ℓ are i.i.d. $\mathcal{N}(0, \sigma^2)$.

The white noise model may seem a bit theoretical but it is standardly used in nonparametric statistics as an approximation to more refined models such as functional regression models (as detailed in the following paragraph) or even density estimation models.

Connection to Functional Data Analysis : Regression analysis. Suppose that our data is composed of n independent curves discretely observed on $[0,1]$ with a grid of step size equal to $\frac{1}{d}$.

$$X_i\left(\frac{j}{d}\right) = \mu_i\left(\frac{j}{d}\right) + \sigma_0 Z_{ij}, \ i = 1, \ldots, n, \ j = 1, \ldots d,$$

where the Z_{ij}'s are i.i.d. Gaussian $\mathcal{N}(0,1)$ variables. The goal again is to locate a change occurring in the functions μ_i (i.e. $\forall i \leq n\tau$, $\mu_i(u) = \mu_-(u)$, and $\forall i > n\tau$, $\mu_i(u) = \mu_+(u)$, $\forall u \in [0,1]$). Then, again, this model can be connected to (1), with the help of an orthonormal basis $(\Phi_\ell)_{\ell \geq 1}$ of $\mathbb{L}^2([0,1])$, writing

$$Y_i^\ell = \frac{1}{d}\sum_{j=1}^{d} X_i\left(\frac{j}{d}\right)\Phi_\ell\left(\frac{j}{d}\right) = \theta_i^\ell + \eta_i^\ell + r_i^\ell, \ i = 1, \ldots, n, \ \ell = 1 \ldots, d,$$

where $\theta_i^\ell = \int_{[0,1]} \mu_i(u)\Phi_\ell(u)du$, the variables η_i^ℓ are i.i.d. $\mathcal{N}(0, \frac{\sigma_0^2}{d})$ and r_i^ℓ are deterministic quantities describing the difference between the Riemann sum and the integral which can be negligible for d large enough and standard regularity assumptions —which will not be discussed here— on the functions μ_i and Φ_ℓ.

Heteroscedasticity. For the sake of simplicity, the covariance matrix of the noise η_i in (1) is chosen to be proportional to identity. In various examples, it could be reasonable to choose a covariance of the form $\sigma^2 J$, where J is a known matrix, different from identity. Similarly, in the FDA example, if the Z_i's are i.i.d. Gaussian processes but not white noise, then taking an orthonormal basis of $\mathbb{L}^2([0,1])$ would not necessarily lead to a covariance matrix proportional to identity. Then, a simple change of variables like $J^{-1/2}Y_i$ would lead to a similar behavior, provided appropriate regularity assumptions are made on the parameters $J^{-1/2}\theta^+$ and $J^{-1/2}\theta^-$.

Variance. We suppose here σ^2 to be known. Note that σ^2 may depend on d or on n. For instance, as in one of the previous examples, it may be of the form σ_0^2/d, where σ_0^2 is an absolute and known constant.

Gaussianity. Considering Gaussian noise is a useful simplification, but this assumption is not crucial. We essentially need concentration inequalities, and similar results may likely be obtained under sub-Gaussian hypotheses on the errors. In that case, the considered estimator of τ is no more a maximum likelihood estimator, but simply a CUSUM estimator.

2.2. *Estimation method*

We are interested in the behavior of the maximum likelihood estimator, also called in this case CUSUM estimator:

$$
\hat{k}(d) = \operatorname*{arg\,min}_{k\in\{2,\dots,n-2\}} \left\{ \sum_{i=1}^{k}\sum_{j=1}^{d}\left(Y_{i,j} - \frac{1}{k}\sum_{\ell=1}^{k}Y_{\ell,j}\right)^2 \right.
$$
$$
\left. + \sum_{i=k+1}^{n}\sum_{j=1}^{d}\left(Y_{i,j} - \frac{1}{n-k}\sum_{\ell=k+1}^{n}Y_{\ell,j}\right)^2 \right\},
$$

To prove some of our results, we will need the following sparsity conditions.

2.2.1. *Condition on the means*

For $s > 0$, we define

$$
\Theta(s,L) := \left\{ \theta \in \mathbb{R}^d,\ \sup_{K\in\mathbb{N}^*} K^{2s}\sum_{k\geq K}(\theta_k)^2 \leq L^2 \right\}.
$$

We will suppose that θ^- and θ^+ are in $\Theta(s,L)$.

Remark 2. This assumption expresses a form of sparsity of the coefficients which is standard in nonparametric settings. It corresponds to conditions which are directly connected to the regularity of the function to be estimated in nonparametric estimation (see [57] for a general introduction). So this condition is easily interpretable in the cases of FDA mentioned above. In the more general setting of a high-dimensional physical observation, it is commonly accepted to solve learning problems by introducing sparsity constraints (see for instance [32] or [29] among many others). These constraints can take various forms: we deliberately chose here the Sobolev type. It is among the simplest forms to handle technically. Note that it reflects an ordering: the first coefficients are supposed to be more important than the last ones. This is quite a reasonable assumption

since, generally, modeling of high-dimensional or functional data via a basis expansion results in such a situation. However, other type of structural sparsity could be investigated (like having a finite support of coefficients) which would lead to different methods in particular for the adaptivity part.

Example 1. Let $s = 1/2$ and $L = 1$. Assume that θ is defined by $\theta_k = \frac{1}{\sqrt{2}k}$ for $k = 1, \ldots, k_{\max}$ and $\theta_k = 0$ for $k > k_{\max}$. Then, for every $K = 1, \ldots, k_{\max}$,

$$K^{2s} \sum_{k \geq K} (\theta_k)^2 \leq L^2 = K \sum_{k=K}^{k_{\max}} \left(\frac{1}{\sqrt{2}k} \right)^2 \leq 1.$$

Hence, $\theta \in \Theta(1/2, 1)$.

Note that there are possible extensions to other kinds of sparsity, considering for instance coefficients belonging to the set

$$\Theta_q(L) := \left\{ \theta \in \mathbb{R}^d, \ \sum_k |\theta_k|^q \leq L \right\},$$

where $q < 1$. Note that this choice requires more sophisticated smoothing algorithms.

To end up this section we introduce the following important parameter:

$$\varepsilon := \min\{\tau, 1 - \tau\}. \tag{2}$$

As we are in a not degenerate case —there is a change, ε is a strictly positive quantity, which measures the potential lack of information at the border of the interval $[0, 1]$. It is important to notice that the theoretical performances of the procedures will depend on ε. However, the procedure is agnostic to ε, which therefore *will not be supposed to have a known lower bound.*

2.3. Dimension reduction for the estimation of τ

Our aim is to determine whether or not it is efficient to perform a dimension reduction when estimating the change-point τ. More specifically, we will investigate the effect of replacing the vectors $Y_i = (Y_{i,1}, \ldots, Y_{i,d})$, $i \leq n$ (called "raw data"), by, for $p < d$, $Y_i(p) := (Y_{i,1}, \ldots, Y_{i,p})$, $i \leq n$, the vectors of $\mathbb{R}^p$ composed of the p first coordinates of Y_i.
For each projection dimension p, we may define:

$$\hat{k}(p) = \operatorname*{arg\,min}_{k \in \{2, \ldots, n-2\}} \left\{ \sum_{i=1}^{k} \sum_{j=1}^{p} \left(Y_{i,j} - \frac{1}{k} \sum_{\ell=1}^{k} Y_{\ell,j} \right)^2 \right.$$

$$\left. + \sum_{i=k+1}^{n} \sum_{j=1}^{p} \left(Y_{i,j} - \frac{1}{n-k} \sum_{\ell=k+1}^{n} Y_{\ell,j} \right)^2 \right\}.$$

We set

$$\hat{\tau}(p) = \frac{\hat{k}(p)}{n}.$$

In the sequel, we will use the notation

$$\Delta^2 := \sum_{j=1}^{d} (\theta_j^- - \theta_j^+)^2 = \|\theta^+ - \theta^-\|^2.$$

We also define, for $p \le d$,

$$\Delta_p^2 := \sum_{j=1}^{p} (\theta_j^- - \theta_j^+)^2, \quad \Psi_n(p, \Delta_p) = \frac{\sigma^2}{n\Delta_p^2} \left(1 \vee \frac{\sigma^2 p}{n\Delta_p^2} \right).$$

Example 2. Let θ^+ and θ^- be defined by $\theta_k^+ = \frac{1}{\sqrt{2k}}$ and $\theta_k^- = -\frac{1}{\sqrt{2k}}$. The rate Ψ_n is plotted as a function of n and p in Figure 1.
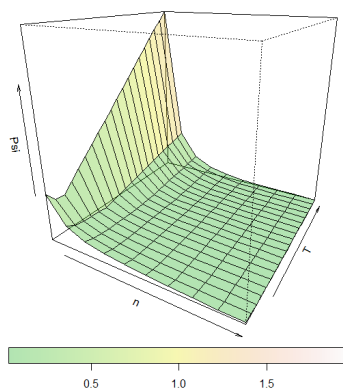

FIG 1. *Example of plot of Ψ_n as a function of n and p ($n = 1, \dots, 10$; $p = 1, \dots, 20$).*

The next result describes the behavior of the estimated change-point $\hat{\tau}(p)$.

Proposition 1. *For any $\gamma > 0$, there exist constants $\kappa(\gamma, \varepsilon)$ and $c(\gamma, \varepsilon)$ such that, if*

$$\Delta_p^2 \ge c(\gamma, \varepsilon) \frac{\sigma^2 \ln(n)}{n},$$

then

$$P\Big(|\hat{\tau}(p) - \tau| \ge \kappa(\gamma, \varepsilon) \ln(n) \Psi_n(p, \Delta_p)\Big) \le cn^{-\gamma},$$

where c is an absolute constant.

Remark 3.

No sparsity required. Note that no condition on the sparsity of θ^+ and θ^- is needed for this result.

Minimaxity. Thanks to [35], one can observe that $\Psi_n(p, \Delta_p)$ is the minimax rate in this framework. Compared to their result, we are apparently loosing a logarithmic factor. However, it is important to stress that in [35], a fixed lower bound on ε is supposed to be known, whereas our estimator $\hat{\tau}(p)$ is adaptive in ε. In other terms, *the procedure $\hat{\tau}(p)$ has minimax rate possibly up to a logarithmic factor*. About this logarithmic factor, it is also worth noticing that $\ln(n)$ could be substituted by any sequence r_n, provided that the factor $n^{-\gamma}$ is simultaneously replaced by $\exp(-c r_n)$ in Proposition 1. (Here, c is a constant which can be made explicit by a closer look into the proof.)

ε-dependence. Looking carefully at the proofs, the constants $c(\varepsilon, \gamma)$ and $\kappa(\varepsilon, \gamma)$ can be taken proportional to $\frac{(\gamma+1)}{\varepsilon^2}$. This remark proves that no condition is required on the proximity of the change point to one extremity of the interval of observation. It also shows that the dependence on ε is of polynomial form. An interesting point, beyond the aim of this paper, would be to investigate whether ε^2 is the optimal rate. One comment that can be made is that our proofs are especially adequate when ε is fixed (does not depend on n).

Fast rate/slow rate. For $p = d$, $\Psi_n(d, \Delta_d) = \Psi_n(d, \Delta)$. The rate is composed of two different regimes: a "fast one" $\frac{\sigma^2 \ln(n)}{n\Delta^2}$, which does not depend on the dimension d and a "slow one" $\frac{\sigma^4 \ln(n) d}{(n\Delta^2)^2}$, which is rapidly deteriorating with the dimension. From the results above, we deduce that if $c(\gamma, \varepsilon)\frac{\sigma^2 \ln(n)}{n} \leq \Delta^2 < \frac{\sigma^2 d}{n}$, the rate of convergence is $\frac{\sigma^4 \ln(n) d}{(n\Delta^2)^2}$ (small gap, slow performances), whereas if $\Delta^2 \geq \frac{\sigma^2 d}{n} \vee c(\gamma, \varepsilon)\frac{\sigma^2 \ln(n)}{n}$, it is $\frac{\sigma^2 \ln(n)}{n\Delta^2}$. This last rate is obviously much better, and with this latter condition on Δ, taking $p = d$ (so raw data) allows to obtain the best rate $\frac{\sigma^2 \ln(n)}{n\Delta^2}$. Taking a smaller p could lead to a reduction of Δ_p damaging the rate.

However this latter condition is quite restrictive on Δ when d is large. In the next paragraph, we will try to refine this condition, gaining on the size p of the projection.

Without assumptions on the behavior of the parameters θ^+ and θ^-, there is nothing much to hope about the way Δ_p is increasing in p. At this stage, it is fruitful to introduce sparsity assumptions. If we assume that the means θ^- and θ^+ belong to $\Theta(s, L)$, then, for p such that $\Delta^2 \geq 8L^2 p^{-2s}$, Δ_p and Δ are comparable, in the sense that $\Delta_p^2 \geq \Delta^2/2$. Indeed, $\Delta^2 - \Delta_p^2 = \sum_{j=p+1}^d (\theta_j^- - \theta_j^+)^2 \leq 4p^{-2s}L^2$, so that

$$\frac{\Delta_p^2}{\Delta^2} \geq 1 - \frac{4p^{-2s}L^2}{\Delta^2} \geq 1/2.$$

This is precisely what is exploited in the first part of Theorem 1 below.

Let us observe that if Δ_p and Δ are comparable (in the sense above), then $\Psi_n(p, \Delta_p) \sim \Psi_n(p, \Delta)$ becomes much easier to analyse. In particular, we see that, again, it is composed of two regimes —a slow one and a fast one— and

the dependence in p becomes easily understandable: $\frac{\sigma^2 \ln(n)}{n\Delta^2}$ for $p \leq \frac{n\Delta^2}{\sigma^2}$, and $\frac{\sigma^4 \ln(n) p}{(n\Delta^2)^2}$ for larger p's.

This corresponds to what may be observed in practical applications: when the dimension p is increasing, one first observes an improved convergence of $\hat{\tau}$, then the rate remains stable for a while, and then convergence gets less good again. Note that two different convergence rates have also been highlighted in other change-point settings, for instance in [58].

2.4. Minimax convergence rate under sparsity condition

The following theorem is an immediate consequence of Proposition 1 in the case where one assumes the sparsity condition on the means.

