Variance Change Point Detection under A Smoothly-changing Mean Trend with Application to Liver Procurement

Zhenguo Gao
Department of Statistics, Virginia Tech,
Zuofeng Shang
Department of Mathematical Sciences, SUNY at Binghamton
and
Pang Du*
Department of Statistics, Virginia Tech,

March 28, 2017

Abstract

Literature on change point analysis mostly require a sudden change in the data distribution, either in a few parameters or the distribution as a whole. We are interested in the scenario that the variance of data may make a significant jump while the mean of data changes in a smooth fashion. It is motivated from a liver procurement experiment with organ surface temperature monitoring. Blindly applying the existing change point analysis methods to the example can yield erratic change point estimates since the smoothly-changing mean violates the sudden-change assumption. In this paper we propose a penalized weighted least squares approach with an iterative estimation procedure that naturally integrates variance change point detection and smooth mean function estimation. Given the variance components the mean function is estimated by smoothing splines as the minimizer of the penalized weighted least

*Du's research was supported by U.S. National Science Foundation under grant DMS-1620945.

squares. Given the mean function, we propose a likelihood ratio test statistic for identifying the variance change point. The null distribution of the test statistic is derived together with the rates of convergence of all the parameter estimates. Simulations show excellent performance of the proposed method. Application analysis offers numerical support to the non-invasive organ viability assessment by surface temperature monitoring.

Keywords: Variance change point; Smoothly-changing mean trend; Hypothesis testing in nonparametric smoothing; Change point consistency; Asymptotic null distribution.

1 Introduction

Change point detection is a classical topic that has attracted a lot of attention for decades. Efforts have mostly focused on detection of sudden changes in a few parameters, such as the mean and/or variance, of the underlying distribution, or the distribution itself as a whole entity. In this paper, we are concerned with variance change point detection under a smoothly-changing mean trend. Particularly, the constantly changing mean trend violates the assumptions of most existing change point detection methods. As demonstrated in the paper, a naïve application of these existing methods to such kind of data would yield erratic change point estimates.

Our method is motivated from an experiment about the procurement of transplant livers. Quality/viability evaluation is a key issue in the procurement of transplant organs. Currently, such evaluations are mostly performed through visual inspection by surgeons or biopsy image assessment by pathologists. Both approaches are subjective judgements. Biopsy is more accurate than surgeons' visual inspection, but it is also invasive and damages the part of the organ where the biopsy sample is collected. And the viability status of the biopsy sample may not represent that of the whole organ. In the experiment considered in the paper, surface temperature of a severed porcine liver was constantly monitored upon the infusion of the perfusion liquid to the organ. The measurements consisted of surface temperatures measured every 10 minutes on a dense grid covering the whole organ for a span of 24 hours. The left panels in Figure 1 were the temperature profiles for three spots on the surface. The temperature of the perfusion liquid was often slightly different from the body temperature. So the temperature of the organ changed in a slow fashion and displayed an overall smooth mean trend. The high oscillations in the first half reflected the resistance of the organ to the abrupt temperature change in the environment. Around the

10th hour, the organ started to lose its viability and this change was reflected in a sudden drop in the variance of the temperature, as shown in the plot of residuals versus time in the right panels of Figure 1. Our goal is to design a testing procedure for identifying the variance change point of the residuals after removing the smoothly changing mean trend.

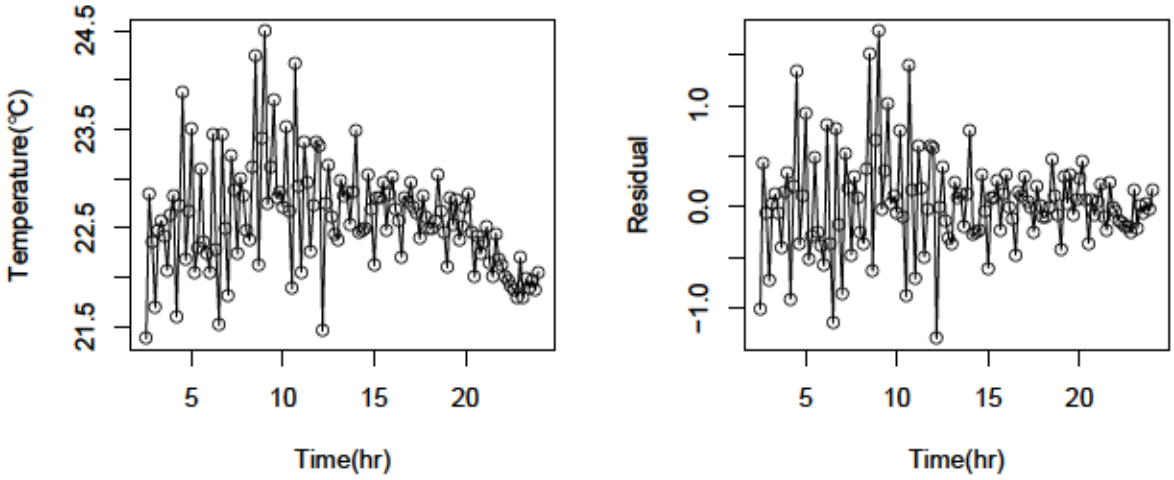

Figure 1: Raw temperature profiles (left panels) and detrended temperature profiles (right panels) at three spots of the liver. The x-axis labels in both panels represent 24 hours.

Note that this phenomenon of having a variance change point underlying a smooth mean trend actually occurs in many other settings besides the liver procurement experiment considered here. For example, seismic activity monitoring often sees a smooth mean trend with small variation and a sudden change in variation could be the early sign of an earthquake; the EEG signal for an epilepsy patient generally shows a smooth mean trend and a sudden variation change in the signal might mean the onset of a seizure; the stock

price for a big company often shows a smooth mean trend and a sudden increase in variation could mean a turmoil on the stock market or stock holders' rising panic about the company's health. So the change point detection procedure proposed here is a new method that arises naturally from our motivating example on liver procurement and can be also applied to many other areas.

The existing literature on change point analysis can be roughly divided into two categories. In the domain of parametric change point analysis, researchers assume that the underlying distribution belongs to some known family and sudden shift changes in the mean, variance, or both are considered. For example, when change of variance is the only concern, two representative approaches are the cumulative sum of squares approach in Inclán and Tiao (1994) and the Schwartz information criterion in Chen and Gupta (1997). When simultaneous shifts in mean and variance are considered, Horváth (1993) and Pan and Chen (2006) studied the theoretical properties of likelihood ratio test and modified information criterion respectively. One can refer to Chen and Gupta (2012) for a comprehensive list of publications in parametric change point analysis. In the domain of nonparametric change point analysis, the assumption is that there is a sudden change in the probability distribution of the data. Various measures for such change have been developed in the literature to describe the differences between probability distributions. For example, Hariz et al. (2007) developed a semi-norm to measure the difference between empirical probability distributions and estimated the change point as the position where a weighted version of such difference is maximized. Matteson and James (2014) used hierarchical clustering to estimate the number of change points and their positions simultaneously for multivariate data. However, none of these existing methods in these two domains can address the problem in our experiment where the variance change happened underneath a smoothly-changing mean trend. Particularly, a smooth mean trend implies that the mean,

and thus the distribution of the data, are constantly changing over time besides the sudden change in variance. Neither the parametric nor the nonparametric change point analysis methods can capture the gradually changing mean trend. As demonstrated in our numerical experiments, erratic behavior occurs when blindly applying these methods to such kind of data ignoring the underlying smooth mean trend.

Nonparametric smoothing and change point detection are often viewed as two conflicting issues in statistics since the former emphasizes on continuity and the latter represents discontinuity. The variance change point detection method proposed here naturally integrates these two domains in both numerical and theoretical senses. There has been other work combining nonparametric regression with change point detection. For example, both Loader (1996) and Grégoire and Hamrouni (2002) considered the problem of detecting jump points in smooth curves. However, they both focused on jumps in the mean curve whereas our application clearly showed a jump in the variance. So the method proposed in this paper is uniquely suited to tackling the change point problem found in our liver procurement experiment.

Our variance change point detection method is formulated under the framework of penalized weighted least squares estimation. Particularly, the estimates of the mean function, the change point, and the variances are a local minimizer of a penalized weighted least squares score whose global minimizer may not exist. This objective functional consists of three parts: the weighted sum of squared errors represents the goodness-of-fit, the roughness penalty on the mean function estimate enforces smoothness on the mean, and the smoothing parameter balances the tradeoff. The optimization of the objective functional is carried out in an iterative fashion starting with a consistent initial mean estimate. When the mean function is given, the variance change point and the corresponding variances are estimated through a testing procedure generalizing the one in Chen and Gupta (1997).

When the variance change point and the variances of two subsequences of data are given, the mean function is estimated by smoothing splines through the standard optimization of the penalized weighted least squares with known weights. The initial mean estimate is the minimizer of the penalized least squares under the working independence assumption.

For theoretical properties, we derive the asymptotic null distribution of our test statistic for the variance change point and we show that our change point estimate is consistent when the function space for the mean function is a periodic Soblev space. We note that these results have their own theoretical values too. Testing procedures under nonparametric null and alternative hypotheses are very difficult problems since both the null and alternative spaces are of infinite dimensions. They become even harder in the penalized estimation scenario since the smoothing parameter in the penalty adds additional complexity to the derivation of asymptotic theory. For example, the rigorous theory for statistical inference with smoothing spline regression under the constant variance assumption was established by Shang and Cheng (2013) only a few years ago. And their work focused on the inference of the mean function. But our work studies hypothesis testing on the variance component. Our consistency result on the mean and variance component estimates is also new. Recognizing that the global minimizer of the penalized weighted least squares may not exist, we have proved the consistency of the estimates obtained from an iterative algorithm starting with a consistent initial mean estimate. This opens a new venue for studying the asymptotic theory of a nonparametric regression model when the random errors are not IID. So the theoretical developments here are novel and nontrivial.

In our simulations, we first demonstrate the pitfall of blindly applying the existing change point procedures without removing the smoothly-changing mean trend when such a trend is present. Then we show the excellent performance of our method in estimating the variance change point, the mean functions and the variances. The application of our

method to the temperature profiles collected in the liver procurement experiment yield critical information about the viability status of the organ. In summary, our method has the following distinguishing features: (1) it is uniquely qualified to address the scientific hypothesis raised in our application experiment; (2) it is an innovative addition to the existing rich literature on change point analysis, (3) it naturally integrates smoothing and change point analysis in a way distinct from others, and (4) its theoretical development opens new fronts for the inference theory of nonparametric smoothing.

The rest of the paper is organized as follows. In Section 2, we introduce in the order: the notation and model, the iterative algorithm, the mean estimation given the variances and change point, the test procedure for variance change point give the mean function, and the theoretical properties of the proposed method. In Section 3 we present all the simulations. We analyze the liver procurement data in Section 4. Discussion in Section 5 concludes the paper. Proofs of the theorems are collected in the Appendix.

2 Method

2.1 Notation and Model

Suppose that y_i are independent observations generated from the following model

$$y_i = f_0(i/n) + \epsilon_i, i = 1, \ldots, n, \tag{1}$$

where f_0 is an unknown smooth function, $\epsilon_i \sim N(0, \sigma_i^2)$ with $\sigma_i = \sigma_0$ when $i \leq \tau_0$ and $\sigma_i = \delta_0$ when $i > \tau_0$, $\sigma_0^2 \neq \delta_0^2$ are unknown variances, and τ_0 is the unknown variance change point. Assume that f_0 belongs to a reproducing kernel Hilbert space $\mathcal{H} = \{f | f : [0,1] \to \mathbb{R}, J(f) < \infty\}$, where J is a semi-norm on $\mathcal{H}$. For example, we consider $J(f) = \int_0^1 \{f^{(m)}(t)\}^2 dt$ in this paper for some positive integer m. We propose to

estimate $(f_0, \tau_0, \sigma_0^2, \delta_0^2)$ through the minimization of the penalized weighted least squares

$$\frac{1}{n}(\mathbf{y}-\mathbf{f})^T\Sigma_{n,\tau,\sigma,\delta}^{-1}(\mathbf{y}-\mathbf{f})+\lambda J(f), \tag{2}$$

where f is a function in $\mathcal{H}$, $\mathbf{y}=(y_1,\ldots,y_n)^T$ and $\mathbf{f}=(f(1/n),f(2/n),\ldots,f(1))^T$ are respectively the vectors of observed responses and fitted values, $\Sigma_{n,\tau,\sigma,\delta}$ is a diagonal matrix with the first τ diagonals equal to σ^2 and the rest equal to δ^2, $J(f)$ acts as a roughness penalty, and $\lambda>0$ is the smoothing parameter balancing the tradeoff between the smoothness of the mean function estimate and the goodness-of-fit represented by the weighted sum of squared errors.