Theorem 1. *We consider Model (1), with the means θ^+ and θ^- in $\Theta(s, L)$. For any $\gamma > 0$, there exist constants $\kappa(\gamma, \varepsilon)$ and $c(\gamma, \varepsilon)$ such that, if*

$$\Delta^2 \geq \left[2c(\gamma, \varepsilon) \frac{\sigma^2 \ln(n)}{n} \vee 8L^2 p^{-2s} \right],$$

then

$$P\Big(|\hat{\tau}(p) - \tau| \geq \kappa(\gamma, \varepsilon) \ln(n) \Psi_n(p, \Delta) \Big) \leq cn^{-\gamma}.$$

If, now,

$$\Delta^2 \geq \left[2c(\gamma, \varepsilon) \frac{\sigma^2 \ln(n)}{n} \vee 8L^2 p^{-2s} \vee \frac{\sigma^2 p}{n} \right], \tag{3}$$

then

$$P\Big(|\hat{\tau}(p) - \tau| \geq \kappa(\gamma, \varepsilon) \frac{\sigma^2 \ln(n)}{n\Delta^2} \Big) \leq cn^{-\gamma}.$$

Here, c is an absolute constant.

It is important to notice that a big difference with Proposition 1 is that the rate is $\Psi_n(p, \Delta)$ instead of $\Psi_n(p, \Delta_p)$, which is more "honest" in a sense. The price to pay is then, as is intuitive, that p should be large enough. In the second statement, we look at the condition on Δ and p to obtain the fast rate. For Δ fixed, we see that p must not be too large or too small.

Condition (3) contains two terms: one is increasing in p, one decreasing. Hence it can be optimized leading to

$$p_{opt} \sim p_s := \left(\frac{8L^2 n}{\sigma^2} \right)^{\frac{1}{1+2s}}. \tag{4}$$

We obtain the next corollary, corresponding to this projection dimension p_s.

Corollary 1. *Under the conditions above, for any $\gamma > 0$, there exist constants $\kappa(\gamma, \varepsilon)$ and $c(\gamma, \varepsilon)$ such that, if*

$$\Delta^2 \geq \left[2c(\gamma, \varepsilon) \frac{\sigma^2 \ln(n)}{n} \vee \left(\frac{\sigma^2}{n} \right)^{\frac{2s}{1+2s}} \left(8L^2 \right)^{\frac{1}{1+2s}} \right],$$

$$P\Big(|\hat{\tau}(p_s) - \tau| \geq \kappa(\gamma,\varepsilon)\frac{\sigma^2\ln(n)}{n\Delta^2}\Big) \leq cn^{-\gamma}.$$

Interpretation is that the quantity $\left[2c(\gamma,\varepsilon)\frac{\sigma^2\ln(n)}{n}\vee\left(\frac{\sigma^2}{n}\right)^{\frac{2s}{1+2s}}(8L^2)^{\frac{1}{1+2s}}\right]$ is the minimal gap between the two regimes ensuring that the "fast" rate $\frac{\sigma^2\ln(n)}{n\Delta^2}$ can be obtained, with an appropriate projection dimension.

Remark 4.   1. We see here that there is an obvious advantage in reducing the dimension, since it allows to obtain the best rate with less restricting conditions on the gap Δ.

2. We observe that the greater Δ^2, the faster the rate of convergence of $\hat{\tau}$, which is quite natural.

3. At first sight, the rate of convergence and the conditions could seem quite unsatisfactory, but observe that very often σ^2 is of the form $\frac{\sigma_0^2}{d}$. In this case, the rate of convergence is of the order $\left(\frac{nd}{\sigma_0^2}\right)^{\frac{-2s}{1+2s}}\Delta^{-2}$.

4. Formula (4) indicates that the optimal p depends on the sparsity constant s, which is rarely known.

5. If we now look for a procedure searching for an optimal p in an adaptive way (without knowing the regularity s), some comments can be made before proposing a solution. In particular, one may ask whether it is possible to optimize individually (on each signal Y_i of $\mathbb{R}^d$), or if it is necessary to perform an off-line preprocessing (requiring the use of all the signals). The form of the optimal projection dimension $p_s \sim \left(\frac{nd}{\sigma_0^2}\right)^{\frac{1}{1+2s}}$ allows to answer this question. Indeed, any adaptive smoothing performed individually on each signal Y_i (such as thresholding, lasso...) would lead at best to a dimension of the form $p_{opt} = \left(\frac{d}{\sigma_0^2}\right)^{\frac{1}{1+2s}}$, which would induce a lost of a polynomial factor in n in the rates. This means that it is obviously more efficient to find a procedure performing the smoothing globally (off-line).

3. Fast rate of convergence: adaptive choice of p

The message of the section above is the following: for a multichannel signal with sparsity conditions, there is a lower bound on Δ above which there exists an efficient choice of p, leading to the fast rate of estimation for the parameter τ, $\frac{\sigma^2\ln(n)}{n\Delta^2}$. However, in Corollary 1, this choice depends on the knowledge of the regularity parameter s. An essential question then is to construct a adaptive procedure, that is, to design a strategy still performing optimally, without knowledge about the regularity.

There are several ways to give an answer to this question and one can for instance look at the procedure introduced in [59], which proves adaptivity under slightly different conditions.

Our preference here will be to take inspiration into nonparametric statistics which provide many adaptive procedures, and in particular Lepski's method.

This will lead to a procedure which is quite simple, and interesting by itself in this context.

3.1. Lepski's procedure

Let us recall, as already mentioned in the Introduction, that the Lepski method ([40, 41, 42]) is a strategy allowing to obtain adaptivity in various functional estimation settings such as white noise model, regression or density estimation. In these models, minimax optimality is linked with the regularity assumptions imposed on the functions which are estimated. In these nonparametric problems, there is a balance to obtain between a "variance" term, typically of the form $\frac{p}{n}$, and a "bias" term, typically of the form p^{-2s}. The Lepski procedure proposes to choose the minimal p among those such that an estimated version of the bias is below a bound.

For the sake of clarity, let us first recall the classical Lepski procedure in the standard Gaussian white noise model. Note that it will not be described in the original form presented in the first papers, corresponding to kernel estimation methods. Here, the procedure is adapted to the orthogonal series estimation methods, which is more suitable for a transposition to our case.

Consider the following model:

$$Z_j = \beta_j + \varepsilon_j, \; j = 1, \ldots, d, \tag{5}$$

where the ε_j's are i.i.d. $\mathcal{N}(0, \nu^2)$. The Lepski procedure for choosing the optimal projection dimension p consists in defining $\hat{p}$ as follows:

$$\hat{p} := \min \left\{ k \geq 1 : \forall d \geq j \geq m \geq k, \sum_{\ell=m}^{j} (Z_\ell)^2 \leq C_{\mathcal{L}} j \nu^2 \ln(d) \right\},$$

where $C_{\mathcal{L}}$ is a tuning constant of the procedure.

In our change-point setting, a transformation of the data is necessary to fall into the frame of Model (5). We will apply the Lepski method to a surrogate data vector built on the whole observation.

3.2. Preprocessing

Using the complete data set (so off-line), we define a surrogate data vector, which will be used to find an optimal $\hat{p}$. We assume, for the sake of simplicity, that n is even; otherwise, the modifications are elementary.

We set:

$$Z_j = \frac{1}{n} \sum_{i=1}^{n} Y_{i,j} - \frac{2}{n} \sum_{i=1}^{n/2} Y_{i,j}, \quad j = 1, \ldots, d.$$

This vector $Z = (Z_j)_{1 \leq j \leq d}$ is a special case of Model (5), where

$$\beta_j = (1-\tau)(\theta_j^+ - \theta_j^-)\mathbf{1}_{\{\tau \geq 1/2\}} + \tau(\theta_j^+ - \theta_j^-)\mathbf{1}_{\{\tau < 1/2\}}, \quad j = 1, \ldots, d$$

$$\varepsilon_j = \sum_{i=1}^{n/2} \frac{-1}{n}\eta_{i,j} + \sum_{i=n/2+1}^{n} \frac{1}{n}\eta_{i,j}, \quad j = 1, \ldots, d$$

$$\nu^2 = \frac{\sigma^2}{n}.$$

3.3. Adaptive convergence rate

We consider the Lepski procedure applied to the vector Z, producing a projection dimension $\hat{p}$. This parameter is then just plugged in the maximum likelihood procedure for estimating $\hat{\tau}$.

It is well-known that estimating the regularity of a signal is impossible without important extraneous assumptions, but Lepski's procedure provides a projection dimension $\hat{p}$ which, with overwhelming probability, is smaller than the optimal p_s (defined in (4) above) and such that the bias $\Delta^2 - \Delta_{\hat{p}}$ is controlled, which is precisely the need here.

The following theorem states that the method leads to an optimal selection, up to logarithmic terms. As announced, Lepski's method allows an adaptive choice of p: though the optimal p_s is unknown, we are able to achieve the same convergence rate as in Corollary 1 with $p = p_s$.

Theorem 2. *We consider Model (1) and assume that θ^+ and θ^- belong to $\Theta(s, L)$. We suppose that there exists a constant $\alpha > 0$ such that*

$$\frac{n}{\sigma^2} \geq \alpha \ln d.$$

For any $\gamma > 0$, there exist constants $C_{\mathcal{L}}$, $\kappa(\gamma, \varepsilon)$, $c(\gamma, \varepsilon)$ and R such that, letting

$$\hat{p} := \min\left\{k \geq 1 : \forall d \geq j \geq m \geq k, \sum_{\ell=m}^{j}(Z_\ell)^2 \leq C_{\mathcal{L}}j\frac{\sigma^2}{n}\ln(d \vee n)\right\},$$

if

$$\Delta^2 \geq 2c(\gamma, \varepsilon)\frac{\sigma^2 \ln(n)}{n} \vee R\left(\frac{\sigma^2 \ln(d \vee n)}{n}\right)^{\frac{2s}{1+2s}},$$

then

$$P\left(|\hat{\tau}(\hat{p}) - \tau| \geq \kappa(\gamma, \varepsilon)\frac{\sigma^2 \ln(n)}{n\Delta^2}\right) \leq cn^{-\gamma}.$$

Remark 5. A thorough examination of the proof in Section 5.2 shows that the constant $C_{\mathcal{L}}$ only needs to be "large enough" (see (10)). Obviously, there is no point in pretending that the bound in (10) is optimal. Hence, $C_{\mathcal{L}}$ is to be considered as a tuning constant of the method. The theorem proves that if the constant is large enough, then an optimal result is obtained.

4. Numerical study

In this section, we provide some simulations illustrating our theoretical results.

4.1. Rate of convergence

In this experiment, we study the rate of convergence of the estimator $\hat{\tau}$. Let $d = 20$, $p = 10$, $\sigma = 1$, $\tau = 0.3$. Let us consider data generated from Model (1) with the means θ^- and θ^+ obtained from the following distribution: $\theta^- \sim \mathcal{N}(0, \frac{1}{20j^2}))$, $\theta^+ \sim \mathcal{N}(-\theta^-, 10^{-4})$.

To get a first insight about the rate of convergence, we simulate 1000 times a sample of length n, for n chosen between 20 and 4000, and plot in Figure 2 the mean and median of the error $|\tau - \hat{\tau}|$ over the 1000 trials in function of n, together with the function $n \mapsto \ln(n)\Psi_n(p, \Delta_p)$ corresponding to the theoretical rate of convergence obtained in Proposition 1. Note that the rate of convergence of $|\tau - \hat{\tau}|$ is given in the proposition up to a constant $\kappa(\gamma, \varepsilon)$. Nevertheless, the figure provides an appropriate illustration of the result as soon as n is large enough.
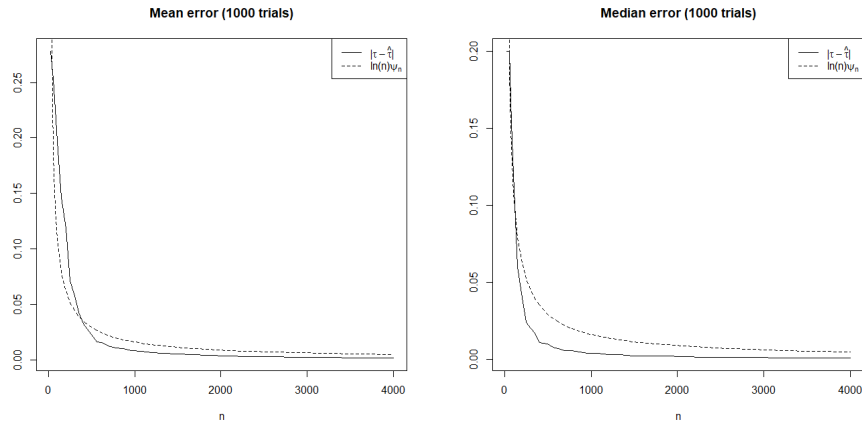

FIG 2. *Plot of $|\hat{\tau} - \tau|$ as a function of n (mean and median over 1000 trials).*

Then, simulating 1000 samples, for each value of the sample size n between 500 and 4000, we try to estimate of the rate of convergence by computing the linear regression of $|\tau - \hat{\tau}|$ by $\ln(n)$: omitting the logarithmic factor, an exponent -1 is to be found, corresponding to the rate of convergence $\frac{1}{n}$. Figure 3 provides an illustration of this linear regression, considering again the mean and the median over the 1000 trials. On this example, the estimated slope of the regression line is -1.172 for the mean and -1.098 for the median.
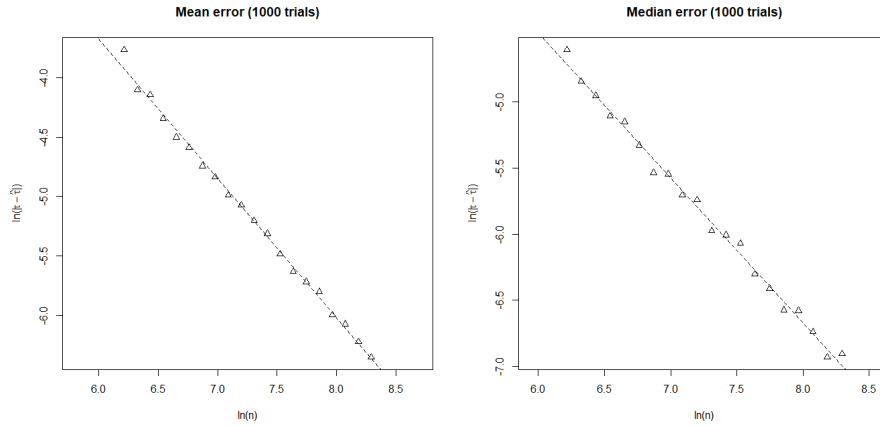
FIG 3. *Plot of* $\ln(|\hat{\tau} - \tau|)$ *as a function of* $\ln(n)$ *(mean and median over 1000 trials).*

4.2. Selection of p

In Theorem 2, we suggest to select p using Lepski's method. Before introducing a practical procedure for the selection of p, let us illustrate the fact that the performance of the estimator $\hat{\tau}$ may indeed vary a lot as a function of p, so that selecting the right p is a crucial issue in the estimation of τ.

We set $d = 200$, $n = 100$, $\sigma = 1$, $\tau = 0.3$. We consider data generated from Model (1) with means θ^- and θ^+ built as follows:

- Case A: $\theta^- \sim \mathcal{N}(0, V)$, $\theta^+ \sim \mathcal{N}(0, V)$, $V = \text{diag}(v_1, \ldots, v_d)$, $v_j = \frac{1}{2j^2}$ for $j = 1, \ldots, d$.
- Case B: θ^- is such that $\theta_j^- \sim \mathcal{N}(0, 1/2)$ for $j = 1, \ldots, 20$, and $\theta_j^- \sim \mathcal{N}(0, \frac{1}{2(j-20)^2})$ for $j = 21, \ldots, d$. θ^+ is such that $\theta_j^+ \, \mathcal{N}(\theta_j^-, 10^{-2})$) for $j = 1, \ldots, 20$, and $\theta_j^+ \sim \mathcal{N}(0, \frac{1}{2(j-20)^2})$ for $j = 21, \ldots, d$.

We simulated 5000 data sets according to Model (1) in each of the two cases. Figure 4 and 5 show the mean and median error $|\hat{\tau} - \tau|$ over the 5000 trials as a function of p. In the first case, the best result is obtained already with $p = 1$, whereas for the second, taking p around 30 is a good choice.
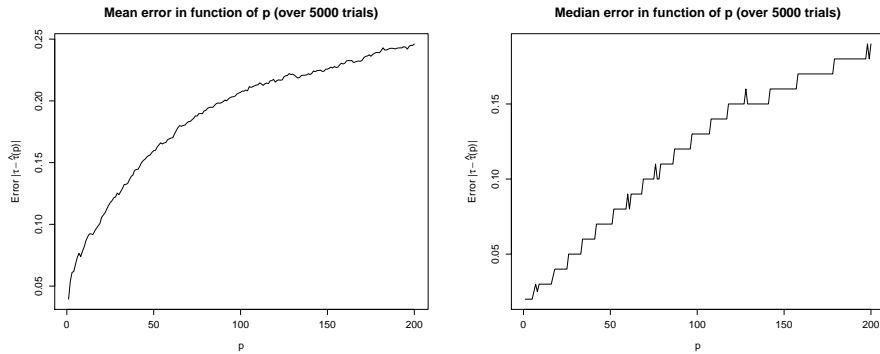
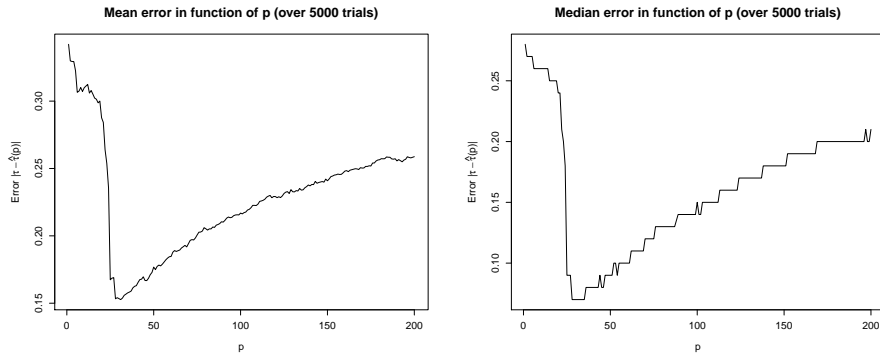FIG 4. *Mean and median of the error over 5000 trials for Model A.*

FIG 5. *Mean and median of the error over 5000 trials for Model B.*

Theorem 2 provides a theoretical way to select p. However, the statement depends on a tuning constant $C_{\mathcal{L}}$. In practice, it is simpler to try to select directly p. In the sequel, two such procedures are investigated, yielding two estimators $\hat{p}_1$ and $\hat{p}_2$.

- Method 1. This method is often used to search for tuning constants in adaptive methods. The idea is to find a division of the set $\{1, \ldots, d\}$ into $\{1, \ldots, \hat{p}_1\}$ and its complementary, where the two subsets are corresponding to two "regimes" for the data, one with "big coefficients", one with small ones.
  Let $\bar{Z}^{(p)} = \frac{1}{p} \sum_{j=1}^{p} Z_j$ and $\bar{Z}^{(-p)} = \frac{1}{d-p} \sum_{j=p+1}^{d} Z_j$, and consider

$$V(p) = \sum_{j=1}^{p} (Z_j - \bar{Z}^{(p)})^2 + \sum_{j=p+1}^{d} (Z_j - \bar{Z}^{(-p)})^2.$$

This quantity V is computed for every $p = 1, \ldots, d$ and the value $\hat{p}_1$ is chosen such that

$$\hat{p}_1 \in \arg \min_{p=1,\ldots,d} V(p).$$

Indeed, this procedure, by searching for a change-point along $Z_1, \ldots, Z_d$, should separate the first most significative differences $\theta_j^- - \theta_j^+$, where $j = 1, \ldots, \hat{p}_1$, from the remaining ones, expected to be less significative for estimating $\hat{\tau}$, in such a way that keeping for the estimation all components until $\hat{p}_1$ seems a reasonable choice.

- Method 2. The second idea is more computationally involved and based on subsampling. When performing subsampling, the indices drawn at random are sorted, so that the parameter of interest τ remains indeed approximatively unchanged. For each $p = 1, \ldots, d$, we compute $\hat{\tau}(p)$ for a collection of subsamples. Then, $\hat{p}_2$ is set to the value of p minimizing the variance of $\hat{\tau}$ over all subsamples. Here, 100 subsamples are built, each of them containing 80% of the initial sample.

Remark 6. Proportions of data from 50% to 90% have also been tried, with quite similar results. Observe that picking a quite small proportion of data for subsampling could be interesting since it provides more variability between the subsamples, but, at the same time, the fact that the ratio between the dimension d and the sample size is modified may be annoying when the aim is to select p. We also considered a version of subsampling where a different subsampling index is drawn for every $p = 1, \ldots, d$: again, this provides more variability in the subsamples, but τ may also vary more than in the classical version. The results were not significantly different.

The performance of the two methods is compared with the result obtained using the value of p minimizing the average value of $|\tau - \hat{\tau}(p)|$ over a large number of trials, called hereafter oracle $p^\star$ (here, $p^\star = 30$ as obtained above for 5000 trials). Of course, $p^\star$ is not available in practice, since it depends on the true τ. However, it is introduced as a benchmark. The results, corresponding to 1000 trials, for Model B, are shown in Figure 6 and Table 1. The performances of the proposed methods could seem unsatisfactory in absolute terms. Nevertheless, the data has deliberately been chosen difficult to segment. Indeed, to illustrate the selection of p, it seems more appropriate to consider a high-dimensional, hard situation, rather than an easy one where the true τ is always found exactly. Observe that the two methods perform very similarly, with a slight advantage of Method 2 over Method 1. However, Method 2 is based on subsampling, and, as such, is more CPU-time consuming.
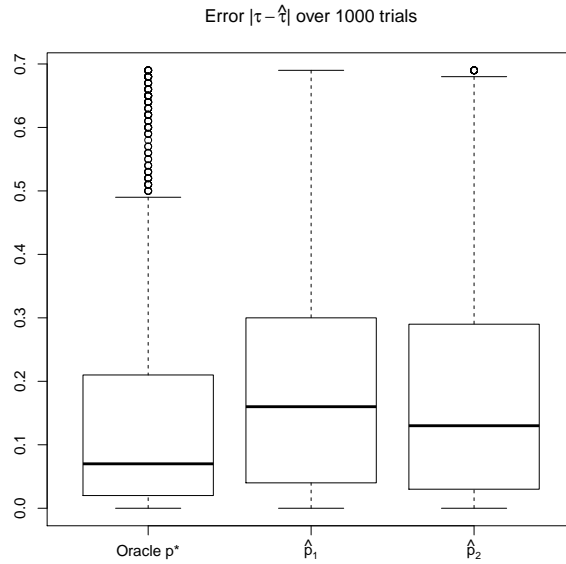
FIG 6. *Error of the two selection procedures over 1000 trials, compared with the error obtained using the oracle $p^\star = 30$.*

Error over 1000 trials	Oracle $p^\star$	$\hat{p}_1$	$\hat{p}_2$
Mean	0.1524	0.2207	0.2047
(Standard deviation)	(0.18735)	(0.21329)	(0.20841)

TABLE 1

Mean and standard deviation over 1000 trials of the error obtained with the oracle $p^\star$ and the two selection methods.

5. Proofs

5.1. Proof of Proposition 1

Our proof will heavily rely on standard concentration inequalities for Gaussian and chi-square distributions, detailed in the Appendix (see Section A).

In the sequel, for the sake of simplicity, we will assume that $n\tau \in \{2, \ldots, n-2\}$. This will not have any consequence on the result but will save us the repeated use of integer parts. Also, in this proof, $\Psi_n(p, \Delta)$ will be replaced by Ψ_n to lighten notation. The proof uses several lemmas, whose proofs are given at the end of the section. First, Lemma 1 shows that $\hat{\tau}$ may be written using Gaussian and chi-square random variables.

Lemma 1. *Under Model* (1), *the estimator* $\hat{\tau}(p)$ *may be written as follows:*

$$\hat{\tau}(p) = \underset{t \in \{\frac{2}{n}, \dots, \frac{n-2}{n}\}}{\arg\min} K^p(t),$$

where

$$K^p(t) = -\sum_{j=1}^{p} \sigma^2 V_j^2(t) - \sum_{j=1}^{p} \sigma^2 W_j^2(t) + \sum_{j=1}^{p} \delta_j^2 \frac{(nt - n\tau)n\tau}{nt} + 2N_1(t) - 2N_2(t).$$

Here, $V_j^2(t)$ *and* $W_j^2(t)$, $j = 1, \dots, p$, *are independent* $\chi^2(1)$ *random variables,* $N_1(\tau) = N_2(\tau) = 0$, *for every* $t \neq \tau$, $N_1(t) \sim \mathcal{N}\left(0, \sum_{j=1}^{p} \sigma^2 (nt - n\tau)\delta_j^2\right)$ *and* $N_2(t) \sim \mathcal{N}\left(0, \sum_{j=1}^{p} \frac{\sigma^2(nt - n\tau)^2 \delta_j^2}{nt}\right)$.

The expression obtained for $\hat{\tau}(p)$ is then used in the sequel for building an upper bound for $P\left(|\hat{\tau}(p) - \tau| \geq \lambda \Psi_n\right)$, where $\lambda > 0$.

Lemma 2. *For* $\lambda > 0$, *we have*

$$P\left(|\hat{\tau}(p) - \tau| \geq \lambda \Psi_n\right)$$

$$\leq P\left(\inf_{\frac{k}{n} - \tau \geq \lambda \Psi_n} K^p\left(\frac{k}{n}\right) < K^p(\tau)\right) + P\left(\inf_{\frac{k}{n} - \tau \leq -\lambda \Psi_n} K^p\left(\frac{k}{n}\right) < K^p\tau\right).$$

We will only evaluate the probability $P\left(\inf_{\frac{k}{n} - \tau \geq \lambda \Psi_n} K^p\left(\frac{k}{n}\right) < K^p(\tau)\right)$ in what follows, since the second term can be treated in a symmetrical way.

Lemma 3. *For* $\lambda > 0$,

$$P\left(\inf_{\frac{k}{n} - \tau \geq \lambda \Psi_n} K^p\left(\frac{k}{n}\right) < K^p(\tau)\right) \leq P^{V,W} + P^N,$$

where

$$P^{V,W} =$$

$$P\Bigg(\exists k \in \{n\tau + n\lambda \Psi_n, \dots, n-2\},$$

$$\sum_{j=1}^{p} V_j^2\left(\frac{k}{n}\right) + \sum_{j=1}^{p} W_j^2\left(\frac{k}{n}\right) - \sum_{j=1}^{p} V_j^2(\tau) - \sum_{j=1}^{p} W_j^2(\tau) > \frac{n\Delta_p^2}{\sigma^2} \frac{(\frac{k}{n} - \tau)n\tau}{2k}\Bigg),$$

$$P^N =$$

$$P\Bigg(\exists k \in \{n\tau + n\lambda \Psi_n, \dots, n-2\}, \frac{|N_1(\frac{k}{n}) - N_2(\frac{k}{n})|}{\sigma^2} > \frac{n\Delta_p^2}{\sigma^2} \frac{(\frac{k}{n} - \tau)n\tau}{4k}\Bigg).$$

Lemma 4 provides a control of the term P^N. For $P^{V,W}$, 3 different cases will be considered, addressed in Lemmas 5 , 6, and 7 below.

Lemma 4 (Control of P^N). *The following inequality holds:*

$$P^N \le 2n \left[\exp\left(-\frac{\varepsilon^2 \lambda}{2^6}\right) + \exp\left(-\frac{\varepsilon^2 n \Delta_p^2}{2^6 \sigma^2}\right) \right].$$

To get an upper bound for the term $P^{V,W}$, let us begin with the case where $n\Delta_p^2 \le 32 p \sigma^2 / \varepsilon^2$. Moreover, the situation where $\frac{\sigma^4 p}{(n\Delta_p^2)^2} \ge \frac{1}{\lambda}$ will be addressed first.

Lemma 5 (Control of $P^{V,W}$, case 1). *Assume that $\frac{\sigma^4 p}{(n\Delta_p^2)^2} \ge \frac{1}{\lambda}$. Then,*

$$P^{V,W} \le 4(n+1) \exp\left(-\frac{\lambda \varepsilon^2}{2^{10}}\right).$$

The more intricate situation where $\frac{\sigma^4 p}{(n\Delta_p^2)^2} \le \frac{1}{\lambda}$ (i.e. $\frac{(n\Delta_p^2)^2}{\sigma^4 p} \ge \lambda$) is considered in the next lemma.

Lemma 6 (Control of $P^{V,W}$, case 2). *Assume that $n\Delta_p^2 \le 32 p \sigma^2 / \varepsilon^2$ and that $\frac{\sigma^4 p}{(n\Delta_p^2)^2} \le \frac{1}{\lambda}$. Then,*

$$P^{V,W} \le 12n \exp\left(-\frac{\lambda \varepsilon^2}{2^{14}}\right) + 12n \exp\left(-\frac{n\Delta_p^2 \varepsilon^2}{2^7 \sigma^2}\right).$$

Then, Lemma 7 provides an upper bound for $P^{V,W}$ in the case where $n\Delta_p^2 \ge 32 p \sigma^2 / \varepsilon^2$.