We note that the global minimizer of (2) does not exist since it approaches zero as σ^2 goes to infinity. Hence, we propose the estimates $(\widehat{f},\widehat{\tau},\widehat{\sigma}^2,\widehat{\delta}^2)$ as the local minimizer of (2) obtained through the following iterative algorithm. We shall show in Section 2.4 that the estimates are consistent with proper rates of convergence.

Algorithm.

1. Initialize $\widehat{f}^{(0)}$ with the mean function estimate assuming constant variance. That is, $\widehat{f}^{(0)}$ minimizes
$$\frac{1}{n}(\mathbf{y}-\mathbf{f})^T(\mathbf{y}-\mathbf{f})+\lambda J(f). \tag{3}$$
Note that when $\sigma^2=\delta^2$, the covariance matrix in (2) reduces to $\sigma^2 I$ and σ^2 can be absorbed into the smoothing parameter λ.

2. Each iteration consists of two steps. At the ιth iteration,

   (a) given the mean estimate $\widehat{f}^{(\iota-1)}$, we first use the testing procedure in Section 2.3 to find an estimate $\widehat{\tau}^{(\iota)}$ for τ_0. Then we estimate the variance parameters respectively by the maximum likelihood variance estimates, $[\widehat{\sigma}^2]^{(\iota)}$ and $[\widehat{\delta}^2]^{(\iota)}$, of the

subsequences of residuals, $\{y_i - \widehat{f}^{(\iota-1)}(i/n) : i = 1, \ldots, \widehat{\tau}^{(\iota)}\}$ and $\{y_i - \widehat{f}^{(\iota-1)}(i/n) : i = \widehat{\tau}^{(\iota)} + 1, \ldots, n\}$.

(b) Now given the estimates $\widehat{\tau}^{(\iota)}$, $[\widehat{\sigma}^2]^{(\iota)}$ and $[\widehat{\delta^2}]^{(\iota)}$, we update the mean estimate by the minimizer of (2) where τ, σ^2 and δ^2 are replaced respectively by their current estimates.

3. Iterate until the algorithm converges.

2.2 Mean Estimation Given τ, σ^2, and δ^2

When τ, σ^2 and δ^2 are given, the mean function f_0 is estimated as the minimizer of the penalized weighted least squares (2) in a reproducing kernel Hilbert space $\mathcal{H}$ of functions on the domain $\mathcal{T}$. A reproducing kernel Hilbert space (RKHS) is a Hilbert space $\mathcal{H}$ where the evaluation functional $[t] : \mathcal{H} \to \mathbb{R}, f \mapsto f(t)$ is continuous for every $t \in \mathcal{T}$. The Riesz Representation Theorem then indicates that for all $t \in \mathcal{T}$ there exists a unique function $R_t \in \mathcal{H}$ with the reproducing property $\langle R_t, f \rangle = [t](f) = f(t)$, where $\langle \cdot, \cdot \rangle$ is the inner product on $\mathcal{H}$. Now the reproducing kernel R of $\mathcal{H}$ is defined as a function $R : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ such that $R(s,t) = \langle R_s, R_t \rangle$. One can show that each RKHS is uniquely associated with a reproducing kernel and vice versa.

Note that the penalty functional J in (2) is a squared semi-norm on $\mathcal{H}$. The null space of J, namely $\mathcal{N}_J = \{f : J(f) = 0\}$, induces a direct sum decomposition $\mathcal{H} = \mathcal{N}_J \oplus \mathcal{H}_J$, where $\mathcal{H}_J$ is the complement of $\mathcal{N}_J$ in $\mathcal{H}$. This then yields a decomposition of the reproducing kernel $R = R_0 + R_J$, where R_0 and R_J are respectively the reproducing kernels on the subspaces $\mathcal{N}_J$ and $\mathcal{H}_J$. See, e.g., Gu (2013, Chapter 2) for more details on RKHSs.

We now introduce an example of cubic smoothing splines to illustrate these concepts. We shall use the cubic smoothing splines in all the numerical studies of the paper.

Example 2.1 (Cubic Smoothing Splines). *Without loss of generality assume* $\mathcal{T} = [0, 1]$. *A choice of* $J(f)$ *is* $\int_0^1 (f'')^2 dt$, *which yields the popular cubic splines. If the inner product in* $\mathcal{N}_J$ *is* $(\int_0^1 f\,dt)(\int_0^1 g\,dt) + (\int_0^1 f'\,dt)(\int_0^1 g'\,dt)$, *then* $\mathcal{H}_J = \mathcal{H} \ominus \mathcal{N}_J = \{f : \int_0^1 f dt = \int_0^1 f' dt = 0, J(f) < \infty\}$ *and the reproducing kernel* $R_J(s,t) = k_2(s)k_2(t) - k_4(|s-t|)$, *where* $k_\nu(t) = B_\nu(t)/\nu!$ *are scaled Bernoulli polynomials for* $t \in [0,1]$. *The null space* $\mathcal{N}_J$ *has a basis* $\{1, k_1(t)\}$ *of 2 functions, where* $k_1(t) = t - 0.5$ *for* $t \in [0,1]$. *See Gu (2013, Section 2.3.3).* □

The RKHS $\mathcal{H}$ is of infinite dimensions, so a direct optimization of (2) on $\mathcal{H}$ seems infeasible. However, since the weighted least squares part in (2) depends on f only through its evaluations at the observation points $t_i, i = 1, \ldots, n$, the Representer Theorem (Wahba, 1990) guarantees that the exact minimizer of (2) actually resides in a finite dimensional subspace of $\mathcal{H}$, namely, $\mathcal{N}_J \oplus \text{span}\{R_J(t_1, \cdot), \ldots, R_J(t_n, \cdot)\}$. Let $\phi_l, l = 1, \ldots, m$ be the basis functions of N_J and $\xi_j = R_J(t_j, \cdot), j = 1, \ldots, n$. Write $f = \boldsymbol{\phi}^T \mathbf{d} + \boldsymbol{\xi}^T \mathbf{c}$, where $\mathbf{c}$ and $\mathbf{d}$ are the corresponding coefficient vectors. Also note that $J(f)$ can be written as a quadratic form $J(f) = \mathbf{c}^T Q \mathbf{c}$, where Q is the $n \times n$ matrix with the (i,j)th entry equal to $R_J(t_i, t_j)$. So for a fixed λ, the objective function (2) is reduced to a quadratic function of the coefficient vectors $\mathbf{c}$ and $\mathbf{d}$. Its minimizer can be obtained analytically. To select the smoothing parameter λ, an outer loop for minimizing the generalized cross-validation (GCV) score is sufficient for the job; see Gu (2013, Chapter 3).

2.3 Variance Change Point Detection Given f

Given $\widehat{f}$, we now introduce a testing procedure to find an estimate $\widehat{\tau}$ for the variance change point τ_0. Then we compute the maximum likelihood estimates for σ^2 and δ^2 respectively by $\widehat{\sigma}^2 = \widehat{\tau}^{-1} \sum_{i=1}^{\widehat{\tau}} \{y_i - \widehat{f}(i/n)\}^2$ and $\widehat{\delta}^2 = (n - \widehat{\tau})^{-1} \sum_{i=\widehat{\tau}+1}^{n} \{y_i - \widehat{f}(i/n)\}^2$. We propose a

testing procedure that generalizes the one introduced by Chen and Gupta (1997) for the parametric case of normal data with a fixed mean.

We want to test the hypothesis

$$H_0 : \sigma_1^2 = \cdots = \sigma_n^2 \text{ versus } H_1 : \sigma_1^2 = \cdots = \sigma_\tau^2 \neq \sigma_{\tau+1}^2 = \cdots = \sigma_n^2, \tag{4}$$

for a potential change point position τ. Let

$$\ell(\tau) = \tau \log\Big[\frac{1}{\tau}\sum_{i=1}^{\tau}\{y_i - \widehat{f}(i/n)\}^2\Big] + (n-\tau)\log\Big[\frac{1}{n-\tau}\sum_{i=\tau+1}^{n}\{y_i - \widehat{f}(i/n)\}^2\Big].$$

Note that $\ell(n) = -2L_0(\widehat{\sigma}^2) - n - n\log 2\pi$ and $\ell(\tau) = -2L_1(\widehat{\sigma}^2, \widehat{\delta}^2) - n - n\log 2\pi$, where L_0 and L_1 are respectively the log likelihood functions under the null and alternative hypotheses of (4). So we define the test statistic to be $\Delta_n^2 = \max_{1<\tau<n}\{\ell(n) - \ell(\tau)\}$.

To gain further insight for the test statistic Δ_n^2, we recap the motivation illustrated in Chen and Gupta (1997) by referring to the Schwartz information criterion (SIC) from Schwarz (1978). As a criterion for model selection, the SIC is defined as $-2\log L(\widehat{\boldsymbol{\theta}}) + p\log n$, where $L(\widehat{\boldsymbol{\theta}})$ is the likelihood function for the model, $\widehat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of the parameter θ, and p is the dimension of $\boldsymbol{\theta}$. In our case, given f and τ we have two models corresponding to the null and alternative hypotheses with their SICs respectively defined by $\mathrm{SIC}(n) = -2L_0(\widehat{\sigma}^2) + \log n$ and $\mathrm{SIC}(\tau) = -2L_1(\widehat{\sigma}^2, \widehat{\delta}^2) + 2\log n$. By the principle of minimum information criterion, we do not reject H_0 if $\mathrm{SIC}(n) \leq \min_\tau \mathrm{SIC}(\tau)$, or equivalently $\ell(n) \leq \min_{1<\tau<n}\ell(\tau)$, and reject H_0 if $\mathrm{SIC}(n) > \mathrm{SIC}(\tau)$ for some τ, or equivalently $\ell(n) > \ell(\tau)$ for some τ. In the case of rejection(s), we estimate the position of change point by $\widehat{\tau} = \arg\min_{1<\tau<n}\ell(\tau)$. So our test statistic can also be written as $\Delta_n^2 = \log n - \min_{1<\tau<n}\{\mathrm{SIC}(\tau) - \mathrm{SIC}(n)\}$. We shall present the asymptotic distribution of Δ_n^2 under the null hypothesis in Section 2.4.

2.4 Theoretical Properties

In this section we present the asymptotic theories for the proposed method. For simplicity, we only consider the special case when $\mathcal{H}$ is the mth order Sobolev space of periodic functions on $[0,1]$ with period 1, namely,

$$\mathcal{H} = S^m \equiv \Big\{ f : f(t) = \sum_{\nu=1}^{\infty} f_\nu \varphi_\nu(t) \text{ with } t \in [0,1] \text{ and } \sum_{\nu=1}^{\infty} f_\nu^2 \gamma_\nu < \infty \Big\},$$

where for $k = 1, 2, \ldots$, $\varphi_{2k-1}(t) = \sqrt{2}\cos(2\pi kt)$, $\varphi_{2k}(t) = \sqrt{2}\sin(2\pi kt)$, and $\gamma_{2k-1} = \gamma_{2k} = (2\pi k)^{2m}$. Note that $J(f) = \int_0^1 \{f^{(m)}(t)\}^2 dt = \sum_{\nu=1}^{\infty} f_\nu^2 \gamma_\nu$ for $f \in S^m$ and $R_J(s,t) = (2\pi m)^{-2m} \sum_{\nu=1}^{\infty} 2\cos\{2\pi\nu(s-t)\}/(2\pi m\nu)^{2m}$.

Let $h = \lambda^{1/(2m)}$, $r_n = \sqrt{\log n/(nh)} + h^{m-1/2}$, and $\widetilde{r}_n = r_n^2 + (nh)^{-3/4} + (\log n)^5 (\log\log n)^2/n + n^{-1/2}$. We shall first show the consistency of the estimates $(\widehat{f}, \widehat{\tau}, \widehat{\sigma}^2, \widehat{\delta}^2)$.

Theorem 2.1 (Consistency of Parameter Estimates). *Under Conditions 1-3 in the Appendix, the estimates $(\widehat{f}, \widehat{\tau}, \widehat{\sigma}^2, \widehat{\delta}^2)$ from the algorithm in Section 2.1 are consistent with the following rates of convergence:*

$$\|\widehat{f} - f_0\|_n^2 = O_P(\lambda + (nh)^{-1} + h^{-1}\widetilde{r}_n^2), \quad |\widehat{\tau} - \tau_0| = O_P((\log n)^4 (\log\log n)^2),$$
$$|\widehat{\sigma}^2 - \sigma_0^2| = O_P(\widetilde{r}_n), \quad |\widehat{\delta}^2 - \delta_0^2| = O_P(\widetilde{r}_n),$$

where $\|f\|_n = \sqrt{\sum_{i=1}^n f(i/n)^2/n}$ is the empirical norm of a function f.