Lemma 7 (Control of $P^{V,W}$, case 3). *Assume that $n\Delta_p^2 \ge 32 p \sigma^2 / \varepsilon^2$. Then,*

$$P^{V,W} \le 8n \exp\left(-\frac{n\Delta_p^2 \varepsilon}{2^6 \sigma^2}\right) + 4n \exp\left(-\frac{\lambda \varepsilon^2}{2^7}\right).$$

End of the proof of Proposition 1. Collecting the results of the different lemmas, we see that $P(|\hat{\tau} - \tau| \ge \lambda \Psi_n)$ may be upper bounded by a sum of terms all of the form

$$c_1 n \exp(-c_2(\varepsilon)\lambda)$$

or

$$c_1 n \exp\left(-c_2(\varepsilon)\frac{n\Delta_p^2}{\sigma^2}\right),$$

where c_1 denote an absolute constant, and c_2 is polynomial in ε. Recalling that $\frac{n\Delta_p^2}{\sigma^2} \ge c(\gamma, \varepsilon) \ln(n)$, and taking $\lambda = \kappa(\gamma, \varepsilon) \ln(n)$, this proves Proposition 1. $\quad\square$

Proof of Lemma 1. Let us denote by $P_{(\tau,\theta^+,\theta^-)}$ the probability distribution associated with Model (1). We will consider the behavior of our estimators under the probability $P_{(\tau,\theta^+,\theta^-)}$. Using the notation $x^+ = (x_1^+, \ldots, x_p^+)$, and $x^- = (x_1^-, \ldots, x_p^-)$, observe that $\hat{\tau}$ may be defined in the following way:

$$\hat{\tau}(p) = \frac{1}{n} \underset{k \in \{2,\ldots,n-2\}}{\arg\min} \left\{ \sum_{i=1}^{k} \sum_{j=1}^{p} \left(Y_{i,j} - \frac{1}{k} \sum_{\ell=1}^{k} Y_{\ell,j} \right)^2 \right.$$

$$\left. + \sum_{i=k+1}^{n} \sum_{j=1}^{p} \left(Y_{i,j} - \frac{1}{n-k} \sum_{\ell=k+1}^{n} Y_{\ell,j} \right)^2 \right\}$$

$$= \underset{t \in \{\frac{2}{n},\ldots,\frac{n-2}{n}\}}{\arg\min} K^p(t).$$

where

$$K^p(t) = \min_{x^-,x^+} L(t, x^-, x^+) - L(\tau, 0, 0).$$

Here, the function L is given (for $t \in \{\frac{2}{n}, \ldots, \frac{n-2}{n}\}$) by

$$L(t, x^-, x^+) = \sum_{i=1}^{nt} \sum_{j=1}^{p} (Y_{i,j} - \theta_j^- - x_j^-)^2 + \sum_{i=nt+1}^{n} \sum_{j=1}^{p} (Y_{i,j} - \theta_j^+ - x_j^+)^2.$$

As an aside, not used in the sequel, note that writing

$$\frac{dP_{(t,\theta_+ + x^+, \theta_- + x^-)}}{dP_{(\tau,\theta_+,\theta_-)}} = \exp\left(-\frac{1}{2\sigma^2} (L(t, x^-, x^+) - L(\tau, 0, 0)) \right)$$

highlights the facts that we are actually searching for a maximum likelihood estimator. Let us consider the case $t \geq \tau$. The other case can be treated in a symmetrical way. For $t \geq \tau$, and under the distribution $P_{(\tau,\theta^+,\theta^-)}$, we may write

$$L(t, x^-, x^+) - L(\tau, 0, 0)$$

$$= \sum_{i=1}^{n\tau} \sum_{j=1}^{p} ((x_j^-)^2 - 2\eta_{i,j}x_j^-) + \sum_{i=nt+1}^{n} \sum_{j=1}^{p} ((x_j^+)^2 - 2\eta_{i,j}x_j^+)$$

$$+ \sum_{i=n\tau+1}^{nt} \sum_{j=1}^{p} \left((\theta_j^+ - \theta_j^- - x_j^-)^2 + 2\eta_{i,j}(\theta_j^+ - \theta_j^- - x_j^-) \right)$$

$$= \sum_{i=1}^{n\tau} \sum_{j=1}^{p} ((x_j^-)^2 - 2\eta_{i,j}x_j^-) + \sum_{i=nt+1}^{n} \sum_{j=1}^{p} ((x_j^+)^2 - 2\eta_{i,j}x_j^+)$$

$$+ \sum_{i=n\tau+1}^{nt} \sum_{j=1}^{p} \left((\delta_j - x_j^-)^2 + 2\eta_{i,j}(\delta_j - x_j^-) \right),$$

so that

$$L(t, x^-, x^+) - L(\tau, 0, 0)$$
$$= \sum_{i=nt+1}^{n} \sum_{j=1}^{p} ((x_j^+)^2 - 2\eta_{i,j} x_j^+) + \sum_{i=1}^{nt} \sum_{j=1}^{p} ((x_j^-)^2 - 2\eta_{i,j} x_j^-)$$
$$+ \sum_{i=n\tau+1}^{nt} \sum_{j=1}^{p} \left(\delta_j^2 - 2\delta_j x_j^- + 2\delta_j \eta_{i,j} \right), \tag{6}$$

where $\delta = (\delta_1, \ldots, \delta_p)$ is the vector $\theta^+ - \theta^-$. Now, we have to minimize in (x^-, x^+) the previous expression. By differentiation, we obtain that the minimum is attained by taking, for every j,

$$\hat{x}_j^+ = \frac{\sum_{i=nt+1}^{n} \eta_{i,j}}{n - nt}, \tag{7}$$
$$\hat{x}_j^- = \frac{\sum_{i=1}^{nt} \eta_{i,j} + (nt - n\tau)\delta_j}{nt}. \tag{8}$$

Plugging the minimizers (7) and (8) into expression (6) leads to the minimum

$$K^p(t) = \sum_{j=1}^{p} \left(-\frac{\left(\sum_{i=nt+1}^{n} \eta_{i,j} \right)^2}{n - nt} - \frac{\left(\sum_{i=1}^{nt} \eta_{i,j} + (nt - n\tau)\delta_j \right)^2}{nt} \right.$$
$$\left. + (nt - n\tau)\delta_j^2 + 2\delta_j \sum_{i=n\tau+1}^{nt} \eta_{i,j} \right).$$

Under $P_{(\tau, \theta^+, \theta^-)}$, $K^p(t)$ can be written in the following way:

$$K^p(t) = -\sum_{j=1}^{p} \sigma^2 V_j^2(t) - \sum_{j=1}^{p} \sigma^2 W_j^2(t) + \sum_{j=1}^{p} \delta_j^2 \frac{(nt - n\tau)n\tau}{nt} + 2N_1(t) - 2N_2(t),$$

where

$$\sigma^2 V_j^2(t) = \frac{\left(\sum_{i=nt+1}^{n} \eta_{i,j} \right)^2}{n - nt},$$
$$\sigma^2 W_j^2(t) = \frac{\left(\sum_{i=1}^{nt} \eta_{i,j} \right)^2}{nt},$$
$$N_1(t) = \sum_{j=1}^{p} \sum_{i=n\tau+1}^{nt} \eta_{i,j} \delta_j,$$
$$N_2(t) = \sum_{j=1}^{p} \frac{\sum_{i=1}^{nt} \eta_{i,j}(nt - n\tau)\delta_j}{nt}, \quad N_2(\tau) = 0.$$

$\square$

Proof of Lemma 2. We write

$$
P\left(|\hat{\tau} - \tau| \geq \lambda \Psi_n\right)
$$

$$
= P\left(\inf_{|\frac{k}{n} - \tau| \geq \lambda \Psi_n} K^p\left(\frac{k}{n}\right) < \inf_{|\frac{k}{n} - \tau| < \lambda \Psi_n} K^p\left(\frac{k}{n}\right)\right)
$$

$$
\leq P\left(\inf_{|\frac{k}{n} - \tau| \geq \lambda \Psi_n} K^p\left(\frac{k}{n}\right) < K^p(\tau)\right)
$$

$$
\leq P\left(\inf_{\frac{k}{n} - \tau \geq \lambda \Psi_n} K^p\left(\frac{k}{n}\right) < K^p(\tau)\right) + P\left(\inf_{\frac{k}{n} - \tau \leq -\lambda \Psi_n} K^p\left(\frac{k}{n}\right) < K^p(\tau)\right).
$$

$\square$

Proof of Lemma 3. We have

$$
P\left(\inf_{\frac{k}{n} - \tau \geq \lambda \Psi_n} K^p\left(\frac{k}{n}\right) < K^p(\tau)\right)
$$

$$
= P\left(\exists k \in \{2, \ldots, n-2\}, \ \frac{k}{n} - \tau \geq \lambda \Psi_n, \ K^p\left(\frac{k}{n}\right) < K^p(\tau)\right)
$$

$$
\leq P\Bigg(\exists k \in \{n\tau + n\lambda \Psi_n, \ldots, n-2\},
$$

$$
- \sum_{j=1}^{p} V_j^2\left(\frac{k}{n}\right) - \sum_{j=1}^{p} W_j^2\left(\frac{k}{n}\right) + \frac{2N_1(\frac{k}{n}) - 2N_2(\frac{k}{n})}{\sigma^2} + \frac{n\Delta_p^2}{\sigma^2}\frac{(\frac{k}{n} - \tau)n\tau}{k}
$$

$$
< - \sum_{j=1}^{p} V_j^2(\tau) - \sum_{j=1}^{p} W_j^2(\tau) + \frac{2N_1(\tau) - 2N_2(\tau)}{\sigma^2}\Bigg)
$$

$$
\leq P\Bigg(\exists k \in \{n\tau + n\lambda \Psi_n, \ldots, n-2\}, \sum_{j=1}^{p} V_j^2\left(\frac{k}{n}\right) + \sum_{j=1}^{p} W_j^2\left(\frac{k}{n}\right)
$$

$$
- \sum_{j=1}^{p} V_j^2(\tau) - \sum_{j=1}^{p} W_j^2(\tau) - \frac{2N_1(\frac{k}{n}) - 2N_2(\frac{k}{n})}{\sigma^2} > \frac{n\Delta_p^2}{\sigma^2}\frac{(\frac{k}{n} - \tau)n\tau}{k}\Bigg),
$$

since $N_1(\tau) = N_2(\tau) = 0$.

Thus,

$$P\left(\inf_{\frac{k}{n}-\tau \geq \lambda \Psi_n} K^p\left(\frac{k}{n}\right) < K^p\left(\tau\right)\right)$$

$$\leq P\Bigg(\exists k \in \{n\tau + n\lambda\Psi_n, \ldots, n-2\},$$

$$\left|\sum_{j=1}^p V_j^2\left(\frac{k}{n}\right) + \sum_{j=1}^p W_j^2\left(\frac{k}{n}\right) - \sum_{j=1}^p V_j^2\left(\tau\right) - \sum_{j=1}^p W_j^2\left(\tau\right)\right| > \frac{n\Delta_p^2}{\sigma^2}\frac{(\frac{k}{n}-\tau)n\tau}{2k}\Bigg)$$

$$+ P\left(\exists k \in \{n\tau + n\lambda\Psi_n, \ldots, n-2\}, \frac{|2N_1(\frac{k}{n}) - 2N_2(\frac{k}{n})|}{\sigma^2} > \frac{n\Delta_p^2}{\sigma^2}\frac{(\frac{k}{n}-\tau)n\tau}{2k}\right).$$

Here, we used the fact that $P(|A-B| > x) \leq P(|A| > x/2) + P(|B| > x/2)$. $\quad\square$

Proof of Lemma 4. Using the Gaussian concentration inequality (12) recalled in the Appendix, we may write

$$P^N \leq \sum_{k \in \{n\tau+n\lambda\Psi_n, \ldots, n-2\}} P\left(\left|N_1\left(\frac{k}{n}\right)\right| > \frac{\Delta_p^2 n(\frac{k}{n}-\tau)n\tau}{8k}\right)$$

$$+ \sum_{k \in \{n\tau+n\lambda\Psi_n, \ldots, n-2\}} P\left(\left|N_2\left(\frac{k}{n}\right)\right| > \frac{\Delta_p^2 n(\frac{k}{n}-\tau)n\tau}{8k}\right)$$

$$\leq \sum_{k \in \{n\tau+n\lambda\Psi_n, \ldots, n-2\}} 2\exp\left(-\frac{\left(\frac{\Delta_p^2 n(\frac{k}{n}-\tau)n\tau}{8k}\right)^2}{2n\Delta_p^2(\frac{k}{n}-\tau)\sigma^2}\right)$$

$$+ \sum_{k \in \{n\tau+n\lambda\Psi_n, \ldots, n-2\}} 2\exp\left(-\frac{\left(\frac{\Delta_p^2 n(\frac{k}{n}-\tau)n\tau}{8k}\right)^2}{2\frac{n\Delta_p^2(\frac{k}{n}-\tau)^2\sigma^2}{\frac{k}{n}}}\right)$$

$$\leq 2n\left[\exp\left(-\frac{\tau^2 n\Delta_p^2 \Psi_n \lambda}{2^6\sigma^2}\right) + \exp\left(-\frac{\tau^2 n\Delta_p^2}{2^6\sigma^2}\right)\right]$$

$$\leq 2n\left[\exp\left(-\frac{\tau^2\lambda}{2^6}\right) \wedge \exp\left(-\frac{\tau^2\lambda p\sigma^2}{2^6 n\Delta_p^2}\right) + \exp\left(-\frac{\tau^2 n\Delta_p^2}{2^6\sigma^2}\right)\right]$$

$$\leq 2n\left[\exp\left(-\frac{\tau^2\lambda}{2^6}\right) + \exp\left(-\frac{\tau^2 n\Delta_p^2}{2^6\sigma^2}\right)\right]$$

$$\leq 2n\left[\exp\left(-\frac{\varepsilon^2\lambda}{2^6}\right) + \exp\left(-\frac{\varepsilon^2 n\Delta_p^2}{2^6\sigma^2}\right)\right].$$

$\square$

Proof of Lemma 5. We have

$$
P^{V,W} \leq P\Bigg(\exists k \in \{n\tau + n\lambda\Psi_n, \dots, n-2\},
$$

$$
\Bigg| \sum_{j=1}^{p} V_j^2\left(\frac{k}{n}\right) - p + \sum_{j=1}^{p} W_j^2\left(\frac{k}{n}\right) - p \Bigg| > \frac{n\Delta_p^2}{\sigma^2} \frac{(\frac{k}{n} - \tau)n\tau}{4k} \Bigg)
$$

$$
+ P\Bigg(\Bigg| \sum_{j=1}^{p} V_j^2(\tau) - p + \sum_{j=1}^{p} W_j^2(\tau) - p \Bigg| > \frac{n\Delta_p^2}{\sigma^2} \frac{\lambda\Psi_n n\tau}{4n} \Bigg).
$$