Note that when $m \geq 1$ and $\lambda \asymp n^{-2m/(2m+1)}$, it can be verified that $\widetilde{r}_n = O(n^{-1/2})$. Then this implies that $\widehat{\sigma}^2$ and $\widehat{\delta}^2$ are $\sqrt{n}$-consistent, and that $\|\widehat{f} - f_0\|_n = O_P(n^{-m/(2m+1)})$ or $\widehat{f}$ achieves the optimal convergence rate of a spline function estimate.

We then derive the asymptotic sampling distribution of the test statistic Δ_n^2 under the null hypothesis H_0 in (4).

Theorem 2.2 (Asymptotic Null Distribution of Test Statistic). *Suppose that as* $n \to 0$, $h \to 0$ *and* $r_n^2 \log n \to 0$. *Under* H_0 *in* (4) *and Conditions 1, 2, and 3' in the Appendix, for any* $t \in \mathbb{R}$,

$$P(a_n(\log n)^{1/2}\Delta_n - b_n \log n \leq t) \to \exp(-2\exp(-t)),$$

where $a_n = (2\log\log n)^{1/2}/\log n$, *and* $b_n = \{2\log\log n + \frac{1}{2}\log\log\log n - \log\Gamma(1/2)\}/\log n$.

The limit distribution turns out to be an extreme value distribution. Based on this result, we propose the following testing rule at the significance level $1-\alpha$:

$$\text{Reject } H_0 \Leftrightarrow a_n(\log n)^{1/2}\Delta_n - b_n \log n > -\log\{-\log(1-\alpha)/2\}.$$

3 Simulations

We compared the change point estimation performance of the proposed variance detection method with two existing change point detection methods, one from the parametric domain and the other from the nonparametric domain. The parametric method is the SIC approach in Chen and Gupta (1997) hereafter denoted by the CG method. We used the implementation in the `changepoint` package of R. The nonparametric method is the hierarchical clustering approach in Matteson and James (2014) hereafter denoted by the MJ method. We used the authors' implementation in their R package `ecp`. Furthermore, we examined the performance of the proposed method in estimating the mean curve and the variances.

We considered two mean functions $f_{01}(t) = 20 + 12t(1-t)$ and $f_{02}(t) = \sin(t) + t^5 - 8t^3 + 10t + 6$. The first function f_{01} had a trend similar to the mean temperature profile in the liver procurement study and the second function f_{02} represented a more complex

smooth trend. Two sample sizes $n = 130$ and 500 were used. The true variance change point was set at $\tau_0 = 65$ when $n = 130$, and $\tau_0 = 250$ when $n = 500$. The true variances were $\sigma_0^2 = 0.219$ and $\delta_0^2 = 0.057$ when f_{01} was the true mean function, and $\sigma_0^2 = 9$ and $\delta_0^2 = 2$ when f_{02} was the true mean function. We simulated 1000 data replicates for each combination of mean function and sample size.

For each data replicate, we applied the three variance change point detection methods to obtain the change point estimate. These estimates were divided by n to rescale them to the range of $(0, 1)$ for easier comparison. For the proposed method, we also obtained the mean function estimate and the two variance estimates. To evaluate their performances, we computed the mean squared error MSE $= n^{-1}\sum_{i=1}^{n}\{\widehat{f}(i/n) - f_0(i/n)\}^2$ and the log ratios $\log(\widehat{\sigma}^2/\sigma_0^2)$ and $\log(\widehat{\delta}^2/\delta_0^2)$.

Figure 2 displayed the boxplots of change point estimates from the three methods. We can clearly see that both the CG and the MJ methods suffered when blindly applied to the data without removing the mean trend. On the other hand, the proposed method did a decent job in estimating the location of the change point. And the estimation accuracy clearly improved as the sample size n increased from 130 to 500.

Figure 3 assesses the performance of mean estimation. The top panels plotted the mean estimates that attained the 25th, 50th and 75th percentiles of the MSEs for sample sizes $n = 130$ and 500. The mean function estimates all matched well with the true functions. The 75th percentile estimate for the true function f_{01} with $n = 130$ was slightly off in the area around the change point, which was reasonable considering the fluctuations in that area. Also, the estimation accuracies improved as the sample size increased.

Figure 4 uses the log ratios of variance estimates versus true variances to assess the estimation performance for both variances. We can see that both variances were accurately estimated with the accuracies also improved as the sample size increased.

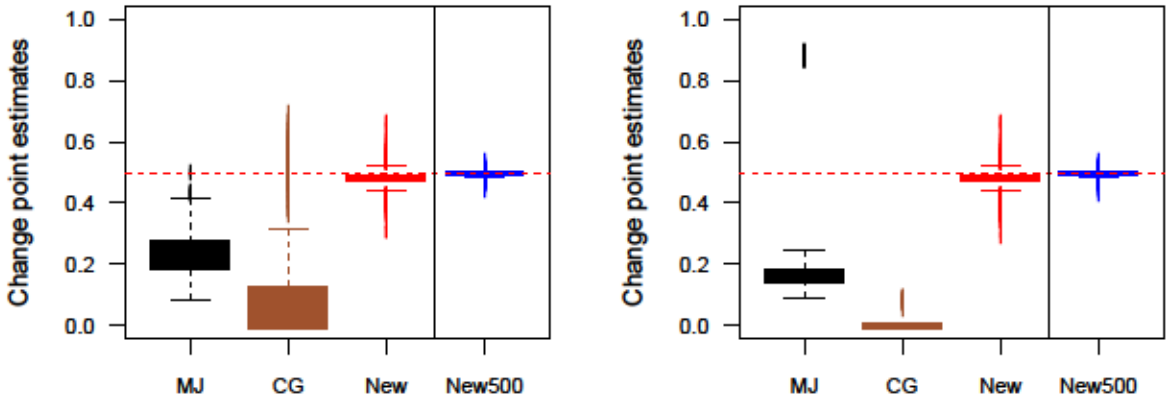

Figure 2: Boxplots of change point estimates. Left panel: Simulations with the true mean function= f_{01}; Right panel: Simulations with the true mean function= f_{02}. The three plots on the left in each panel were the change point estimates with $n = 130$ respectively for the methods in Matteson and James (2014) (MJ), Chen and Gupta (1997) (CG), and the newly proposed method (New). The rightmost plot in each panel was the proposed method with $n = 500$ (New500). The red dashed line is the true change point $\tau_0/n = 0.5$

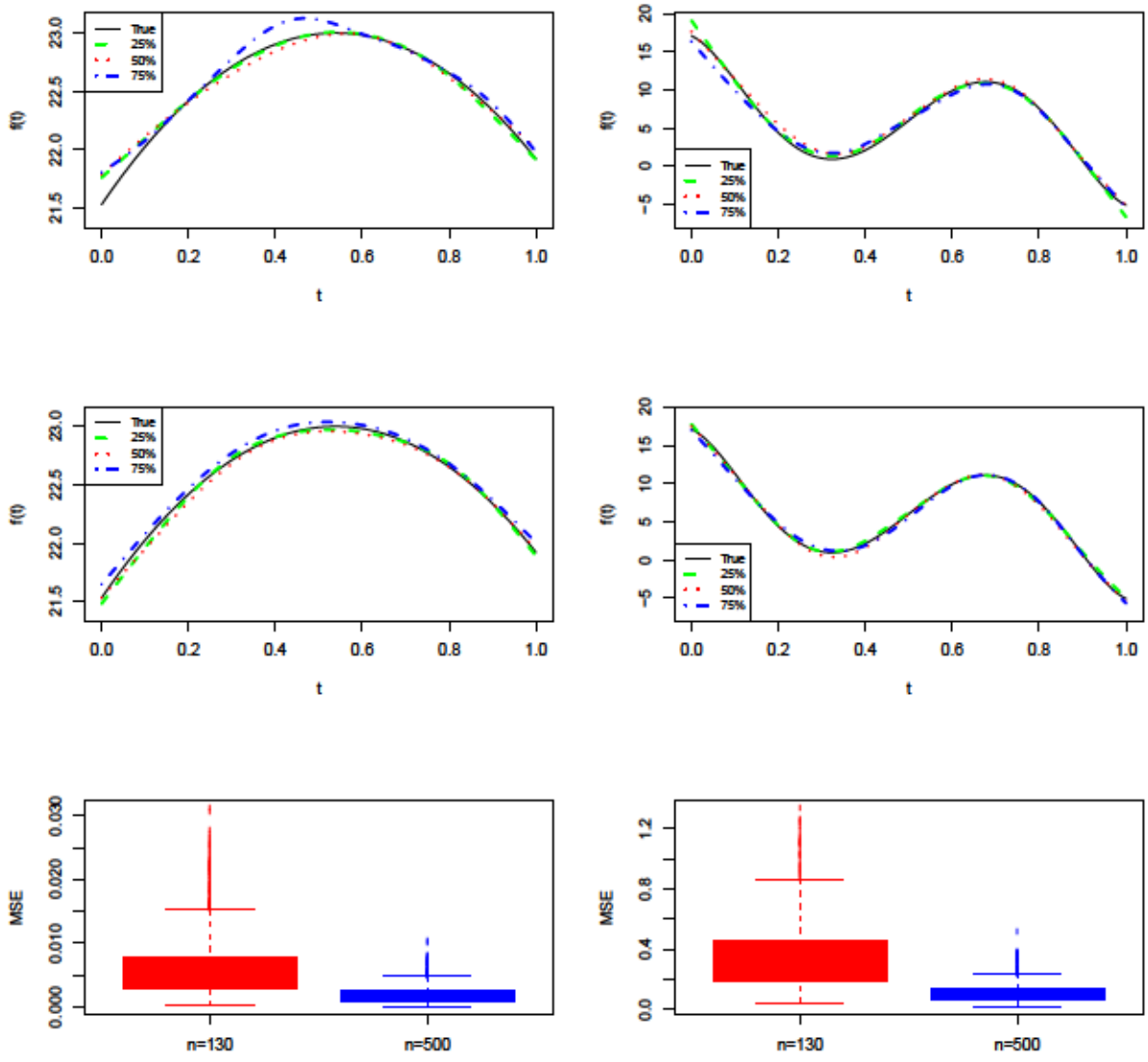

Figure 3: Plots for assessing mean estimation performance. Left panels: Simulations with the true mean function= f_{01}; Right panels: Simulations with the true mean function= f_{02}. Top: True mean function (solid black) versus the mean estimates with $n = 130$ whose MSE were the 25th (dashed green), 50th (dotted red), and 75th (dot-dashed blue) percentiles of the 1000 MSEs obtained in each setting. Middle: same as top but with $n = 500$. Bottom: boxplots of the 1000 MSEs in each setting.

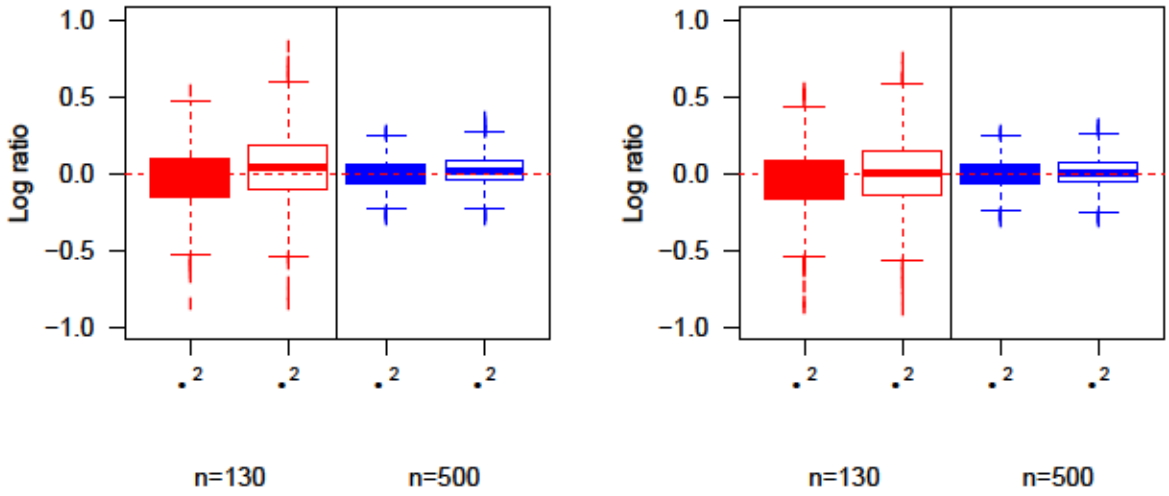

Figure 4: Boxplots of the log ratios of variance estimates versus true variances. Left panels: Simulations with the true mean function= f_{01}; Right panels: Simulations with the true mean function= f_{02}. Red: $n = 130$; Blue: $n = 500$. Filled boxes: σ^2; Unfilled boxes: δ^2.