Using concentration results for the chi-square distribution recalled in Corollary 2 in the Appendix, we obtain

$$
P\Bigg(\exists k \in \{n\tau + n\lambda\Psi_n, \dots, n-2\},
$$

$$
\Bigg| \sum_{j=1}^{p} V_j^2\left(\frac{k}{n}\right) - p + \sum_{j=1}^{p} W_j^2\left(\frac{k}{n}\right) - p \Bigg| > \frac{n\Delta_p^2}{\sigma^2} \frac{(\frac{k}{n} - \tau)n\tau}{4k} \Bigg)
$$

$$
\leq 2 \sum_{k \in \{n\tau + n\lambda\Psi_n, \dots, n-2\}} P\Bigg(\Bigg| \sum_{j=1}^{p} V_j^2\left(\frac{k}{n}\right) - p \Bigg| > \frac{n\Delta_p^2}{\sigma^2} \frac{(\frac{k}{n} - \tau)n\tau}{8k} \Bigg)
$$

$$
\leq 4 \sum_{k \in \{n\tau + n\lambda\Psi_n, \dots, n-2\}} \Bigg[\exp\Bigg(-\Bigg(\frac{n\Delta_p^2}{\sigma^2} \frac{(\frac{k}{n} - \tau)n\tau}{8k} \Bigg)^2 \frac{1}{16p} \Bigg)
$$

$$
\vee \exp\Bigg(-\frac{n\Delta_p^2}{\sigma^2} \frac{(\frac{k}{n} - \tau)n\tau}{8k \times 4} \Bigg) \Bigg]
$$

$$
\leq 4n \Bigg[\exp\Bigg(-\Bigg(\frac{n\Delta_p^2 \lambda\Psi_n \tau}{\sigma^2} \Bigg)^2 \frac{1}{2^{10}p} \Bigg) \vee \exp\Bigg(-\frac{n\Delta_p^2 \lambda\Psi_n \tau}{2^5 \sigma^2} \Bigg) \Bigg]
$$

$$
\leq 4n \Bigg[\exp\Bigg(-\frac{\lambda^2 \tau^2 \sigma^4 p}{2^{10}(n\Delta_p^2)^2} \Bigg) \vee \exp\Bigg(-\frac{\lambda\tau}{2^5} \Bigg) \Bigg]
$$

$$
\leq 4n \Bigg[\exp\Bigg(-\frac{\lambda\tau^2}{2^{10}} \Bigg) \vee \exp\Bigg(-\frac{\lambda\tau}{2^5} \Bigg) \Bigg]
$$

$$
\leq 4n \exp\Bigg(-\frac{\lambda\varepsilon^2}{2^{10}} \Bigg).
$$

The last but one inequality follows from the assumption $\frac{\sigma^4 p}{(n\Delta_p^2)^2} \geq \frac{1}{\lambda}$. Likewise,

$$
P\Bigg(\Bigg| \sum_{j=1}^{p} (V_j^2(\tau) - 1) + \sum_{j=1}^{p} (W_j^2(\tau) - 1) > \frac{n\Delta_p^2}{\sigma^2} \frac{\lambda\Psi_n \tau}{4} \Bigg| \Bigg) \leq 4\exp\Bigg(-\frac{\lambda\varepsilon^2}{2^{10}} \Bigg).
$$

$\square$

Proof of Lemma 6. We have

$$
P^{V,W} \leq P\Bigg(\exists k \in \{n\tau + n\lambda\Psi_n, \dots, n-2\},
$$
$$
\left| \sum_{j=1}^{p} V_j^2 \left(\frac{k}{n}\right) - \sum_{j=1}^{p} V_j^2 (\tau) \right| > \frac{n\Delta_p^2}{\sigma^2} \frac{(\frac{k}{n} - \tau)n\tau}{4k} \Bigg)
$$
$$
+ P\Bigg(\exists k \in \{n\tau + n\lambda\Psi_n, \dots, n-2\},
$$
$$
\left| \sum_{j=1}^{p} W_j^2 \left(\frac{k}{n}\right) - \sum_{j=1}^{p} W_j^2 (\tau) \right| > \frac{n\Delta_p^2}{\sigma^2} \frac{(\frac{k}{n} - \tau)n\tau}{4k} \Bigg)
$$
$$
:= P^V + P^W.
$$

Let us compute

$$
\sigma^2 \left[W_j^2 \left(\frac{k}{n}\right) - W_j^2 (\tau) \right]
$$
$$
= \left(\sum_{i=1}^{n\tau} \eta_{i,j} \right)^2 \left(\frac{1}{k} - \frac{1}{n\tau} \right) + \left(\sum_{i=n\tau+1}^{k} \eta_{i,j} \right)^2 \frac{1}{k}
$$
$$
+ \frac{2}{k} \sum_{i=1}^{n\tau} \eta_{i,j} \sum_{i=n\tau+1}^{k} \eta_{i,j}
$$
$$
= \frac{k - n\tau}{k} \left[\left(\sum_{i=n\tau+1}^{k} \eta_{i,j} \right)^2 \frac{1}{k - n\tau} - \left(\sum_{i=1}^{n\tau} \eta_{i,j} \right)^2 \frac{1}{n\tau} \right]
$$
$$
+ \frac{2}{k} \sum_{i=1}^{n\tau} \eta_{i,j} \sum_{i=n\tau+1}^{k} \eta_{i,j}.
$$

Similarly,

$$\sigma^2\left[V_j^2\left(\frac{k}{n}\right) - V_j^2(\tau)\right]$$

$$= \left(\sum_{i=k+1}^{n} \eta_{i,j}\right)^2 \left(\frac{1}{n-k} - \frac{1}{n-n\tau}\right) - \left(\sum_{i=n\tau+1}^{k} \eta_{i,j}\right)^2 \frac{1}{n-n\tau}$$

$$- \frac{2}{n-n\tau} \sum_{i=k+1}^{n} \eta_{i,j} \sum_{i=n\tau+1}^{k} \eta_{i,j}$$

$$= \frac{k-n\tau}{n-n\tau}\left[\left(\sum_{i=k+1}^{n} \eta_{i,j}\right)^2 \frac{1}{n-k} - \left(\sum_{i=n\tau+1}^{k} \eta_{i,j}\right)^2 \frac{1}{k-n\tau}\right]$$

$$- \frac{2}{n-n\tau} \sum_{i=k+1}^{n} \eta_{i,j} \sum_{i=n\tau+1}^{k} \eta_{i,j}.$$

As a consequence, we get

$$P^W \leq \sum_{k\in\{n\tau+n\lambda\Psi_n,\ldots,n-2\}}$$

$$P\left(\left|\sum_{j=1}^{p} \frac{k-n\tau}{k}\left[\left(\sum_{i=n\tau+1}^{k} \eta_{i,j}\right)^2 \frac{1}{k-n\tau} - \left(\sum_{i=1}^{n\tau} \eta_{i,j}\right)^2 \frac{1}{n\tau}\right]\right|\right.$$

$$\left. > n\Delta_p^2 \frac{(\frac{k}{n}-\tau)n\tau}{8k}\right)$$

$$+ \sum_{k\in\{n\tau+\lambda\Psi_n,\ldots,n-2\}} P\left(\left|\sum_{j=1}^{p} \frac{2}{k} \sum_{i=1}^{n\tau} \eta_{i,j} \sum_{i=n\tau+1}^{k} \eta_{i,j}\right| > n\Delta_p^2 \frac{(\frac{k}{n}-\tau)n\tau}{8k}\right)$$

$$:= P_1^W + P_2^W.$$

Also,

$$P^V \leq \sum_{k\in\{n\tau+\lambda\Psi_n,\ldots,n-2\}} P\left(\left|\sum_{j=1}^{p} \frac{k-n\tau}{n-n\tau}\left[\left(\sum_{i=k+1}^{n} \eta_{i,j}\right)^2 \frac{1}{n-k}\right.\right.\right.$$

$$\left.\left.\left. - \left(\sum_{i=n\tau+1}^{k} \eta_{i,j}\right)^2 \frac{1}{k-n\tau}\right]\right| > n\Delta_p^2 \frac{(\frac{k}{n}-\tau)n\tau}{8k}\right)$$

$$+ \sum_{k\in\{n\tau+\lambda\Psi_n,\ldots,n-2\}} P\left(\left|\sum_{j=1}^{p} \frac{2}{n-n\tau} \sum_{i=k+1}^{n} \eta_{i,j} \sum_{i=n\tau+1}^{k} \eta_{i,j}\right|\right.$$

$$\left. > n\Delta_p^2 \frac{(\frac{k}{n}-\tau)n\tau}{8k}\right)$$

$$:= P_1^V + P_2^V.$$

Now,

$$
\begin{aligned}
P_1^W &= \sum_{k \in \{n\tau + n\lambda\Psi_n, \ldots, n-2\}} P\Bigg(\Bigg| \sum_{j=1}^{p} \frac{k-n\tau}{k} \Bigg[\Bigg(\sum_{i=n\tau+1}^{k} \eta_{i,j} \Bigg)^2 \frac{1}{\sigma^2(k-n\tau)} \\
&\qquad\qquad\qquad - \Bigg(\sum_{i=1}^{n\tau} \eta_{i,j} \Bigg)^2 \frac{1}{\sigma^2 n\tau} \Bigg] \Bigg| > \frac{n\Delta_p^2}{\sigma^2} \frac{(\frac{k}{n}-\tau)n\tau}{8k} \Bigg) \\
&\leq \sum_{k \in \{n\tau + n\lambda\Psi_n, \ldots, n-2\}} P\Bigg(\Bigg| \sum_{j=1}^{p} \Bigg(\sum_{i=n\tau+1}^{k} \eta_{i,j} \Bigg)^2 \frac{1}{\sigma^2(k-n\tau)} - p \Bigg| > \frac{n\Delta_p^2}{\sigma^2} \frac{\tau}{16} \Bigg) \\
&\quad + \sum_{k \in \{n\tau + n\lambda\Psi_n, \ldots, n-2\}} P\Bigg(\Bigg| \sum_{j=1}^{p} \Bigg(\sum_{i=1}^{n\tau} \eta_{i,j} \Bigg)^2 \frac{1}{\sigma^2 n\tau} - p \Bigg| > \frac{n\Delta_p^2}{\sigma^2} \frac{\tau}{16} \Bigg) \\
&\leq 4n \Bigg[\exp\Bigg(-\frac{(n\Delta_p^2\tau)^2}{16^2\sigma^4 \times 16p} \Bigg) \vee \exp\Bigg(-\frac{n\Delta_p^2\tau}{16\sigma^2 \times 4} \Bigg) \Bigg] \\
&\leq 4n \Bigg[\exp\Bigg(-\frac{\lambda\tau^2}{2^{12}} \Bigg) \vee \exp\Bigg(-\frac{n\Delta_p^2\tau}{2^6\sigma^2} \Bigg) \Bigg] \\
&\leq 4n \Bigg[\exp\Bigg(-\frac{\lambda\varepsilon^2}{2^{12}} \Bigg) \vee \exp\Bigg(-\frac{n\Delta_p^2\varepsilon}{2^6\sigma^2} \Bigg) \Bigg].
\end{aligned}
$$

In the last two bounds, we have first applied Corollary 2, then used the fact that we are in the case $\frac{(n\Delta_p^2)^2}{\sigma^4 p} \geq \lambda$.

Similarly,

$$
\begin{aligned}
P_1^V &= \sum_{k \in \{n\tau + \lambda\Psi_n, \ldots, n-2\}} P\Bigg(\Bigg| \sum_{j=1}^{p} \frac{k-n\tau}{n-n\tau} \Bigg[\Bigg(\sum_{i=k+1}^{n} \eta_{i,j} \Bigg)^2 \frac{1}{\sigma^2(n-k)} \\
&\qquad\qquad\qquad - \Bigg(\sum_{i=n\tau+1}^{k} \eta_{i,j} \Bigg)^2 \frac{1}{\sigma^2(k-n\tau)} \Bigg] \Bigg| > \frac{n\Delta_p^2}{\sigma^2} \frac{(\frac{k}{n}-\tau)n\tau}{8k} \Bigg) \\
&\leq \sum_{k \in \{n\tau + n\lambda\Psi_n, \ldots, n-2\}} P\Bigg(\Bigg| \sum_{j=1}^{p} \Bigg(\sum_{i=k+1}^{n} \eta_{i,j} \Bigg)^2 \frac{1}{\sigma^2(n-k)} - p \Bigg| > \frac{n\Delta_p^2\tau(1-\tau)}{16\sigma^2} \Bigg) \\
&\quad + \sum_{k \in \{n\tau + n\lambda\Psi_n, \ldots, n-2\}} P\Bigg(\Bigg| \sum_{j=1}^{p} \Bigg(\sum_{i=n\tau+1}^{k} \eta_{i,j} \Bigg)^2 \frac{1}{\sigma^2(k-n\tau)} - p \Bigg| \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad > \frac{n\Delta_p^2\tau(1-\tau)}{16\sigma^2} \Bigg).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
P_1^V &\leq 4n\left[\exp\left(-\frac{(n\Delta_p^2)^2\tau^2(1-\tau)^2}{16^2\sigma^4\times 16p}\right)\vee\exp\left(-\frac{n\Delta_p^2\tau(1-\tau)}{4\times 16\sigma^2}\right)\right]\\
&\leq 4n\left[\exp\left(-\frac{\tau^2(1-\tau)^2\lambda}{2^{12}}\right)\vee\exp\left(-\frac{n\Delta_p^2\tau(1-\tau)}{2^6\sigma^2}\right)\right]\\
&\leq 4n\left[\exp\left(-\frac{\lambda\varepsilon^2}{2^{14}}\right)\vee\exp\left(-\frac{n\Delta_p^2\varepsilon}{2^7\sigma^2}\right)\right].
\end{aligned}
$$

For P_2^W, let us denote by $\mathscr{F}$, the $\sigma-$algebra spanned by the variables $\{\eta_{i,j},\ i\leq n\tau,\ j\leq p\}$. We write

$$
P_2^W =
$$

$$
\sum_{k\in\{n\tau+n\lambda\Psi_n,\ldots,n-2\}}E\left[P\left(\left|\sum_{j=1}^{p}\sum_{i=1}^{n\tau}\eta_{i,j}\sum_{i=n\tau+1}^{k}\eta_{i,j}\right|>n\Delta_p^2\frac{(\frac{k}{n}-\tau)n\tau}{16}\Big|\mathscr{F}\right)\right].
$$

Conditionally on $\mathscr{F}$, the random variable $\sum_{j=1}^{p}\sum_{i=1}^{n\tau}\eta_{i,j}\sum_{i=n\tau+1}^{k}\eta_{i,j}$ follows a centered normal distribution $\mathcal{N}(0,\sigma^2(k-n\tau)\sum_{j=1}^{p}(\sum_{i=1}^{n\tau}\eta_{i,j})^2)$, that is

$$
\frac{\sum_{j=1}^{p}\sum_{i=1}^{n\tau}\eta_{i,j}\sum_{i=n\tau+1}^{k}\eta_{i,j}}{\sigma(k-n\tau)^{1/2}(\sum_{j=1}^{p}(\sum_{i=1}^{n\tau}\eta_{i,j})^2)^{1/2}}\sim\mathcal{N}(0,1).
$$