4 Application: Temperature Monitoring in Liver Procurement

Viability assessment is a critical step in organ transplant procedures. The current assessment procedure purely relies on visual inspection of physicians or biopsy. While the former suffers from subjective judgement, the latter is an intrusive approach that destroys the part of organ where the biopsy sample is collected. Aimed to find a new noninvasive way of assessing the viability of organs, a biomedical engineering team at Virginia Tech designed a temperature monitoring system such that the surface temperature of a perfused organ can be densely and continuously monitored. In the experiment considered in this paper, a lobe of porcine liver, as shown in Figure 5, was perfused in a standard kind of perfusion fluid. Its surface temperature was intensively monitored for a continuing period of 24 hours. The liver lobe was divided into a dense grid of 36,795 spots with each spot producing a 24-hour temperature profile. The temperature measurements were collected every 10 minutes, yielding a total of 145 points in each profile. The first 2.5 hours of data were discarded since it took about one to two hours for the perfusion fluid to completely soak the liver. The data before the liver getting soaked were not of interest. So we had $n = 130$ points left in each profile.

We applied the proposed variance change point detection method to the 36,795 temperature profiles in the data. Since a large number of hypothesis tests were involved here, we considered the Benjamini-Hochberg-Yekutieli (BHY) procedure (Benjamini and Yekutieli, 2001) to address the multiple comparison issue with the control of false discovery rate. This procedure is an extension of the well-known Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) to the case of dependent tests. Due to the positive correlation between our temperature profiles, we used the positive dependency version of the procedure with the

false discovery rate controlled at level 0.05. The largest p-value among all the 36,795 tests of variance change points was 0.019. Hence all the change points were legitimate following the principle of the BHY procedure.

The heat map of all the estimated change points were plotted in Figure 5. Note that an earlier change point in variance meant an earlier drop in the viability of the cells around the spot. We can see that the top half and the middle bottom parts of the liver mostly failed around 12 hours while the bottom left and right portions of the liver lasted beyond 14 hours. There were also a couple of clearly visible straight-line type of boundaries between the early and late failure areas. These might be the part where the porcine liver lobe was bent between the time of severing and perfusion.

Figure 6 plotted the mean estimates and variance change point estimates at three randomly selected spots, imposed respectively on the raw and de-trended temperature profiles. All the mean estimates matched well with the trends shown in the data. As we can see, the mean temperature increased at different paces at the three spots in the first 12 hours or so and shared a common trend of a quicker drop in the second half of the 24-hour period. The variance change points at the three points were all between 12 and 15 hours.

5 Conclusion

In this article, we have presented a new variance change point detection method when the underlying mean trend changes smoothly. Motivated from a liver procurement experiment, the proposed method naturally integrates the seemingly conflicting goals of estimating a smooth mean and detecting a jump point in variance under the framework of penalized weighted least squares. As demonstrated in the simulations, this is not something that can be handled by the existing change point detection methods. Furthermore, the testing pro-

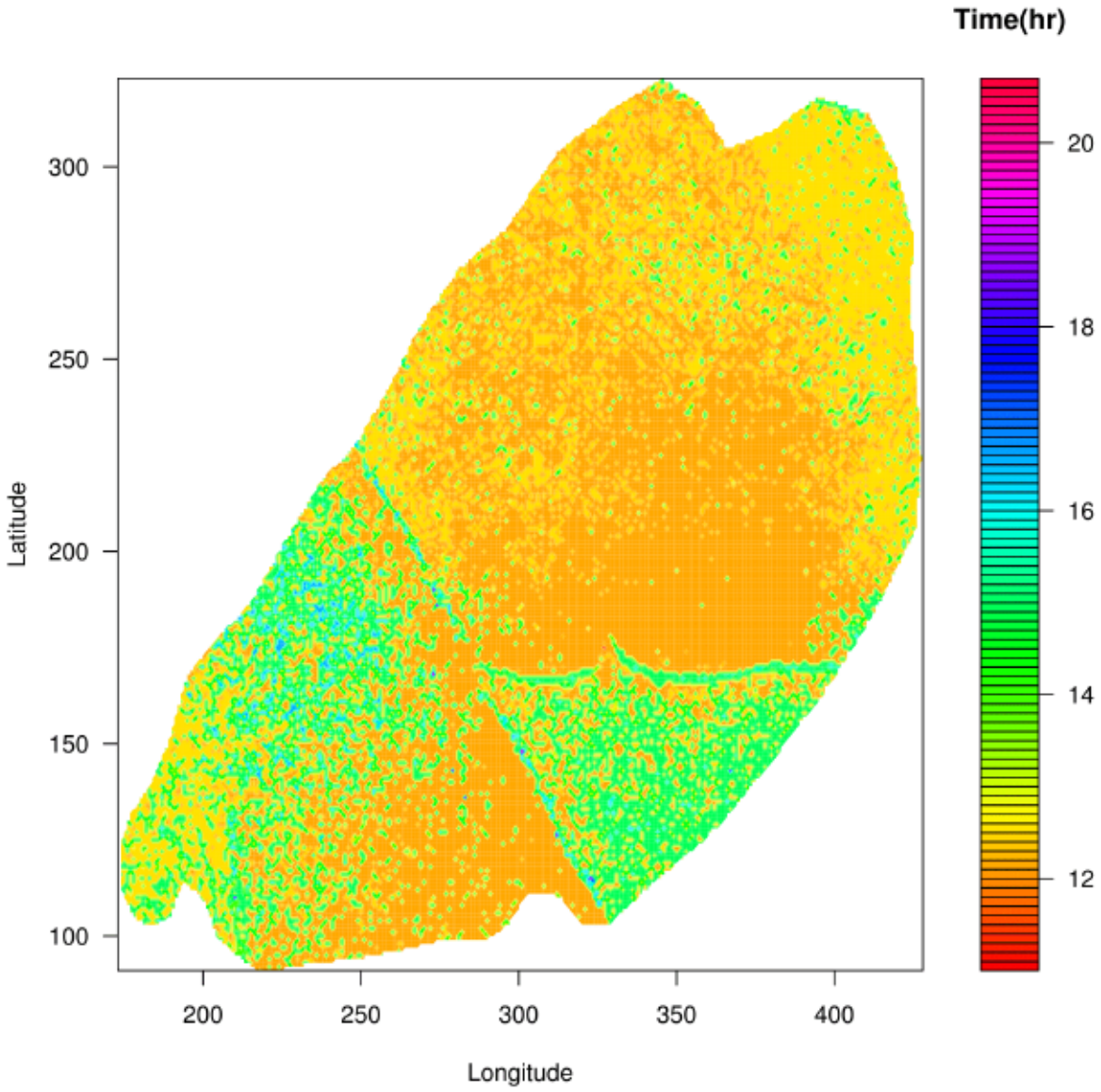

Figure 5: The heat map of estimated variance change points of temperatures on the lobe of liver in the procurement experiment.

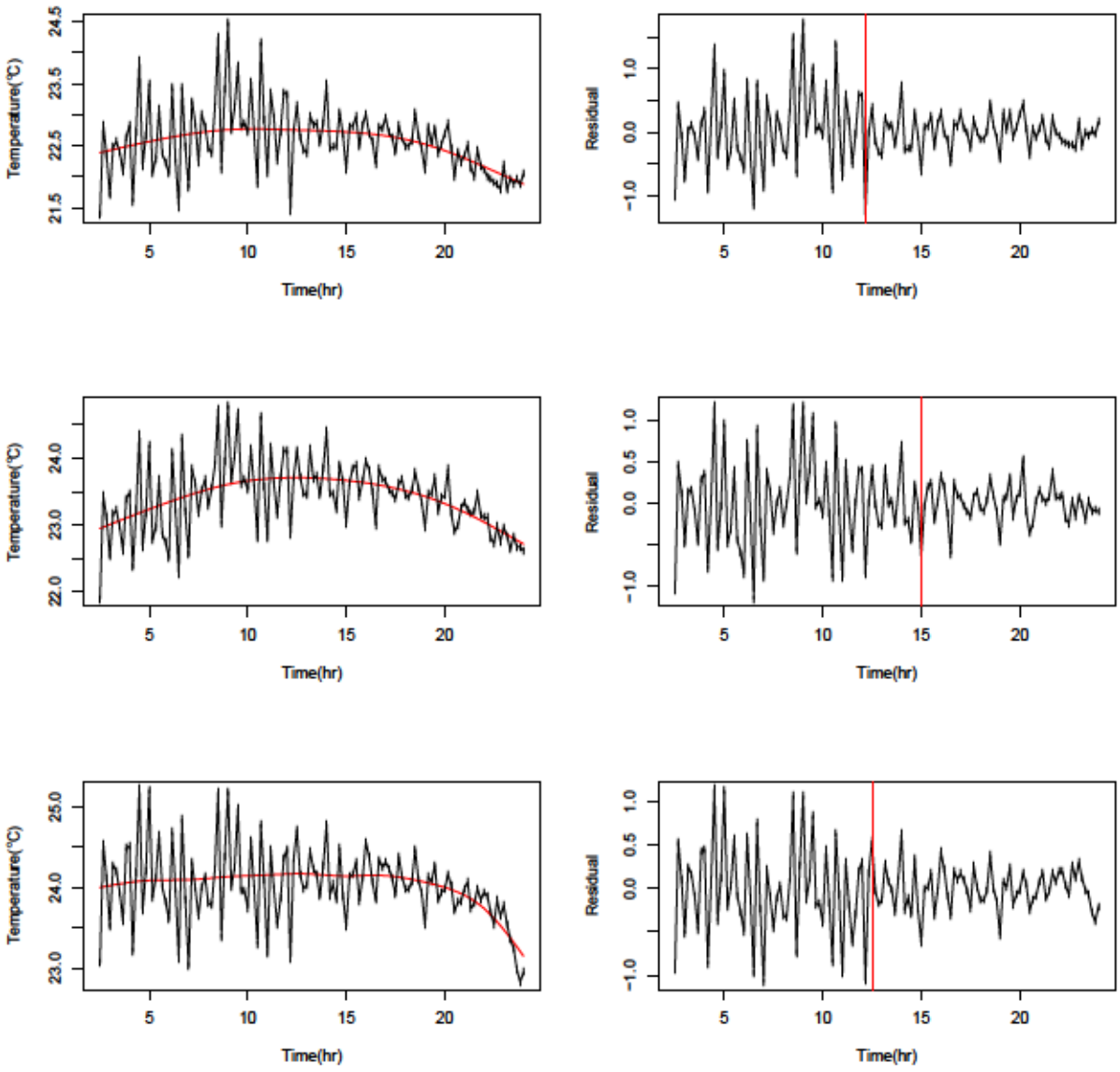

Figure 6: Mean and variance change point estimates imposed respectively on the raw and de-trended temperature profiles at three randomly selected spots.

cedure under our nonparametric smoothing setting is shown to have theoretical properties similar to that under a parameter model. The consistency result also has its own innovation in the perspective of nonparametric regression with non-IID errors. The application of our method to the liver procurement experiment provided critical information about the viability status of the liver lobe at different locations. A direction that merit further investigation is the development of an online version of our procedure. This can be derived with a combination of a proper characterization of in-control data.

SUPPLEMENTARY MATERIAL

The supplementary material collects all the conditions and technical proofs for the theoretical results in Section 2.4.

A.1 Conditions and Technical Lemmas

Conditions:

1. Suppose that when there is a variance change point the true change point $\tau_0 \in [cn/\log n, n - cn/\log n]$ for some $c > 0$. And assume that $\tau_0/n \to q_0 \in (0,1)$ as $n \to \infty$.

2. The true mean function $f_0 \in S^m$, the mth order Sobolev space of periodic functions on $[0,1]$ with period 1.

3. The random errors $\epsilon_i, i = 1, \ldots, n$ are independent normal random variables with mean 0 and variance σ_i^2, where $\sigma_i = \sigma_0$ when $i \leq \tau_0$ and $\sigma_i = \delta_0$ when $i > \tau_0$.

3'. The random errors $\epsilon_i, i = 1, \ldots, n$ are independent and identically distributed normal random variables with mean 0 and variance σ_0^2.