Thus,

$$
\begin{aligned}
P_2^W &\leq 2\sum_{k\in\{n\tau+n\lambda\Psi_n,\ldots,n-2\}}\\
&\quad E\left[\exp\left(-\frac{(n\Delta_p^2(\frac{k}{n}-\tau)n\tau)^2}{16^2}\frac{1}{2\sigma^2(k-n\tau)\sum_{j=1}^{p}(\sum_{i=1}^{n\tau}\eta_{i,j})^2}\right)\right]\\
&\leq 2\sum_{k\in\{n\tau+n\lambda\Psi_n,\ldots,n-2\}}\\
&\quad E\left[\exp\left(-\frac{(n\Delta_p^2(\frac{k}{n}-\tau)n\tau)^2}{16^2}\frac{1}{2\sigma^2(k-n\tau)\sum_{j=1}^{p}(\sum_{i=1}^{n\tau}\eta_{i,j})^2}\right)\right.\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\left.\times\mathbf{1}_{\left\{\sum_{j=1}^{p}\frac{(\sum_{i=1}^{n\tau}\eta_{i,j})^2}{\sigma^2n\tau}\leq 8p\right\}}\right]\\
&\quad +2nP\left(\sum_{j=1}^{p}\frac{\left(\sum_{i=1}^{n\tau}\eta_{i,j}\right)^2}{\sigma^2n\tau}\geq 8p\right).
\end{aligned}
$$

Hence,

$$P_2^W \le 2n \exp\left(-\frac{(n\Delta_p^2)^2 \lambda \Psi_n \tau}{2^{12}\sigma^4 p}\right) + 2n \exp\left(-\frac{7p}{4}\right)$$

$$\le 2n \exp\left(-\frac{\lambda\tau}{2^{12}}\right) + 2n \exp\left(-\frac{n\Delta_p^2 \varepsilon^2}{2^5 \sigma^2}\right)$$

$$\le 2n \exp\left(-\frac{\lambda\varepsilon}{2^{12}}\right) + 2n \exp\left(-\frac{n\Delta_p^2 \varepsilon^2}{2^5 \sigma^2}\right).$$

We used here $n\Delta_p^2 \le 32p\sigma^2/\varepsilon^2$ together with Corollary 2. To end the proof, we investigate the term P_2^V. For $k \in \{n\tau + n\lambda\Psi_n, \dots, n-2\}$, let $\mathscr{F}_k$ denote the $\sigma-$algebra spanned by the variables $\{\eta_{i,j}, \ i > k, \ j \le p\}$.

Conditionally on $\mathscr{F}_k$, the random variable $\sum_{j=1}^p \frac{1}{n-n\tau} \sum_{i=k+1}^n \eta_{i,j} \sum_{i=n\tau+1}^k \eta_{i,j}$ follows a centered normal distribution $\mathcal{N}(0, \frac{\sigma^2(k-n\tau)}{n^2(1-\tau)^2} \sum_{j=1}^p (\sum_{i=k+1}^n \eta_{i,j})^2)$, that is

$$\frac{\sum_{j=1}^p \sum_{i=k+1}^n \eta_{i,j} \sum_{i=n\tau+1}^k \eta_{i,j}}{\sigma(k-n\tau)^{1/2}(\sum_{j=1}^p (\sum_{i=k+1}^n \eta_{i,j})^2)^{1/2}} \sim \mathcal{N}(0,1).$$

Using Gaussian concentration (12), we write:

$$P_2^V = \sum_{k \in \{n\tau + n\lambda\Psi_n, \dots, n-2\}}$$
$$E\left[P\left(\left|\sum_{j=1}^p \frac{1}{n-n\tau} \sum_{i=k+1}^n \eta_{i,j} \sum_{i=n\tau+1}^k \eta_{i,j}\right| > n\Delta_p^2 \frac{(\frac{k}{n}-\tau)n\tau}{16k}\Big|\mathscr{F}_k\right)\right]$$
$$\le 2 \sum_{k \in \{n\tau + n\lambda\Psi_n, \dots, n-2\}}$$
$$E\left[\exp\left(-\frac{(n\Delta_p^2(k-n\tau)(1-\tau)n\tau)^2}{16^2 k^2} \frac{1}{2\sigma^2(k-n\tau)\sum_{j=1}^p (\sum_{i=k+1}^n \eta_{i,j})^2}\right)\right.$$
$$\left. \times \mathbf{1}_{\left\{\sum_{j=1}^p \frac{(\sum_{i=k+1}^n \eta_{i,j})^2}{\sigma^2(n-k)} \le 8p\right\}}\right]$$
$$+ 2nP\left(\sum_{j=1}^p \frac{(\sum_{i=k+1}^n \eta_{i,j})^2}{\sigma^2(n-k)} \ge 8p\right).$$

Hence,

$$
\begin{aligned}
P_2^V &\leq 2n \exp\left(-\frac{(n\Delta_p^2)^2 \lambda \Psi_n \tau^2 (1-\tau)^2}{2^{12}\sigma^4 p}\right) + \exp\left(-\frac{7p}{4}\right) \\
&\leq 2n \exp\left(-\frac{\lambda \tau^2 (1-\tau)^2}{2^{12}}\right) + 2n \exp\left(-\frac{n\Delta_p^2 \varepsilon^2}{2^5 \sigma^2}\right) \\
&\leq 2n \exp\left(-\frac{\lambda \varepsilon^2}{2^{14}}\right) + 2n \exp\left(-\frac{n\Delta_p^2 \varepsilon^2}{2^5 \sigma^2}\right).
\end{aligned}
$$

Again, we used the assumption $n\Delta_p^2 \leq 32p\sigma^2/\varepsilon^2$ and Corollary 2. $\qquad\square$

Proof of Lemma 7. The proof is very similar to the proof of Lemma 6. Recall that

$$
\begin{aligned}
P^{V,W} &\leq P\Bigg(\exists k \in \{n\tau + \lambda\Psi_n, \ldots, n-2\}, \\
&\qquad \sum_{j=1}^p V_j^2\left(\frac{k}{n}\right) - \sum_{j=1}^p V_j^2(\tau) > \frac{n\Delta_p^2}{\sigma^2}\frac{(\frac{k}{n}-\tau)n\tau}{4k}\Bigg) \\
&\quad + P\Bigg(\exists k \in \{n\tau + n\lambda\Psi_n, \ldots, n-2\}, \\
&\qquad \sum_{j=1}^p W_j^2\left(\frac{k}{n}\right) - \sum_{j=1}^p W_j^2(\tau) > \frac{n\Delta_p^2}{\sigma^2}\frac{(\frac{k}{n}-\tau)n\tau}{4k}\Bigg) \\
&:= P^V + P^W.
\end{aligned}
$$

To bound P^V and P^W, we will use the next inequalities:

$$
\begin{aligned}
\sigma^2\left(W_j^2\left(\frac{k}{n}\right) - W_j^2(\tau)\right) &= \left(\sum_{i=1}^{n\tau}\eta_{i,j}\right)^2\left(\frac{1}{k}-\frac{1}{n\tau}\right) + \left(\sum_{i=n\tau+1}^{k}\eta_{i,j}\right)^2\frac{1}{k} \\
&\quad + \frac{2}{k}\sum_{i=1}^{n\tau}\eta_{i,j}\sum_{i=n\tau+1}^{k}\eta_{i,j} \\
&\leq \left(\sum_{i=n\tau+1}^{k}\eta_{i,j}\right)^2\frac{1}{k} + \frac{2}{k}\sum_{i=1}^{n\tau}\eta_{i,j}\sum_{i=n\tau+1}^{k}\eta_{i,j},
\end{aligned}
$$

and

$$\sigma^2 \left(V_j^2 \left(\frac{k}{n} \right) - V_j^2 (\tau) \right)$$

$$= \left(\sum_{i=k+1}^n \eta_{i,j} \right)^2 \left(\frac{1}{n-k} - \frac{1}{n-n\tau} \right) - \left(\sum_{i=n\tau+1}^k \eta_{i,j} \right)^2 \frac{1}{n-n\tau}$$

$$- \frac{2}{n-n\tau} \sum_{i=k+1}^n \eta_{i,j} \sum_{i=n\tau+1}^k \eta_{i,j}$$

$$\leq \left(\sum_{i=k+1}^n \eta_{i,j} \right)^2 \frac{k-n\tau}{(n-k)(n-n\tau)}$$

$$- \frac{2}{n-n\tau} \sum_{i=k+1}^n \eta_{i,j} \sum_{i=n\tau+1}^k \eta_{i,j}.$$

As a consequence, we get

$$P^W \leq \sum_{k \in \{n\tau+n\lambda\Psi_n,\dots,n-2\}} P\left(\left| \sum_{j=1}^p \frac{1}{\sigma^2(k-n\tau)} \left(\sum_{i=n\tau+1}^k \eta_{i,j} \right)^2 - p \right| > \frac{n\Delta_p^2}{\sigma^2} \frac{\tau}{8} \right)$$

$$+ P\left(\left| \sum_{j=1}^p \frac{2}{\sigma^2 k} \sum_{i=1}^{n\tau} \eta_{i,j} \sum_{i=n\tau+1}^k \eta_{i,j} \right| > \frac{n\Delta_p^2}{\sigma^2} \frac{(\frac{k}{n}-\tau)n\tau}{8k} \right)$$

$$:= P_1^W + P_2^W.$$

Likewise,

$$P^V \leq \sum_{k \in \{n\tau+n\lambda\Psi_n,\dots,n-2\}} P\left(\left| \sum_{j=1}^p \frac{1}{\sigma^2(n-k)} \left(\sum_{i=k+1}^n \eta_{i,j} \right)^2 - p \right| \right.$$

$$\left. > \frac{n\Delta_p^2}{\sigma^2} \frac{n\tau(1-\tau)}{8k} \right)$$

$$+ P\left(\left| \sum_{j=1}^p \frac{2}{\sigma^2(n-n\tau)} \sum_{i=k+1}^n \eta_{i,j} \sum_{i=n\tau+1}^k \eta_{i,j} \right| > \frac{n\Delta_p^2}{\sigma^2} \frac{(\frac{k}{n}-\tau)n\tau}{8k} \right)$$

$$:= P_1^V + P_2^V.$$

Now, using Corollary 2, we get, since $n\Delta_p^2 \geq 32p\sigma^2/\varepsilon^2$,

$$P_1^W \leq 2n \exp\left(-\frac{n\Delta_p^2 \tau}{2^5 \sigma^2} \right) \leq 2n \exp\left(-\frac{n\Delta_p^2 \varepsilon}{2^5 \sigma^2} \right).$$

Also, using $n\Delta_p^2 \geq 32p\sigma^2/\varepsilon^2$ again, we have

$$P_1^V \leq nP\left(\left|\sum_{j=1}^p \frac{1}{\sigma^2(n-k)}\left(\sum_{i=k+1}^n \eta_{i,j}\right)^2 - p\right| > \frac{n\Delta_p^2}{\sigma^2}\frac{\tau(1-\tau)}{8}\right)$$

$$\leq 2n\exp\left(-\frac{n\Delta_p^2\tau(1-\tau)}{2^5\sigma^2}\right) \leq 2n\exp\left(-\frac{n\Delta_p^2\varepsilon}{2^6\sigma^2}\right).$$

Now, to address the term P_2^W, let us denote by $\mathscr{F}$ the $\sigma-$algebra spanned by the variables $\{\eta_{i,j},\ i \leq n\tau,\ j \leq p\}$. Using Gaussian concentration (12) as in Lemma 6, we write

$$P_2^W = \sum_{k \in \{n\tau+n\lambda\Psi_n,\ldots,n-2\}}$$

$$E\left[P\left(\frac{1}{\sigma^2}\left|\sum_{j=1}^p \sum_{i=1}^{n\tau}\eta_{i,j}\sum_{i=n\tau+1}^k \eta_{i,j}\right| > \frac{n\Delta_p^2}{\sigma^2}\frac{(\frac{k}{n}-\tau)n\tau}{16}\bigg|\mathscr{F}\right)\right]$$

$$\leq 2\sum_{k \in \{n\tau+n\lambda\Psi_n,\ldots,n-2\}}$$

$$E\left[\exp\left(-\frac{(n\Delta_p^2)^2(k-n\tau)^2\tau^2}{16^2}\frac{1}{2\sigma^2(k-n\tau)\sum_{j=1}^p(\sum_{i=1}^{n\tau}\eta_{i,j})^2}\right)\right]$$

$$\leq 2\sum_{k \in \{n\tau+n\lambda\Psi_n,\ldots,n-2\}}$$

$$E\left[\exp\left(-\frac{(n\Delta_p^2)^2(k-n\tau)\tau^2}{2^9\sigma^2\sum_{j=1}^p(\sum_{i=1}^{n\tau}\eta_{i,j})^2}\right) \times \mathbf{1}_{\left\{\sum_{j=1}^p \frac{(\sum_{i=1}^{n\tau}\eta_{i,j})^2}{n\tau\sigma^2} \leq \frac{n\Delta_p^2\tau}{8\sigma^2}\right\}}\right]$$

$$+ 2nP\left(\sum_{j=1}^p \frac{(\sum_{i=1}^{n\tau}\eta_{i,j})^2}{n\tau\sigma^2} - p \geq \frac{n\Delta_p^2\tau}{8\sigma^2} - p\right)$$

$$\leq 2n\exp\left(-\frac{(n\Delta_p^2)^2 n\Psi_n\lambda\tau^2 \times 8}{2^9\sigma^2 \times n^2\tau^2\Delta_p^2}\right) + 2n\exp\left(-\frac{n\Delta_p^2\tau}{2^5\sigma^2} + \frac{p}{4}\right)$$

$$\leq 2n\exp\left(-\frac{\lambda}{2^6}\right) + 2n\exp\left(-\frac{n\Delta_p^2\varepsilon}{2^6\sigma^2}\right).$$