Condition 1 is common in change point analysis literature. It basically ensures that the change point is away from the boundaries. Condition 2 restricts our theory to the case when $\mathcal{H} = S^m$. A more general function space $\mathcal{H}$ is possible, but the matrix calculation involved in the technical proof would be much harder. Condition 3 spells our assumption about the error distribution and variances. The normality assumption in not necessary here. Any distribution with sub-Gaussian tails would be sufficient but the proof would be more tedious, though not necessarily harder. Condition 3' is the corresponding assumption about the error distribution under the null hypothesis that there is no variance change point.

We first show two technical lemmas that will be used in the proofs of our main theorems. Let $\delta_i = \widehat{f}^{(0)}(i/n) - E\{\widehat{f}^{(0)}\}(i/n)$ and $\delta_i^0 = E\{\widehat{f}^{(0)}\}(i/n) - f_0(i/n)$.

Lemma A.1. *There exists constant c_m (depending only on m) s.t.*

$$\|E\{\widehat{f}^{(0)}\} - f_0\|_{\sup} \leq c_m \sqrt{J(f_0)}(h^{m-1/2} + (nh)^{-1/2}), \tag{5}$$

Lemma A.2. *Suppose hypothesis H_1 holds true. Then it holds that*

$$\max_{1 \leq k_1 < k_2 \leq n} (k_2 - k_1)^{-1/2} \Big| \sum_{i=k_1+1}^{k_2} [\epsilon_i^2 - E(\epsilon_i^2)] \Big| = O_P(\log n), \tag{6}$$

$$\max_{1 \leq k \leq n} \frac{1}{\sqrt{k}} \Big| \sum_{i=1}^{k} \epsilon_i (\widehat{f}^{(0)}(i/n) - f_0(i/n)) \Big| = O_P(n^{-1/4} h^{-3/4}), \tag{7}$$

$$\max_{1 \leq k \leq n} \frac{1}{\sqrt{n-k}} \Big| \sum_{i=k+1}^{n} \epsilon_i (\widehat{f}^{(0)}(i/n) - f_0(i/n)) \Big| = O_P(n^{-1/4} h^{-3/4}), \tag{8}$$

$$\max_{1 \leq k \leq n} |\widehat{f}^{(0)}(i/n) - f_0(i/n)| = O_P(\sqrt{\log n/(nh)} + h^{m-1/2}). \tag{9}$$

The above results (7), (6), (8) also hold true under hypothesis H_0.

Proof of Lemma A.1. Since $\widehat{f}^{(0)}$ is the minimizer of (3), we have

$$-\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{f}^{(0)}(i/n))R_J(i/n,\cdot) + \lambda\widehat{f}^{(0)} = 0, \tag{10}$$

where R_J is the reproducing kernel associated with S^m (and J), and $R_J(x,\cdot)$ denotes the univariate function derived from R_J with its first argument fixed at x. Taking expectations, we get that

$$\frac{1}{n}\sum_{i=1}^{n}(\bar{f}(i/n) - f_0(i/n))R_J(i/n,\cdot) + \lambda\bar{f} = 0, \tag{11}$$

where $\bar{f} = E\{\widehat{f}^{(0)}\}$. Therefore, $\bar{f}$ is the minimizer to the following functional

$$\ell_0(f) = \frac{1}{n}\sum_{i=1}^{n}(f(i/n) - f_0(i/n))^2 + \lambda J(f).$$

Since $\ell_0(\bar{f}) \le \ell_0(f_0)$, we get

$$\frac{1}{n}\sum_{i=1}^{n}(\bar{f}(i/n) - f_0(i/n))^2 + \lambda J(\bar{f}) \le \lambda J(f_0).$$

This means that $J(\bar{f}) \le J(f_0)$. Let $g(t) = (\bar{f}(t) - f_0(t))^2$. Meanwhile, by Eggermont and LaRiccia (2009, Lemma (2.24), pp. 58) we get that

$$\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^{n}g(i/n) - \int_0^1 g(t)dt\right| &\le \frac{1}{n}\int_0^1 |g'(t)|dt \\
&= \frac{2}{n}\int_0^1 |\bar{f}'(t) - f_0'(t)| \times |\bar{f}(t) - f_0(t)|dt \\
&\le \frac{2}{n}\|\bar{f}' - f_0'\|_{L^2[0,1]}\|\bar{f} - f_0\|_{L^2[0,1]} \\
&\le \frac{2}{n}\|\bar{f}^{(m)} - f_0^{(m)}\|^2_{L^2[0,1]} \le \frac{8}{n}J(f_0).
\end{aligned} \tag{12}$$

In the meantime, (11) leads to

$$\frac{1}{n}\sum_{i=1}^{n}g(i/n) + \lambda J(\bar{f} - f_0) = -\lambda J(f_0, \bar{f} - f_0),$$

so that

$$\begin{aligned}\|\bar{f}-f_0\|^2 &\equiv \|\bar{f}-f_0\|_{L^2}^2+\lambda J(\bar{f}-f_0)\\ &= -\lambda J(f_0,\bar{f}-f_0)+\int_0^1 g(t)dt-\frac{1}{n}\sum_{i=1}^n g(i/n)\\ &\le 2J(f_0)\lambda+8J(f_0)/n=2J(f_0)(\lambda+4/n).\end{aligned}$$

It follows from Eggermont and LaRiccia (2009) that

$$\|\bar{f}-f_0\|_{\sup}\le c_m' h^{-1/2}\|\bar{f}-f_0\|\le c_m\sqrt{J(f_0)}(h^{m-1/2}+(nh)^{-1/2}),$$

where c_m, c_m' are positive constants depending only on m. Thus (5) holds. □

Proof of Lemma A.2. Let $\sigma_i^2=E(\epsilon_i^2)$. Without loss of generality, assume $\sigma^2=\sigma_1^2=\cdots=\sigma_{k_0}^2<\sigma_{k_0+1}=\cdots=\sigma_n^2=\delta^2$. Since $\epsilon_i^2-\sigma_i^2$ are independent centered sub-exponential random variables, by Vershynin (2012), there exist constants $c,d>0$ such that, for any $1\le k_2<k_2\le n$,

$$\begin{aligned}&P\left(|\sum_{i=k_1+1}^{k_2}[\epsilon_i^2-\sigma_i^2]|\ge C\sqrt{k_2-k_1}\log n\right)\\ \le\ & 2\exp\left(-c\min\left\{C^2(k_2-k_1)(\log n)^2/(d^2(k_2-k_1)), C\sqrt{k_2-k_1}\log n/d\right\}\right)\\ \le\ & 2\exp\left(-c\min\{C^2/d^2, C/d\}\log n\right)\le 2\exp(-3\log n)=2/n^3,\end{aligned}$$

where $C=\max\{\sqrt{3d^2/c}, 3d/c\}>0$. Hence, as $n\to\infty$,

$$P\left(\max_{1\le k_1<k_2\le n}(k_2-k_1)^{-1/2}|\sum_{i=k_1+1}^{k_2}[\epsilon_i^2-\sigma_i^2]|\ge C\log n\right)\le 2/n\to 0.$$

This shows (6).

Next we show (7). We only prove the results under H_1. The results under H_0 can be proved similarly. Define $\Omega=(\Omega_1^T,\ldots,\Omega_n^T)^T$ with $\Omega_i=(R_J(1/n,i/n),\ldots,R_J(n/n,i/n))/n$.

Then by the representer theorem (Wahba, 1990) it can be shown that $(\widehat{f}^{(0)}(1/n), \ldots, \widehat{f}^{(0)}(n/n))^T = \Omega(\Omega + \lambda I_n)^{-1}\mathbf{y}$.

From Wahba (1990) we know that

$$R_J(x,y) = \sum_{\nu=1}^{\infty} \frac{\varphi_\nu(x)\varphi_\nu(y)}{\gamma_\nu} = 2\sum_{k=1}^{\infty} \frac{\cos(2\pi k(x-y))}{(2\pi k)^{2m}}, \ x, y \in \mathbb{I}.$$

For $0 \le l \le n-1$, let $c_l = 2/n\sum_{k=1}^{\infty}\cos(2\pi kl/n)/(2\pi k)^{2m}$. Since $c_l = c_{n-l}$ for $l = 1, 2, \ldots, n-1$, Ω is symmetric circulant of order n.

Let $\zeta = \exp(2\pi\sqrt{-1}/n)$. The normalized eigenvectors of Ω can be specified as

$$\mathbf{x}_k = \frac{1}{\sqrt{n}}(1, \zeta^k, \zeta^{2k}, \ldots, \zeta^{(n-1)k})^T, \ k = 0, 1, \ldots, n-1.$$

Let $M = (\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{n-1})$. Denote M^* as the conjugate transpose of M. Clearly, $MM^* = I_n$ and Ω admits the decomposition $\Omega = M\Lambda M^*$, where $\Lambda = \mathrm{diag}(\lambda_0, \lambda_1, \ldots, \lambda_{n-1})$ with $\lambda_l = c_0 + c_1\zeta^l + \ldots + c_{n-1}\zeta^{(n-1)l}$.

Direct calculations show that

$$\lambda_l = \begin{cases} 2\sum_{k=1}^{\infty}\frac{1}{(2\pi kn)^{2m}}, & l = 0, \\ \sum_{k=1}^{\infty}\frac{1}{[2\pi(kn-l)]^{2m}} + \sum_{k=0}^{\infty}\frac{1}{[2\pi(kn+l)]^{2m}}, & 1 \le l \le n-1. \end{cases}$$

It is easy to examine that $\lambda_0 = 2\bar{c}_m(2\pi n)^{-2m}$ where $\bar{c}_m := \sum_{k=1}^{\infty}k^{-2m}$, and for $1 \le l \le n-1$,

$$\begin{aligned} \lambda_l &= \frac{1}{[2\pi(n-l)]^{2m}} + \frac{1}{(2\pi l)^{2m}} \\ &\quad + \sum_{k=2}^{\infty}\frac{1}{[2\pi(kn-l)]^{2m}} + \sum_{k=1}^{\infty}\frac{1}{[2\pi(kn+l)]^{2m}}. \end{aligned} \tag{13}$$

Let $\underline{c}_m = \sum_{k=2}^{\infty}k^{-2m}$. Then

$$\underline{c}_m(2\pi n)^{-2m} \le \sum_{k=2}^{\infty}\frac{1}{[2\pi(kn-l)]^{2m}} \le \bar{c}_m(2\pi n)^{-2m},$$

$$\underline{c}_m(2\pi n)^{-2m} \le \sum_{k=1}^{\infty}\frac{1}{[2\pi(kn+l)]^{2m}} \le \bar{c}_m(2\pi n)^{-2m}.$$

Let $\boldsymbol{e} \sim N(0, I_n)$ be a vector of independent standard normal random variables such that we can write $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T = D_0 \boldsymbol{e}$, where D_0 is the square-root of the true covariance matrix Σ_0 of $\boldsymbol{\epsilon}$, that is, D_0 is a diagonal matrix with the first τ_0 diagonals equal to σ_0 and the remaining diagonals equal to δ_0. Let $\boldsymbol{\delta}_{(k)} = (\delta_1, \ldots, \delta_k)^T$, $\Omega_{(k)} = (\Omega_1^T, \ldots, \Omega_k^T)^T$, $\boldsymbol{\epsilon}_{(k)} = (\epsilon_1, \ldots, \epsilon_k)^T$, $\boldsymbol{\epsilon}_{*(n-k)} = (\epsilon_{k+1}, \ldots, \epsilon_n)^T$, $\boldsymbol{e}_{(k)} = (e_1, \ldots, e_k)^T$, $\boldsymbol{e}_{*(n-k)} = (e_{k+1}, \ldots, e_n)^T$. Then

$$\begin{aligned}
\boldsymbol{\delta}_{(k)} &= \Omega_{(k)}(\Omega + \lambda I_n)^{-1}\boldsymbol{\epsilon} \\
&= (\mathbf{x}_0, \ldots, \mathbf{x}_{k-1})^T \Lambda M^\star M (\Lambda + \lambda I_n)^{-1} M^\star \boldsymbol{\epsilon} \\
&= (\mathbf{x}_0, \ldots, \mathbf{x}_{k-1})^T \Lambda (\Lambda + \lambda I_n)^{-1} M^\star \boldsymbol{\epsilon} \\
&= (\mathbf{x}_0, \ldots, \mathbf{x}_{k-1})^T \Lambda (\Lambda + \lambda I_n)^{-1} (\bar{\mathbf{x}}_0, \ldots, \bar{\mathbf{x}}_{k-1}) \boldsymbol{\epsilon}_{(k)} \\
&\quad + (\mathbf{x}_0, \ldots, \mathbf{x}_{k-1})^T \Lambda (\Lambda + \lambda I_n)^{-1} (\bar{\mathbf{x}}_k, \ldots, \bar{\mathbf{x}}_{n-1}) \boldsymbol{\epsilon}_{*(n-k)},
\end{aligned}$$

where $\bar{\mathbf{x}}_k$ is the conjugate of $\mathbf{x}_k$.

Define, for $k \leq \tau_0$, $D_k = \mathrm{diag}(\underbrace{\sigma_0, \ldots, \sigma_0}_{k \text{ items}})$; for $k > \tau_0$, $D_k = \mathrm{diag}(\underbrace{\sigma_0, \ldots, \sigma_0}_{\tau_0 \text{ items}}, \underbrace{\delta, \ldots, \delta}_{k-\tau_0 \text{ items}})$. Define, for $k \leq n - \tau_0$, $D_{*k} = \mathrm{diag}(\underbrace{\delta_0, \ldots, \delta_0}_{k \text{ items}})$; for $k > n - \tau_0$, $D_{*k} = \mathrm{diag}(\underbrace{\sigma_0, \ldots, \sigma_0}_{k-n+\tau_0}, \underbrace{\delta_0, \ldots, \delta_0}_{n-\tau_0})$. It is easy to see that $\boldsymbol{\epsilon}_{(k)} = D_k \boldsymbol{e}_{(k)}$ and $\boldsymbol{\epsilon}_{*(k)} = D_{*k} \boldsymbol{e}_{*(k)}$. Let $\widetilde{A}_k = (\mathbf{x}_0, \ldots, \mathbf{x}_{k-1})^T \Lambda (\Lambda + \lambda I_n)^{-1} (\bar{\mathbf{x}}_0, \ldots, \bar{\mathbf{x}}_{k-1})$ and $\widetilde{B}_k = (\mathbf{x}_0, \ldots, \mathbf{x}_{k-1})^T \Lambda (\Lambda + \lambda I_n)^{-1} (\bar{\mathbf{x}}_k, \ldots, \bar{\mathbf{x}}_{n-1})$. Define $A_k = D_k \widetilde{A}_k D_k$ and $B_k = D_k \widetilde{B}_k D_{*n-k}$. Then

$$\boldsymbol{\epsilon}_{(k)}^T \boldsymbol{\delta}_{(k)} = \boldsymbol{e}_{(k)}^T A_k \boldsymbol{e}_{(k)} + \boldsymbol{e}_{(k)}^T B_k \boldsymbol{e}_{*(n-k)}.$$

By the Hanson-Wright inequality, for any $k = 1, \ldots, n$,

$$P\left(|\boldsymbol{e}_{(k)}^T A_k \boldsymbol{e}_{(k)} - E\{\boldsymbol{e}_{(k)}^T A_k \boldsymbol{e}_{(k)}\}| \geq C_n \sqrt{k/(nh)} \right) \leq 2 \exp\left(- \min\left\{ \frac{c^2 C_n^2 k/(nh)}{\|A_k\|_F^2}, \frac{c C_n k/(nh)}{\|A_k\|_{op}} \right\} \right), \tag{14}$$

where $\|\cdot\|_F$ and $\|\cdot\|_{op}$ respectively denote the Frobenius norm and operator norm of a matrix, and $C_n > 0$ is a constant depending only n and $c > 0$ is a constant. Let $\Gamma = \Lambda(\Lambda + \lambda I_n)^{-1}$. Since Γ is a diagonal matrix, we can write $\Gamma = \mathrm{diag}(\gamma_0, \ldots, \gamma_{n-1})$. Let $M_{(k)} = (\mathbf{x}_0, \ldots, \mathbf{x}_{k-1})^T$. Since $M^\star M = I_n$, $M_{(k)}^\star M_{(k)} \leq I_n$. Let $a_0 = \max\{\sigma^2, \delta^2\}$. We know that

$$\begin{aligned}
\|A_k\|_F^2 &\leq a_0^2 \mathrm{Tr}(\widetilde{A}_k^\star \widetilde{A}_k) = a_0^2 \mathrm{Tr}(M_{(k)}\Gamma M_{(k)}^\star M_{(k)}\Gamma M_{(k)}^\star) \\
&\leq a_0^2 \mathrm{Tr}(M_{(k)}\Gamma^2 M_{(k)}^\star) = a_0^2 \sum_{l=0}^{k-1} \mathbf{x}_l^T \mathrm{diag}(\gamma_0^2, \ldots, \gamma_{n-1}^2)\bar{\mathbf{x}}_l \\
&= \frac{a_0^2 k}{n} \sum_{r=0}^{n-1} \gamma_r^2 = O(\frac{k}{nh}), \text{uniformly for } k.
\end{aligned}$$

This also shows that $\|A_k\|_{op} \leq \|A_k\|_F = O(\sqrt{k/(nh)})$ uniformly for k. So for $C_n > 1$, (14) becomes

$$P\left(|\boldsymbol{e}_{(k)}^T A_k \boldsymbol{e}_{(k)} - E\{\boldsymbol{e}_{(k)}^T A_k \boldsymbol{e}_{(k)}\}| \geq C_n\sqrt{k/(nh)}\right) \leq 2\exp(-cC_n).$$

This shows that

$$P\left(\max_{1\leq k\leq n} \frac{|\boldsymbol{e}_{(k)}^T A_k \boldsymbol{e}_{(k)} - E\{\boldsymbol{e}_{(k)}^T A_k \boldsymbol{e}_{(k)}\}|}{\sqrt{k/(nh)}} \geq C_n\right) \leq 2n\exp(-cC_n).$$

Taking $C_n = (2/c)\log n$, we have shown that

$$\max_{1\leq k\leq n} \frac{|\boldsymbol{e}_{(k)}^T A_k \boldsymbol{e}_{(k)} - E\{\boldsymbol{e}_{(k)}^T A_k \boldsymbol{e}_{(k)}\}|}{\sqrt{k/(nh)}} = O_P(\log n).$$

In the meantime,

$$P\left(\max_{1\leq k\leq n} \frac{|\sum_{i=1}^k \delta_i^0 \epsilon_i|}{\sqrt{\sum_{i=1}^k (\delta_i^0\sigma_i)^2}} \geq C\right) \leq nP(|Z| > C) = O(n\exp(-C^2/2)),$$

which implies

$$\max_{1\le k\le n}\frac{|\sum_{i=1}^k \delta_i^0\epsilon_i|}{\sqrt{\sum_{i=1}^k(\delta_i^0\sigma_i)^2}} = O_P(\sqrt{\log n}).$$

So by Lemma A.1 and conditions $h(\log n)^2 = o(1)$ and $nh^{4m+1}(\log n)^2 = o(1)$, we have

$$\begin{aligned}
&\frac{1}{\sqrt{k}}\sum_{i=1}^k \epsilon_i(\delta_i+\delta_i^0)\\
=\;& \frac{1}{\sqrt{k}}\boldsymbol{\epsilon}_{(k)}^T\boldsymbol{\delta}_{(k)} + \frac{1}{\sqrt{k}}\sum_{i=1}^k\epsilon_i\delta_i^0\\
=\;& \frac{1}{\sqrt{k}}\left(\boldsymbol{e}_{(k)}^T A_k \boldsymbol{e}_{(k)} + \boldsymbol{e}_{(k)}^T B_k \boldsymbol{e}_{*(n-k)}\right) + \frac{1}{\sqrt{k}}\sum_{i=1}^k\epsilon_i\delta_i^0\\
=\;& O_P\left(\frac{\sqrt{k}}{nh} + \frac{\log n}{\sqrt{nh}} + \sqrt{(\frac{1}{nh}+\sqrt{\frac{1}{knh}})(\frac{n-k}{nh}+\sqrt{\frac{n-k}{nh}})} + \sqrt{\log n}(h^{m-1/2}+(nh)^{-1/2})\right)\\
=\;& O_P\left(\frac{1}{\sqrt{n}h} + \frac{\log n}{\sqrt{nh}} + \sqrt{(\frac{1}{nh}+\sqrt{\frac{1}{nh}})(\frac{1}{h}+\sqrt{\frac{1}{h}})} + \sqrt{\log n}(h^{m-1/2}+(nh)^{-1/2})\right)\\
=\;& O_P\left(\frac{1}{\sqrt{n}h} + \frac{\log n}{\sqrt{nh}} + \left(\frac{1}{nh^3}\right)^{1/4} + \sqrt{\log n}(h^{m-1/2}+(nh)^{-1/2})\right)\\
=\;& O_P(n^{-1/4}h^{-3/4}),
\end{aligned}$$

where the O_P term is uniformly valid for $1 \le k \le n$.

Next we will handle $\|\widehat{f}^{(0)} - \bar{f}\|_{\sup}$. It can be seen by the representer theorem that $\widehat{f}^{(0)} - \bar{f} = (R_J(1/n,\cdot),\ldots,R_J(n/n,\cdot))(\Omega+\lambda I_n)^{-1}\epsilon/n$. It is easy to see that, with Ω_i being the ith row of Ω, $\delta_i = \Omega_i(\Omega+\lambda I_n)^{-1}\epsilon \sim N(0, \Omega_i(\Omega+\lambda I_n)^{-2}\Omega_i^T\sigma_i^2)$. Note that

$$\Omega_i(\Omega+\lambda I_n)^{-2}\Omega_i^T = \mathbf{x}_{i-1}^T\Lambda(\Lambda+\lambda I_n)^{-2}\Lambda\bar{\mathbf{x}}_{i-1} = \frac{1}{n}\sum_{r=0}^{n-1}\frac{\lambda_r^2}{(\lambda+\lambda_r)^2} \asymp \frac{1}{nh}.$$

Therefore, as $n \to \infty$,

$$P\left(\max_{1\le k\le n}|\delta_i| \ge C\sqrt{\log n}/\sqrt{nh}\right) \le nP(|\delta_i| \ge C\sqrt{\log n}/\sqrt{nh}) \le n\exp(-C\log n) \to 0.$$

This shows that $\max_{1\le k\le n}|\delta_i| = O_P(\sqrt{\log n/(nh)})$. The result follows from Lemma A.1.

□

A.2 Proof of Theorem 2.1

The consistency result in Theorem 2.1 is proved in three steps: (1) the consistency of the initial mean function estimate, (2) the consistency of variance change point estimate and variance estimates given a consistent mean function estimate, and (3) the consistency of the mean estimate given consistent variance change point estimate and variance estimates. Particularly, we shall prove the following results.

1. $\max_{1\le i\le n}|\widehat{f}^{(0)}(i/n) - f_0(i/n)| = O_P(r_n)$.
2. Given that $\max_{1\le i\le n}|\widehat{f}^{(\iota-1)}(i/n) - f_0(i/n)| = O_P(r_n)$, we have $|\widehat{\tau}^{(\iota)} - \tau_0| = O_P((\log n)^4(\log\log n)^2)$, $[\widehat{\sigma}^2]^{(\iota)} = \sigma_0^2 + O_P(\widetilde{r}_n)$ and $[\widehat{\delta}^2]^{(\iota)} = \delta_0^2 + O_P(\widetilde{r}_n)$.
3. Given that $|\widehat{\tau}^{(\iota)} - \tau_0| = O_P((\log n)^4(\log\log n)^2)$, $[\widehat{\sigma}^2]^{(\iota)} = \sigma_0^2 + O_P(\widetilde{r}_n)$ and $[\widehat{\delta}^2]^{(\iota)} = \delta_0^2 + O_P(\widetilde{r}_n)$, we have $\max_{1\le i\le n}|\widehat{f}^{(\iota)}(i/n) - f_0(i/n)| = O_P(r_n)$.

These results, combined together, immediately guarantees the consistency result in Theorem 2.1. For simplicity of notation, we shall drop the superscripts $(\iota-1)$ and (ι) in this section of proof.

STEP 1. Consistency of the initial mean function estimate $\widehat{f}^{(0)}$.

This follows directly from Lemma A.2.

STEP 2. Consistency of the estimates of τ_0, σ_0^2 and δ_0^2 given a consistent mean estimate.