Here, we used the inequality $n\Delta_p^2 \geq 32p\sigma^2/\varepsilon^2$, and the fact that $\varepsilon \leq \frac{1}{2}$, and applied again Corollary 2. Similarly, for the term P_2^V, let $\mathscr{F}_k$ be the $\sigma-$algebra spanned by the variables $\{\eta_{i,j},\ i > k,\ j \leq p\}$. We write:

$$P_2^V = \sum_{k \in \{n\tau + n\lambda\Psi_n, \ldots, n-2\}}$$

$$E\left[P\left(\left| \sum_{j=1}^p \frac{1}{n - n\tau} \sum_{i=k+1}^n \eta_{i,j} \sum_{i=n\tau+1}^k \eta_{i,j} \right| > n\Delta_p^2 \frac{(\frac{k}{n} - \tau)n\tau}{16k} \Big| \mathscr{F}_k \right) \right]$$

$$\leq 2 \sum_{k \in \{n\tau + n\lambda\Psi_n, \ldots, n-2\}}$$

$$E\left[\exp\left(- \frac{(n\Delta_p^2)^2 (k - n\tau)^2 n^2 (1-\tau)^2 \tau^2}{16^2 k^2} \right. \right.$$

$$\left. \left. \times \frac{1}{2\sigma^2 (k - n\tau) \sum_{j=1}^p (\sum_{i=k+1}^n \eta_{i,j})^2} \right) \right]$$

$$\leq 2 \sum_{k \in \{n\tau + n\lambda\Psi_n, \ldots, n-2\}}$$

$$E\left[\exp\left(- \frac{(n\Delta_p^2)^2 (k - n\tau)(1-\tau)^2 \tau^2}{2^9 \sigma^2 \sum_{j=1}^p (\sum_{i=k+1}^n \eta_{i,j})^2} \right) \right.$$

$$\left. \times \mathbf{1}_{\left\{ \sum_{j=1}^p \frac{(\sum_{i=k+1}^n \eta_{i,j})^2}{\sigma^2(n-k)} \leq \frac{n\Delta_p^2 \tau}{8\sigma^2} \right\}} \right]$$

$$+ 2n P\left(\sum_{j=1}^p \frac{(\sum_{i=k+1}^n \eta_{i,j})^2}{\sigma^2 (n-k)} - p \geq \frac{n\Delta_p^2 \tau}{8\sigma^2} - p \right)$$

$$\leq 2n \exp\left(- \frac{n\Delta_p^2 \lambda \Psi_n \tau (1-\tau)^2 \times 8}{2^9 \sigma^2} \right) + 2n \exp\left(- \frac{n\Delta_p^2 \tau}{2^5 \sigma^2} + \frac{p}{4} \right)$$

$$\leq 2n \exp\left(- \frac{\lambda\tau(1-\tau)^2}{2^6} \right) + 2n \exp\left(- \frac{n\Delta_p^2 \varepsilon}{2^6 \sigma^2} \right)$$

$$\leq 2n \exp\left(- \frac{\lambda\varepsilon^2}{2^7} \right) + 2n \exp\left(- \frac{n\Delta_p^2 \varepsilon}{2^6 \sigma^2} \right).$$

$$\square$$

5.2. Proof of Theorem 2

Let us in this section define $p_s := \left(\frac{\sigma^2 \ln(d \vee n)}{n} \right)^{\frac{-1}{1+2s}}$. The following Lemma is essential in the sequel.

Lemma 8. *We consider Model (1), and assume that θ^+ and θ^- belong to*

$\Theta(s, L)$. *We suppose that there exists a constant $\alpha > 0$ such that*

$$\frac{n}{\sigma^2} \geq \alpha \ln d.$$

Then, for any γ, if $C_{\mathcal{L}}$ is large enough (see condition (10) below), there exists a constant $R = R(\gamma, L, C_{\mathcal{L}}, \varepsilon)$ (see condition (11)) such that, if

$$\Delta^2 \geq R \left(\frac{\sigma^2 \ln(d \vee n)}{n} \right)^{\frac{2s}{1+2s}}, \tag{9}$$

then we have

$$P \left(\left\{ \Delta_{\hat{p}}^2 \geq \frac{\Delta^2}{2} \right\} \cap \{ \hat{p} \leq p_s \} \right) \geq 1 - c'n^{-\gamma},$$

where c' is an absolute constant.

The proof is based on an intermediate lemma, stating that, with large probability, $\hat{p} \leq p_s$.

Lemma 9. *Under the conditions above, for any γ, if we have*

$$C_{\mathcal{L}} \geq 16 \vee 4L^2 \vee \frac{20(\gamma+2)}{\alpha^{\frac{1}{1+2s}}} \vee \frac{2^4 L(\gamma+2)^{1/2}}{\alpha^{\frac{1}{2(1+2s)}}}, \tag{10}$$

then

$$P(\hat{p} > p_s) \leq 3(d \vee n)^{-\gamma}.$$

Recall that Z is defined by

$$Z_j = \frac{1}{n} \sum_{i=1}^{n} Y_i^j - \frac{2}{n} \sum_{i=1}^{n/2} Y_i^j, \quad j = 1, \dots, d,$$

that is $Z_j = \beta_j + \varepsilon_j$, $j = 1, \dots, d$, where $\beta_j = (1-\tau)(\theta_j^+ - \theta_j^-)\mathbf{1}_{\{\tau \geq 1/2\}} + \tau(\theta_j^+ - \theta_j^-)\mathbf{1}_{\{\tau < 1/2\}}$ and $\varepsilon_j \sim \mathcal{N}(0, \frac{\sigma^2}{n})$.

Lemma 10. *For $p_s \leq \ell \leq k$,*

$$P \left(\left| \sum_{j=\ell}^{k} (Z_j)^2 - \sum_{j=\ell}^{k} (\beta_j)^2 \right| > x \right) \leq 2 \exp \left(-\frac{x^2}{2^6 L^2 p_s^{-2s} \frac{\sigma^2}{n}} \right) + \exp \left(-\frac{nx}{10\sigma^2} \right),$$

as soon as $x \geq 10(k - \ell + 1)\sigma^2/n$.

Proof of Lemma 10.

$$P\left(\left|\sum_{j=\ell}^{k}(Z_j)^2 - \sum_{j=\ell}^{k}(\beta_j)^2\right| > x\right)$$

$$\leq P\left(\sum_{j=\ell}^{k}(\varepsilon_j)^2 + 2\left|\sum_{j=\ell}^{k}\varepsilon_j\beta_j\right| > x\right)$$

$$\leq P\left(\sum_{j=\ell}^{k}(\varepsilon_j)^2 > x/2\right) + P\left(\left|\sum_{j=\ell}^{k}\varepsilon_j\beta_j\right| > x/4\right).$$

Observe that $\sum_{j=\ell}^{k}\varepsilon_j\beta_j$ follows a Gaussian distribution $\mathcal{N}\left(0, \frac{\sigma^2}{n}\sum_{j=\ell}^{k}(\beta_j)^2\right)$, so that, using the concentration of the Gaussian distribution (see (12) in the Appendix) and the fact that $\sum_{j=\ell}^{k}(\beta_j)^2 \leq 2L^2 p_s^{-2s}$, since θ^+ and θ^- are in $\Theta(s, L)$, we obtain

$$P\left(\left|\sum_{j=\ell}^{k}\varepsilon_j\beta_j\right| > x/4\right) \leq 2\exp\left(-\frac{x^2}{2^6 L^2 p_s^{-2s}\frac{\sigma^2}{n}}\right).$$

Using Corollary 2 in the Appendix, we get

$$P\left(\sum_{j=\ell}^{k}(\varepsilon_j)^2 > x/2\right) \leq P\left(\sum_{j=\ell}^{k}\frac{n}{\sigma^2}(\varepsilon_j)^2 - k + \ell - 1 > \frac{nx}{2\sigma^2} - k + \ell - 1\right)$$

$$\leq \exp\left(-\frac{nx}{10\sigma^2}\right),$$

as soon as $x \geq 10(k - \ell + 1)\sigma^2/n$. $\qquad\square$

Proof of Lemma 9. We have

$$P(\hat{p} > p_s) \leq P\left(\exists k \geq \ell \geq p_s, \sum_{j=\ell}^{k}(Z_j)^2 > C_{\mathcal{L}}k\frac{\sigma^2}{n}\ln(d \vee n)\right).$$

Now, since $k \geq p_s$,

$$\sum_{j=\ell}^{k}(\beta_j)^2 \leq 2L^2 p_s^{-2s} = 2L^2\left(\frac{\sigma^2 \ln(d \vee n)}{n}\right)^{\frac{2s}{1+2s}} = 2L^2 p_s \frac{\sigma^2 \ln(d \vee n)}{n}.$$

Thus, if $C_{\mathcal{L}} \geq 4L^2$, we have $\sum_{j=\ell}^{k}(\beta_j)^2 \leq (C_{\mathcal{L}}/2)k\frac{\sigma^2}{n}\ln(d \vee n)$. We get, with $2x := C_{\mathcal{L}}k\frac{\sigma^2}{n}\ln(d \vee n)$, the following inequality:

$$P\left(\sum_{j=\ell}^{k}(Z_j)^2 > 2x\right) \leq P\left(\left|\sum_{j=\ell}^{k}(Z_j)^2 - \sum_{j=\ell}^{k}(\beta_j)^2\right| > x\right).$$

Now, from Lemma 10, as soon as $C_{\mathcal{L}} k \ln(d \vee n) \geq 20(k - \ell + 1)$ (which is always true for instance if $C_{\mathcal{L}} \geq 20$), we have

$$
\begin{aligned}
P(\hat{p} > p_s) & \\
&\leq \sum_{k \geq \ell \geq p_s} P\left(\sum_{j=\ell}^{k} (Z_j)^2 > C_{\mathcal{L}} k \frac{\sigma^2}{n} \ln(d \vee n)\right) \\
&\leq \sum_{k \geq \ell \geq p_s} 2\exp\left(-\frac{(C_{\mathcal{L}} k \frac{\sigma^2}{n} \ln(d \vee n))^2}{2^8 L^2 p_s^{-2s} \frac{\sigma^2}{n}}\right) + \exp\left(-C_{\mathcal{L}} k \ln(d \vee n)/20\right) \\
&\leq d^2 \left[2\exp\left(-\frac{(C_{\mathcal{L}} p_s \frac{\sigma^2}{n} \ln(d \vee n))^2}{2^8 L^2 p_s^{-2s} \frac{\sigma^2}{n}}\right) + \exp\left(-C_{\mathcal{L}} p_s \ln(d \vee n)/20\right)\right] \\
&\leq d^2 \left[2\exp\left(-\frac{C_{\mathcal{L}}^2}{2^8 L^2}\left(\frac{\sigma^2}{n}\right)^{-\frac{1}{1+2s}} (\ln(d \vee n))^{\frac{2s}{1+2s}}\right)\right. \\
&\qquad \left. + \exp\left(-\frac{C_{\mathcal{L}}}{20}\left(\frac{\sigma^2}{n}\right)^{-\frac{1}{1+2s}} (\ln(d \vee n))^{\frac{2s}{1+2s}}\right)\right] \\
&\leq d^2 \left[2\exp\left(-\frac{C_{\mathcal{L}}^2}{2^8 L^2} \alpha^{\frac{1}{1+2s}} \ln(d \vee n)\right) + \exp\left(-\frac{C_{\mathcal{L}}}{20} \alpha^{\frac{1}{1+2s}} \ln(d \vee n)\right)\right].
\end{aligned}
$$

This last term is less than $3(d \vee n)^{-\gamma}$, as, by assumption, $\frac{C_{\mathcal{L}} \alpha^{\frac{1}{1+2s}}}{20} > \gamma + 2$, and $\frac{C_{\mathcal{L}}^2 \alpha^{\frac{1}{1+2s}}}{2^8 L^2} > \gamma + 2$. $\qquad \square$

Equipped with Lemma 9, let us go back to the proof of Lemma 8.

Proof of Lemma 8. To simplify the exposition, let us suppose that $\tau \geq 1/2$, the other case can be treated similarly, with elementary modifications.
We may write

$$
\Delta_{\hat{p}}^2 = \Delta^2 - \sum_{j=\hat{p}+1}^{p_s} (\theta_j^+ - \theta_j^-)^2 - \sum_{j=p_s+1}^{d} (\theta_j^+ - \theta_j^-)^2.
$$

Yet,

$$
\sum_{k=p_s+1}^{d} (\theta_j^+ - \theta_j^-)^2 \leq 4L^2 p_s^{-2s} = 4L^2 \left(\frac{\sigma^2 \ln(d \vee n)}{n}\right)^{\frac{2s}{1+2s}} \leq \Delta^2/4,
$$

as soon as $R \geq 16L^2$, since $\Delta^2 \geq R\left(\frac{\sigma^2 \ln(d \vee n)}{n}\right)^{\frac{2s}{1+2s}}$ (condition (9)). then,

$$
P\left(\{\hat{p} \leq p_s\} \cap \{\Delta_{\hat{p}}^2 \leq \Delta^2/2\}\right) \leq P\left(\{\hat{p} \leq p_s\} \cap \left\{\sum_{j=\hat{p}+1}^{p_s} (\theta_j^+ - \theta_j^-)^2 \geq \Delta^2/4\right\}\right).
$$

Now,

$$P\left(\{\hat{p} \le p_s\} \cap \left\{\sum_{j=\hat{p}+1}^{p_s} (\theta_j^+ - \theta_j^-)^2 \ge \Delta^2/4\right\}\right)$$

$$\le P\left(\{\hat{p} \le p_s\} \cap \left\{\sum_{j=\hat{p}+1}^{p_s} \frac{(Z^j)^2}{(1-\tau)^2} + \left[(\theta_+^j - \theta_-^j)^2 - \frac{(Z^j)^2}{(1-\tau)^2}\right] \ge \Delta^2/4\right\}\right).$$

From the construction of $\hat{p}$, we know that

$$\sum_{j=\hat{p}+1}^{p_s} \frac{(Z^j)^2}{(1-\tau)^2} \le \frac{C_{\mathcal{L}} p_s}{(1-\tau)^2} \frac{\sigma^2}{n} \ln(d \vee n) = \frac{C_{\mathcal{L}}}{(1-\tau)^2} \left(\frac{\sigma^2 \ln(d \vee n)}{n}\right)^{\frac{2s}{1+2s}}.$$

Thus, if $C_{\mathcal{L}}/(1-\tau)^2 \le R/8$,

$$\sum_{j=\hat{p}+1}^{p_s} \frac{(Z^j)^2}{(1-\tau)^2} \le \Delta^2/8.$$

Consequently,

$$P\left(\{\hat{p} \le p_s\} \cap \{\Delta_{\hat{p}}^2 \le \Delta^2/2\}\right)$$

$$\le P\left(\{\hat{p} \le p_s\} \cap \left\{\sum_{j=\hat{p}+1}^{p_s} \left[(\theta_+^j - \theta_-^j)^2 - \frac{(Z^j)^2}{(1-\tau)^2}\right] \ge \Delta^2/8\right\}\right).$$

$$\le \sum_{k=1}^{p_s} P\left(\{\hat{p} = k\} \cap \left\{\left|\sum_{j=k+1}^{p_s} \left[(\theta_+^j - \theta_-^j)^2 - \frac{(Z^j)^2}{(1-\tau)^2}\right]\right| \ge \Delta^2/8\right\}\right)$$

$$\le \sum_{k=1}^{p_s} P\left(\left|\sum_{j=k+1}^{p_s} \left[(\beta_j)^2 - (Z^j)^2\right]\right| \ge (1-\tau)^2 \Delta^2/8\right)$$

$$\le \sum_{k=1}^{p_s} P\left(\sum_{j=k+1}^{p_s} (\varepsilon_j)^2 + 2\left|\sum_{j=k+1}^{p_s} \varepsilon_j \beta_j\right| \ge (1-\tau)^2 \Delta^2/8\right).$$