Without loss of generality assume $\sigma_0^2 < \delta_0^2$. The idea is to show that, $\ell(k) > \ell(\tau_0)$ uniformly for $k \in [cn/\log n, n - cn/\log n]$ with $|k - \tau_0| \geq (\log n)^4(\log\log n)^2$. We only consider $cn/\log n \leq k < \tau_0 - (\log n)^4(\log\log n)^2$ since the case for $n - cn/\log n \geq k > \tau_0 + (\log n)^4(\log\log n)^2$ is similar. Define $\eta_i = \widehat{f}(i/n) - f_0(i/n)$ and let

$$\widehat{\sigma}_0^2 = \frac{1}{\tau_0}\sum_{i=1}^{\tau_0}(\eta_i + \epsilon_i)^2,\ \widehat{\sigma}_n^2 = \frac{1}{n-\tau_0}\sum_{i=\tau_0+1}^{n}(\eta_i + \epsilon_i)^2,\ \widehat{\sigma}_k^2 = \frac{1}{\tau_0 - k}\sum_{i=k+1}^{\tau_0}(\eta_i + \epsilon_i)^2.$$

It follows by Lemmas A.1 and A.2 that $\frac{1}{\tau_0-k}\sum_{i=k+1}^{\tau_0}\eta_i\epsilon_i = O_P(r_n\sqrt{\log n})$ uniformly for $k \leq \tau_0 - (\log n)^4(\log\log n)^2$. Hence we have

$$\begin{aligned}
\widehat{\sigma}_k^2 &= \frac{1}{\tau_0 - k}\sum_{i=k+1}^{\tau_0}\eta_i^2 + \frac{2}{\tau_0 - k}\sum_{i=k+1}^{\tau_0}\eta_i\epsilon_i + \frac{1}{\tau_0 - k}\sum_{i=k+1}^{\tau_0}\epsilon_i^2 \\
&= \sigma^2 + O_P\left(r_n^2 + r_n\sqrt{\log n} + (\log n \log\log n)^{-1}\right) \\
&= \sigma^2 + O_P(r_{1n}^2),
\end{aligned}$$

where $r_{1n}^2 = r_n^2 + r_n\sqrt{\log n} + (\log n \log\log n)^{-1}$ which is $o(1)$ by assumptions. Meanwhile, using similar argument we have $\widehat{\sigma}_0^2 = \sigma^2 + O_P(r_{1n}^2)$, $\widehat{\sigma}_n^2 = \delta^2 + O_P(r_{1n}^2)$.

Therefore, with probability approaching one, uniformly for $k \leq \tau_0 - (\log n)^4(\log\log n)^2$,

$$|\widehat{\sigma}_k^2/\widehat{\sigma}_n^2 - \sigma^2/\delta^2| = O_P(r_{1n}^2),\ \ |\widehat{\sigma}_0^2/\widehat{\sigma}_n^2 - \sigma^2/\delta^2| = O_P(r_{1n}^2),$$

$$\left|1 - \frac{\widehat{\sigma}_k^2}{\widehat{\sigma}_0^2}\right| = \left|\frac{\widehat{\sigma}_0^2 - \widehat{\sigma}_k^2}{\widehat{\sigma}_0^2}\right| = O_P(r_{1n}^2).$$

It is easy to see that

$$\frac{1}{k}\sum_{i=1}^{k}(y_i - \widehat{f}(i/n))^2 = \frac{1}{k}\sum_{i=1}^{k}(\delta_i + \epsilon_i)^2 = \widehat{\sigma}_0^2 + \frac{\tau_0 - k}{k}(\widehat{\sigma}_0^2 - \widehat{\sigma}_k^2),$$

$$\frac{1}{n-k}\sum_{i=k+1}^{n}(y_i - \widehat{f}(i/n))^2 = \frac{1}{n-k}\sum_{i=k+1}^{n}(\delta_i + \epsilon_i)^2 = \widehat{\sigma}_n^2 + \frac{\tau_0 - k}{n-k}(\widehat{\sigma}_k^2 - \widehat{\sigma}_n^2).$$

Therefore,

$$\begin{aligned}
& \ell(k) \\
= \; & \frac{k}{n}\log\left(\widehat{\sigma}_0^2+\frac{\tau_0-k}{k}(\widehat{\sigma}_0^2-\widehat{\sigma}_k^2)\right)+\frac{n-k}{n}\log\left(\widehat{\sigma}_n^2+\frac{\tau_0-k}{n-k}(\widehat{\sigma}_k^2-\widehat{\sigma}_n^2)\right) \\
= \; & \ell(\tau_0)-\frac{\tau_0-k}{n}\log\left(\frac{\widehat{\sigma}_0^2}{\widehat{\sigma}_n^2}\right)+\frac{n-k}{n}\log\left(1+\frac{\tau_0-k}{n-k}\left(\frac{\widehat{\sigma}_k^2}{\widehat{\sigma}_n^2}-1\right)\right)+\frac{k}{n}\log\left(1+\frac{\tau_0-k}{k}\left(1-\frac{\widehat{\sigma}_k^2}{\widehat{\sigma}_0^2}\right)\right).
\end{aligned}$$

Note that

$$\begin{aligned}
& \frac{n-k}{n}\log\left(1+\frac{\tau_0-k}{n-k}\left(\frac{\widehat{\sigma}_k^2}{\widehat{\sigma}_n^2}-1\right)\right) \\
= \; & \frac{n-k}{n}\log\left(1+\frac{\tau_0-k}{n-k}\left(\frac{\widehat{\sigma}_0^2}{\widehat{\sigma}_n^2}-1+\frac{\widehat{\sigma}_k^2-\widehat{\sigma}_0^2}{\widehat{\sigma}_n^2}\right)\right) \\
= \; & \frac{n-k}{n}\log\left(1+\frac{\tau_0-k}{n-k}\left(\frac{\widehat{\sigma}_0^2}{\widehat{\sigma}_n^2}-1\right)\right)+\frac{\tau_0-k}{n}O_P(r_{1n}^2).
\end{aligned}$$

Therefore, with probability approaching one, uniformly for $k<\tau_0-(\log n)^4(\log\log n)^2$, we have

$$\begin{aligned}
& \ell(k)-\ell(\tau_0) \\
= \; & -\frac{\tau_0-k}{n}\log\left(\frac{\widehat{\sigma}_0^2}{\widehat{\sigma}_n^2}\right)+\frac{n-k}{n}\log\left(1+\frac{\tau_0-k}{n-k}\left(\frac{\widehat{\sigma}_k^2}{\widehat{\sigma}_n^2}-1\right)\right)+\frac{k}{n}\log\left(1+\frac{\tau_0-k}{k}\left(1-\frac{\widehat{\sigma}_k^2}{\widehat{\sigma}_0^2}\right)\right) \\
= \; & -\frac{\tau_0-k}{n}\log\left(\frac{\widehat{\sigma}_0^2}{\widehat{\sigma}_n^2}\right)+\frac{n-k}{n}\log\left(1+\frac{\tau_0-k}{n-k}\left(\frac{\widehat{\sigma}_0^2}{\widehat{\sigma}_n^2}-1\right)\right)+\frac{\tau_0-k}{n}O_P(r_{1n}^2) \\
= \; & \frac{\tau_0-k}{n}\sum_{l=1}^{\infty}\frac{1}{l}\left(1-\frac{\widehat{\sigma}_0^2}{\widehat{\sigma}_n^2}\right)^l\left[1-\left(\frac{\tau_0-k}{n-k}\right)^{l-1}\right]+\frac{\tau_0-k}{n}O_P(r_{1n}^2) \\
\geq \; & \frac{\tau_0-k}{2n}\left(1-\frac{\widehat{\sigma}_0^2}{\widehat{\sigma}_n^2}\right)^2\frac{n-\tau_0}{n-k}+\frac{\tau_0-k}{n}O_P(r_{1n}^2) \\
\geq \; & \frac{\tau_0-k}{2n}\left[(1-\sigma^2/\delta^2)^2(1-q_0)+O_P(r_{1n}^2)\right]>0,
\end{aligned}$$

where the last inequality follows by $r_{1n}^2=o(1)$. This means that $\widehat{\tau}\geq\tau_0-(\log n)^4(\log\log n)^2$ with probability approaching one. Similarly, it can be shown that with probability approaching one, $\ell(k)-\ell(\tau_0)>0$ uniformly for $k>\tau_0+(\log n)^4(\log\log n)^2$, which implies

$\widehat{\tau} \le \tau_0 + (\log n)^4(\log\log n)^2$. Therefore,

$$|\widehat{\tau} - \tau_0| = O_P((\log n)^4(\log\log n)^2). \tag{15}$$

To show the consistency of $\widehat{\sigma}^2$ and $\widehat{\delta}^2$, note that by Lemma A.2 we have that

$$\begin{aligned}
\widehat{\sigma}^2 &= \frac{1}{\widehat{\tau}}\sum_{i=1}^{\widehat{\tau}} \eta_i^2 + \frac{2}{\widehat{\tau}}\sum_{i=1}^{\widehat{\tau}} \epsilon_i\eta_i + \frac{1}{\widehat{\tau}}\sum_{i=1}^{\widehat{\tau}} \epsilon_i^2 \\
&= O_P\left(\frac{\log n}{nh} + h^{2m-1} + (nh)^{-3/4}\right) + \frac{1}{\widehat{\tau}}\sum_{i=1}^{\widehat{\tau}} \epsilon_i^2 \\
&= O_P\left(r_n^2 + (nh)^{-3/4}\right) + \frac{1}{\widehat{\tau}}\left(\sum_{i=1}^{\widehat{\tau}} \epsilon_i^2 - \sum_{i=1}^{\tau_0} \epsilon_i^2\right) + \frac{1}{\widehat{\tau}}\sum_{i=1}^{\tau_0} \epsilon_i^2.
\end{aligned}$$

By Lemma A.2 and (15), we have

$$\begin{aligned}
\frac{1}{\widehat{\tau}}\left|\sum_{i=1}^{\widehat{\tau}} \epsilon_i^2 - \sum_{i=1}^{\tau_0} \epsilon_i^2\right| &\le \frac{1}{\widehat{\tau}} \sum_{i=\tau_0+1-|\widehat{\tau}-\tau_0|}^{\tau_0+1+|\widehat{\tau}-\tau_0|} \epsilon_i^2 \\
&= \frac{2|\widehat{\tau}-\tau_0|+1}{\widehat{\tau}} \times \frac{1}{2|\widehat{\tau}-\tau_0|+1} \sum_{i=\tau_0+1-|\widehat{\tau}-\tau_0|}^{\tau_0+1+|\widehat{\tau}-\tau_0|} \epsilon_i^2 \\
&= \frac{2|\widehat{\tau}-\tau_0|+1}{\widehat{\tau}} \times O_P\left(1 + \frac{\sqrt{|\widehat{\tau}-\tau_0|}\log n}{|\widehat{\tau}-\tau_0|+1}\right) \\
&= O_P((\log n)^5(\log\log n)^2/n), \\
\text{and} \quad \frac{1}{\widehat{\tau}}\sum_{i=1}^{\tau_0} \epsilon_i^2 &= \frac{\tau_0}{\widehat{\tau}}\frac{1}{\tau_0}\sum_{i=1}^{\tau_0} \epsilon_i^2 = \sigma_0^2 + O_P(n^{-1/2}).
\end{aligned}$$

Therefore, we have proved that

$$\widehat{\sigma}^2 = \sigma_0^2 + O_P(r_n^2 + (nh)^{-3/4} + (\log n)^5(\log\log n)^2/n + n^{-1/2}) = \sigma_0^2 + O_P(\widetilde{r}_n).$$

The proof for $\widehat{\delta}^2 = \delta_0^2 + O_P(\widetilde{r}_n)$ is similar.

STEP 3. Consistency of the mean function estimate given the consistent estimates of τ_0, σ_0^2 and δ_0^2.