Now, we use again Gaussian and chi-square concentration results, as in the proof of Lemma 10.
As soon as $(1-\tau)^2 n\Delta^2/(80\sigma^2) \ge p_s$, which is always true if $R \ge \frac{80}{(1-\tau)^2}$, we obtain:

$$\sum_{k=1}^{p_s} P\left(\sum_{j=k+1}^{p_s} (\varepsilon_j)^2 \ge (1-\tau)^2 \Delta^2/16\right) \le p_s \exp\left(-\frac{(1-\tau)^2 n\Delta^2}{80\sigma^2}\right).$$

Moreover, since $\text{Var}\left(\sum_{j=k+1}^{p_s} \varepsilon_j \beta_j\right) = \frac{\sigma^2}{n} \sum_{j=k+1}^{p_s} (\beta_j)^2 \le \frac{\sigma^2 \Delta^2 (1-\tau)^2}{n}$,

$$\sum_{k=1}^{p_s} P\left(\left|\sum_{j=k+1}^{p_s} \varepsilon_j \beta_j\right| \ge (1-\tau)^2 \Delta^2/2^5\right) \le 2p_s \exp\left(-\frac{(1-\tau)^2 n\Delta^2}{2^{11}\sigma^2}\right).$$

Now, $\frac{n\Delta^2}{\sigma^2} \geq R\ln(d \vee n)p_s$. Hence, for R large enough, the right-hand terms admit an upper bound of the order $n^{-\gamma}$. Combining this bounds with Lemma 9, we get the desired result, as soon as Conditions (10) and

$$R \geq 16L^2 \vee \frac{8C_{\mathcal{L}}}{\varepsilon^2} \vee \frac{2^{11}(\gamma+1)}{\varepsilon^2} \tag{11}$$

are satisfied. $\qquad\square$

Proof of Theorem 2. We use Lemma 8, the definition of p_s, and Proposition 1. For any γ, γ', we have

$$P\left(|\hat{\tau}(\hat{p}) - \tau| \geq \kappa(\gamma, \varepsilon)\frac{\sigma^2 \ln(n)}{n\Delta^2}\right)$$

$$\leq P\left(\left\{|\hat{\tau}(\hat{p}) - \tau| \geq \kappa(\gamma, \varepsilon)\frac{\sigma^2 \ln(n)}{n\Delta^2}\right\} \cap \left\{\Delta_{\hat{p}}^2 \geq \Delta^2/2\right\} \cap \{\hat{p} \leq p_s\}\right) + c'n^{-\gamma}$$

$$\leq \sum_{p=1}^{p_s} P\left(\left\{|\hat{\tau}(\hat{p}) - \tau| \geq \kappa(\gamma, \varepsilon)\frac{\sigma^2 \ln(n)}{n\Delta^2}\right\} \cap \left\{\Delta_{\hat{p}}^2 \geq \Delta^2/2\right\} \cap \{\hat{p} = p\}\right) + c'n^{-\gamma'}$$

$$\leq \sum_{p=1}^{p_s} P\left(\left\{|\hat{\tau}(p) - \tau| \geq \kappa(\gamma, \varepsilon)\frac{\sigma^2 \ln(n)}{2n\Delta_p^2}\right\} \cap \left\{\Delta_p^2 \geq \Delta^2/2\right\} \cap \{\hat{p} = p\}\right) + c'n^{-\gamma'}$$

$$\leq \sum_{p=1}^{p_s} P\bigg(\left\{|\hat{\tau}(p) - \tau| \geq \kappa(\gamma, \varepsilon)\frac{\sigma^2 \ln(n)}{2n\Delta^2}\right\} \cap \left\{\Delta_p^2 \geq \frac{Rp\sigma^2 \ln(d \vee n)}{2n}\right\}$$

$$\cap \{\hat{p} = p\}\bigg) + c'n^{-\gamma'}$$

$$\leq p_s cn^{-\gamma} + c'n^{-\gamma'},$$

as soon as R is large enough, which proves the theorem. $\qquad\square$

Appendix A: Appendix: concentration inequalities

Gaussian concentration   If $N \sim \mathcal{N}(0,1)$, then it is well known that, for $x > 0$,

$$P(|N| > x) \leq 2\exp\left(-\frac{x^2}{2}\right). \tag{12}$$

Concentration for the chi-square distribution (moderate and large deviations)   The next lemma is proved in [37].

Lemma 11. *Let k be a positive integer and U be a χ^2 distribution with k degrees of freedom. For $z > 0$,*

$$P(U - k \geq 2\sqrt{kz} + 2z) \leq \exp(-z),$$
$$P(k - U \geq 2\sqrt{kz}) \leq \exp(-z).$$

In this paper, the following form of the result is used.

Corollary 2. *Let k be a positive integer and U be a χ^2 distribution with k degrees of freedom. For $0 < x \leq 4k$,*

$$P(U - k \geq x) \leq \exp\left(-\frac{x^2}{16k}\right).$$

For $x \geq 4k$,

$$P(U - k \geq x) \leq \exp\left(-\frac{x}{4}\right).$$

For $x > 0$,

$$P(k - U \geq x) \leq \exp\left(-\frac{x^2}{4k}\right).$$

Consequently,

$$P(|U - k| \geq x) \leq 2\left[\exp\left(-\frac{x^2}{16k}\right) \vee \exp\left(-\frac{x}{4}\right)\right].$$

References

[1] S. Arlot and A. Celisse. Segmentation of the mean of heteroscedastic data via cross-validation. *Stat. Comput.*, 21(4):613–632, 2011. ISSN 0960-3174.

[2] J. A. D. Aston and C. Kirch. Evaluating stationarity via change-point alternatives with applications to fmri data. *Ann. Appl. Stat.*, 6(4):1906–1948, 2012.

[3] A. Aue, S. Hörmann, L. Horváth, M. Reimherr, et al. Break detection in the covariance structure of multivariate time series models. *The Annals of Statistics*, 37(6B):4046–4087, 2009.

[4] J. Bai. Least square estimation of a shift in linear processes. *Journal of Time Series Analysis*, 15:453–472, 1994.

[5] J. Bai. Common breaks in means and variances for panel data. *Journal of Econometrics*, 157(1):78–92, 2010.

[6] J. Bai and P. Perron. Estimating and testing linear models with multiple structural changes. *Econometrica*, 66:47–78, 1998.

[7] M. Basseville and I. Nikiforov. *Detection of abrupt changes: Theory and applications*. Lecture Notes in Mathematics. Prentice Hall, 1993.

[8] K. Bleakley and J. P. Vert. The group fused Lasso for multiple change-point detection, 2011. Available at https://hal.inria.fr/hal-00602121.

[9] S. Bourguignon, D. Mary, and E. Slezak. Restoration of astrophysical spectra with sparsity constraints: models and algorithms. *J. Sel. Topics Signal Processing*, 5:1002–1013, 2011.

[10] B. E. Brodsky and B. S. Darkhovsky. *Nonparametric methods in change-point problems*. The Netherlands, 1993.

[11] A. Bücher, I. Kojadinovic, T. Rohmer, and J. Segers. Detecting changes in cross-sectional dependence in multivariate time series. *Journal of Multivariate Analysis*, 132:111–128, 2014.

[12] E. Carlstein, H. Müller, and D. Siegmund, editors. *Change-point problems*, volume 23. 1994.

[13] H. Chen and N. Zhang. Graph-based change-point detection. *Ann. Statist.*, 43:139–176, 2015.

[14] H. Cho. Change-point detection in panel data via double cusum statistic. *Electron. J. Statist.*, 10(2):2000–2038, 2016.

[15] H. Cho and P. Fryzlewicz. Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):475–507, 2015.

[16] I. Cribben and Y. Yu. Estimating whole-brain dynamics by using spectral clustering. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(3):607–627.

[17] M. Csörgő and L. Horváth. *Limit theorems in change-point analysis*. 1997.

[18] F. Enikeeva and Z. Harchaoui. High-dimensional change-point detection with sparse alternatives, 2017. Available at https://hal.inria.fr/hal-00933185.

[19] J. Fiosina and M. Fiosins. Resampling-based change point estimation. In J. Gama, E. Bradley, and J. Hollmén, editors, *Advances in Intelligent Data Analysis X*, pages 150–161, Berlin, Heidelberg, 2011. Springer.

[20] K. Frick, A. Munk, and H. Sieling. Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):495–580, 2014.

[21] P. Fryzlewicz et al. Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281, 2014.

[22] Z. Harchaoui, F. Vallet, A. Lung-Yut-Fong, and O. Cappe. A regularized kernel-based approach to unsupervised audio segmentation. In *ICASSP*, pages 1665–1668, 2009.

[23] S. Hariz and J. J. Wylie. Rates of convergence for the change-point estimator for long-range dependent sequences. *Statistics & Probability Letters*, 73:155–164, 06 2005.

[24] S. B. Hariz, J. J. Wylie, and Q. Zhang. Optimal rate of convergence for nonparametric change-point estimators for nonstationary sequences. *Ann. Statist.*, 35:1802–1826, 2007.

[25] S. S. Henry, D. and R. J. Patton. Fault detection and diagnosis for aeronautic and aerospace missions. In T. L. C. Edwards and H. Smaili, editors, *Fault Tolerant Flight Control – a Benchmark Challenge*, pages 91–128, Berlin, 2010. Springer.

[26] L. Horváth and M. Hušková. Change-point detection in panel data. *Journal of Time Series Analysis*, 33(4):631–648, 2012.

[27] L. Horváth and G. Rice. Extensions of some classical methods in change point analysis. *Test*, 23(2):219–255, 2014.

[28] L. Horváth, P. Kokoszka, and J. Steinebach. Testing for changes in multivariate dependent observations with an application to temperature changes. *Journal of Multivariate Analysis*, 68(1):96–119, 1999.

[29] J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity, 2009.

[30] B. Jin, G. Pan, Q. Yang, and W. Zhou. On high-dimensional change point problem. *Science China Mathematics*, 59:2355–2378, 2016.

[31] M. Jirak et al. Uniform change point tests in high dimension. *The Annals of Statistics*, 43(6):2451–2483, 2015.

[32] I. M. Johnstone and B. W. Silverman. Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.*, 32 (4):1594–1649, 2004. ISSN 0090-5364.

[33] R. Killick, P. Fearnhead, and I. A. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

[34] C. Kirch, B. Muhsal, and H. Ombao. Detection of changes in multivariate time series with application to eeg data. *Journal of the American Statistical Association*, 110(511):1197–1216, 2015.

[35] A. Korostelev and O. Lepski. On a multi-channel change-point problem. *Math. Methods Statist.*, 17(3):187–197, Sep 2008.

[36] A. P. Korostelev. On minimax estimation of a discontinuous signal. *Theory Probab. Appl.*, 32:727– 730, 1987.

[37] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28:1302–1338, 2000.

[38] M. Lavielle and E. Moulines. Least-squares estimation of an unknown number of shifts in a time series. *Journal of Time Series Analysis*, 21: 33–59, 2000.

[39] M. Lavielle and G. Teyssière. Detection of multiple change-points in multivariate time series. *Lithuanian Mathematical Journal*, 46:287–306, 2006.

[40] O. Lepski. On a problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.*, 35:454–466, 1991.

[41] O. Lepski. Asymptotically minimax adaptive estimation. I: Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.*, 36:682–697, 1992.

[42] O. Lepski. Asymptotically minimax adaptive estimation. II. Schemes without optimal adaptation: Adaptive estimators. *Theory Probab. Appl.*, 37:

433–448, 1993.

[43] C. Levy-Leduc and F. Roueff. Detection and localization of change-points in high-dimensional network traffic data. *Ann. Appl. Stat.*, 3:637–662, 2009.

[44] A. Lung-Yut-Fong, C. Levy-Leduc, and O. Cappe. Distributed detection/localization of change-points in high-dimensional network traffic data. *Stat. Comput.*, 22:485–496, 2012.

[45] C. Meillier, F. Chatelain, O. Michel, R. Bacon, L. Piqueras, R. Bacher, and H. Ayasso. SELFI: an object-based, Bayesian method for faint emission line source detection in MUSE deep eld data cubes. *Astronomy and Astrophysics*, 588:140A, 2016.

[46] H. Müller. Change-points in nonparametric regression. *Ann. Statist.*, 20: 737–671, 1992.

[47] A. B. Olshen, E. Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5(4):557–572, 2004.

[48] H. Ombao, R. Von Sachs, and W. Guo. Slex analysis of multivariate non-stationary time series. *Journal of the American Statistical Association*, 100 (470):519–531, 2005.

[49] E. Page. A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3/4):523–527, 1955.

[50] P. Preuss, R. Puchstein, and H. Dette. Detection of multiple structural breaks in multivariate time series. *Journal of the American Statistical Association*, 110(510):654–668, 2015.

[51] Y. Ritov. Asymptotic efficient estimation of the change point with unknown distributions. *The Annals of Statistics*, 18:1829–1839, 1990.

[52] X. Shi, Y. Wu, and C. R. Rao. Consistent and powerful graph-based change-point test for high-dimensional data. *Proceedings of the National Academy of Sciences*, 114:3873–3878, 2017.

[53] A. N. Shiryaev. *Optimal Stopping Rules*. New York, 1978.

[54] Y. S. Soh and V. Chandrasekaran. High-dimensional change-point estimation: Combining filtering with convex optimization. *Applied and Computational Harmonic Analysis*, 43(1):122 − 147, 2017.

[55] R. Sparks, T. Keighley, and D. Muscatello. Early warning cusum plans for surveillance of negative binomial daily disease counts. *J. Appl. Statist.*, 37: 1911–1930, 2010.

[56] R. F. R. Suleiman, D. Mary, and A. Ferrari. Dimension reduction for hypothesis testing in worst-case scenarios. *IEEE Transactions on Signal Processing*, 62:5973–5986, 2014.

[57] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 2008. ISBN 0387790519, 9780387790510.

[58] D. Wang, Y. Yu, and A. Rinaldo. Optimal change point detection and localization in sparse dynamic networks. 2018. Available at https://arxiv.org/abs/1809.09602.

[59] T. Wang and R. J. Samworth. High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 80:57–83, 2018.

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