Recall from the proof of Lemma A.2, we write $\boldsymbol{\epsilon} = D_0\mathbf{e}$, where D_0 is the square-root of the true covariance matrix of $\boldsymbol{\epsilon}$ and $\mathbf{e}$ is a vector of n independent standard normal random variables. By the representer theorem, the estimates $\widehat{f}^{(0)}$ and $\widehat{f}$ have explicit expressions

$$\widehat{f}^{(0)} = \Omega(\Omega + \lambda I)^{-1} f_0 + \Omega(\Omega + \lambda I)^{-1} D_0\mathbf{e},$$

$$\widehat{f} = \Omega(\Omega + \lambda_{new}\widehat{\Sigma})^{-1} f_0 + \Omega(\Omega + \lambda_{new}\widehat{\Sigma})^{-1} D_0\mathbf{e}.$$

Without loss of generality, assume $\sigma^2 < \delta^2$. Let $\widehat{c} = \widehat{\delta}^2/\widehat{\sigma}^2$ and $c_0 = \delta^2/\sigma^2$. By consistency of $\widehat{\sigma}^2$, $\widehat{\delta}^2$, and $\widehat{\tau}$, with probability approaching one, $\mathcal{E}_n \equiv \{\widehat{\sigma}^2 < \widehat{\delta}^2, |\widehat{c} - c_0| \leq C\widetilde{r}_n\}$ holds, where $\varepsilon > 0$ is arbitrarily small. Let 1_k be a vector of k 1's and 0_k be a vector of k 0's. It is easy to see that on $\mathcal{E}_n$,

$$\begin{aligned}
\lambda_{new}\widehat{\Sigma} &= \lambda \mathrm{diag}(\widehat{c}1_{\widehat{\tau}}, 1_{n-\widehat{\tau}}) \\
&= \lambda \mathrm{diag}(c_0 1_{\widehat{\tau}}, 1_{n-\widehat{\tau}}) + \lambda(\widehat{c} - c_0)\mathrm{diag}(1_{\widehat{\tau}}, 0_{n-\widehat{\tau}}) \\
&\equiv \lambda\Gamma + \lambda(\widehat{c} - c_0)E,
\end{aligned}$$

where $\Gamma = \mathrm{diag}(c_0 1_{\widehat{\tau}}, 1_{n-\widehat{\tau}})$ and $E = \mathrm{diag}(1_{\widehat{\tau}}, 0_{n-\widehat{\tau}})$. Then by the Sherman-Woordbury formula,

$$\begin{aligned}
&(\Omega + \lambda_{new}\widehat{\Sigma})^{-1} - (\Omega + \lambda\Gamma)^{-1} \\
= \; & -(\widehat{c} - c_0)\lambda(\Omega + \lambda\Gamma)^{-1}E(I + (\widehat{c} - c_0)\lambda E(\Omega + \lambda\Gamma)^{-1}E)^{-1}E(\Omega + \lambda\Gamma)^{-1} \equiv -(\widehat{c} - c_0)\Delta,
\end{aligned}$$

where $\Delta = \lambda(\Omega + \lambda\Gamma)^{-1}E(I + (\widehat{c} - c_0)\lambda E(\Omega + \lambda\Gamma)^{-1}E)^{-1}E(\Omega + \lambda\Gamma)^{-1}$. Hence, on $\mathcal{E}_n$, $0 \leq \Delta \leq \frac{\lambda}{1-\varepsilon}(\Omega + \lambda\Gamma)^{-2} \leq \frac{1}{1-\varepsilon}(\Omega + \lambda I)^{-1}$.

Notice that $\lambda I \le \lambda_{new}\widehat{\Sigma} \le \lambda\widehat{c}I$. Hence, it holds that

$$\begin{aligned}
&(\widehat{f}-f_0)'(\widehat{f}-f_0)\\
\le\ & 2f_0'(\Omega+\lambda_{new}\widehat{\Sigma})^{-1}(\lambda_{new}\widehat{\Sigma})^2(\Omega+\lambda_{new}\widehat{\Sigma})^{-1}f_0 + 2\mathbf{e}^T D_0(\Omega+\lambda_{new}\widehat{\Sigma})^{-1}\Omega^2(\Omega+\lambda_{new}\widehat{\Sigma})^{-1}D_0\mathbf{e}\\
\le\ & 2\widehat{c}^2\lambda^2 f_0'(\Omega+\lambda I)^{-2}f_0 + 4\mathbf{e}^T D_0(\Omega+\lambda\Gamma)^{-1}\Omega^2(\Omega+\lambda\Gamma)^{-1}D_0\mathbf{e} + 4(\widehat{c}-c_0)^2\mathbf{e}^T D_0\Delta\Omega^2\Delta D_0\mathbf{e}.
\end{aligned}$$

We will handle the three terms respectively. The first term is bounded by $2\widehat{c}^2(E\{\widehat{f}^{(0)}\} - f_0)'(E\{\widehat{f}^{(0)}\} - f_0) = O_P(n\lambda)$ by Wahba (1990). To handle the second term, note that

$$\begin{aligned}
E\{\mathbf{e}^T D_0(\Omega+\lambda\Gamma)^{-1}\Omega^2(\Omega+\lambda\Gamma)^{-1}D_0\mathbf{e}\} &= \mathrm{Tr}(D_0(\Omega+\lambda\Gamma)^{-1}\Omega^2(\Omega+\lambda\Gamma)^{-1}D_0)\\
&= \mathrm{Tr}(\Omega(\Omega+\lambda\Gamma)^{-1}D_0^2(\Omega+\lambda\Gamma)^{-1}\Omega)\\
&\le \delta^4\mathrm{Tr}(\Omega(\Omega+\lambda\Gamma)^{-2}\Omega)\\
&\le \delta^4\mathrm{Tr}(\Omega(\Omega+\lambda I)^{-2}\Omega) = O(h^{-1}),
\end{aligned}$$

so the second term is $O_P(h^{-1})$. As for the third term, notice that

$$\begin{aligned}
\mathbf{e}^T D_0\Delta\Omega^2\Delta D_0\mathbf{e} &\le \mathbf{e}^T\mathbf{e}\mathrm{Tr}(D_0\Delta\Omega^2\Delta D_0)\\
&\le \delta^4\mathbf{e}^T\mathbf{e}\mathrm{Tr}(\Omega\Delta^2\Omega)\\
&\le \delta^4\mathbf{e}^T\mathbf{e}\times\frac{1}{(1-\varepsilon)^2}\mathrm{Tr}(\Omega(\Omega+\lambda I)^{-2}\Omega) = O_P(nh^{-1}),
\end{aligned}$$

hence the last term is $O_P(nh^{-1}\widetilde{r}_n^2)$. Therefore, $\|\widehat{f}-f_0\|_n^2 = O_P(\lambda + (nh)^{-1} + h^{-1}\widetilde{r}_n^2)$.

A.3 Proof of Theorem 2.2

Under H_0, the samples Y_i come from conventional nonparametric model with Gaussian errors of equal variance. Without loss of generality, assume that the variance of ϵ_i is one. Note that the mean function estimate under H_0 is $\widehat{f}^{(0)}$.

Recall that $\delta_i = \widehat{f}^{(0)}(i/n) - E\{\widehat{f}^{(0)}\}(i/n)$ and $\delta_i^0 = E\{\widehat{f}^{(0)}\}(i/n) - f_0(i/n)$. Let $\eta_i = \delta_i + \delta_i^0 = \widehat{f}^{(0)}(i/n) - f_0(i/n)$. For any $1 < k < n$, we have by Taylor's expansion and results from Lemmas A.1 and A.2 that

$$\begin{aligned}
\ell(k) - \ell(n) &= n\log\left(1 + \frac{\sum_{i=1}^n(\epsilon_i^2 - 1) + 2\sum_{i=1}^n \eta_i\epsilon_i + \sum_{i=1}^n \eta_i^2}{n}\right) \\
&\quad -k\log\left(1 + \frac{\sum_{i=1}^k(\epsilon_i^2 - 1) + 2\sum_{i=1}^k \eta_i\epsilon_i + \sum_{i=1}^k \eta_i^2}{k}\right) \\
&\quad -(n-k)\log\left(1 + \frac{\sum_{i=k+1}^n(\epsilon_i^2 - 1) + 2\sum_{i=k+1}^n \eta_i\epsilon_i + \sum_{i=k+1}^n \eta_i^2}{n-k}\right) \\
&= -\frac{1}{2n}\left(\sum_{i=1}^n(\epsilon_i^2 - 1) + 2\sum_{i=1}^n \eta_i\epsilon_i + \sum_{i=1}^n \eta_i^2\right)^2 \\
&\quad +\frac{1}{2k}\left(\sum_{i=1}^k(\epsilon_i^2 - 1) + 2\sum_{i=1}^k \eta_i\epsilon_i + \sum_{i=1}^k \eta_i^2\right)^2 \\
&\quad +\frac{1}{2(n-k)}\left(\sum_{i=k+1}^n(\epsilon_i^2 - 1) + 2\sum_{i=k+1}^n \eta_i\epsilon_i + \sum_{i=k+1}^n \eta_i^2\right)^2 \\
&\quad +O_P(n[\frac{\sum_{i=1}^n(\epsilon_i^2 - 1) + 2\sum_{i=1}^n \eta_i\epsilon_i + \sum_{i=1}^n \eta_i^2}{n}]^3) \\
&\quad +O_P(k[\frac{\sum_{i=1}^k(\epsilon_i^2 - 1) + 2\sum_{i=1}^k \eta_i\epsilon_i + \sum_{i=1}^k \eta_i^2}{k}]^3) \\
&\quad +O_P((n-k)[\frac{\sum_{i=k+1}^n(\epsilon_i^2 - 1) + 2\sum_{i=k+1}^n \eta_i\epsilon_i + \sum_{i=k+1}^n \eta_i^2}{n-k}]^3)
\end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{2n}[\sum_{i=1}^{n}(\epsilon_i^2-1)]^2+\frac{1}{2k}[\sum_{i=1}^{k}(\epsilon_i^2-1)]+\frac{1}{2(n-k)}[\sum_{i=k+1}^{n}(\epsilon_i^2-1)]^2 \\
&\quad -\frac{1}{n}\sum_{i=1}^{n}(\epsilon_i^2-1)[2\sum_{i=1}^{n}\eta_i\epsilon_i+\sum_{i=1}^{n}\eta_i^2]+\frac{1}{k}\sum_{i=1}^{k}(\epsilon_i^2-1)[2\sum_{i=1}^{k}\eta_i\epsilon_i+\sum_{i=1}^{k}\eta_i^2] \\
&\quad +\frac{1}{n-k}\sum_{i=k+1}^{n}(\epsilon_i^2-1)[2\sum_{i=k+1}^{n}\eta_i\epsilon_i+\sum_{i=k+1}^{n}\eta_i^2] \\
&\quad -\frac{1}{2n}[2\sum_{i=1}^{n}\eta_i\epsilon_i+\sum_{i=1}^{n}\eta_i^2]^2+\frac{1}{2k}[2\sum_{i=1}^{k}\eta_i\epsilon_i+\sum_{i=1}^{k}\eta_i^2]^2+\frac{1}{2(n-k)}[2\sum_{i=k+1}^{n}\eta_i\epsilon_i+\sum_{i=k+1}^{n}\eta_i^2]^2 \\
&\quad +O_P(n[\frac{\sum_{i=1}^{n}(\epsilon_i^2-1)+2\sum_{i=1}^{n}\eta_i\epsilon_i+\sum_{i=1}^{n}\eta_i^2}{n}]^3) \\
&\quad +O_P(k[\frac{\sum_{i=1}^{k}(\epsilon_i^2-1)+2\sum_{i=1}^{k}\eta_i\epsilon_i+\sum_{i=1}^{k}\eta_i^2}{k}]^3) \\
&\quad +O_P((n-k)[\frac{\sum_{i=k+1}^{n}(\epsilon_i^2-1)+2\sum_{i=k+1}^{n}\eta_i\epsilon_i+\sum_{i=k+1}^{n}\eta_i^2}{n-k}]^3) \\
&= -n\log\left(\frac{\sum_{i=1}^{n}\epsilon_i^2}{n}\right)+k\log\left(\frac{\sum_{i=1}^{k}\epsilon_i^2}{k}\right)+(n-k)\log\left(\frac{\sum_{i=k+1}^{n}\epsilon_i^2}{n-k}\right) \\
&\quad +O_P(\log n(n^{-1/4}h^{-3/4}+n^{-1/2}h^{-1}\log n+n^{1/2}h^{2m-1})) \\
&= \mathrm{SIC}(k)-\mathrm{SIC}(n)+O_P(r_n),
\end{aligned}
$$

where the O_P term holds uniformly for k and $r_n = \log n(n^{-1/4}h^{-3/4} + n^{-1/2}h^{-1}\log n + n^{1/2}h^{2m-1})$. It then follows

$$\max_{1\le k\le n}[\ell(k)-\ell(n)] = \max_{1\le k\le n}[\mathrm{SIC}(k)-\mathrm{SIC}(n)]+O_P(r_n).$$

By Chen and Gupta (1997) we have for any $x \in \mathbb{R}$,

$$P(a_n(\log n)^{1/2}\lambda_{\star n}-b_n\log n\le x)\to\exp(-2\exp(-x)).$$

Since r_n satisfies $r_n\log^2 n = o(1)$, we have $a_n(\log n)(\lambda_n-\lambda_{\star n}) = o_P(1)$. Therefore,

$$P(a_n(\log n)^{1/2}\lambda_n-b_n\log n\le x)\to\exp(-2\exp(-x)).$$

References

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B 57*, 289–300.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist. 29*(4), 1165–1188.

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

Chen, J. and A. K. Gupta (2012). *Parametric statistical change point analysis (2nd Ed.)*. Basel; Cambridge, MA: Birkhäuser Verlag.

Eggermont, P. and V. LaRiccia (2009). *Maximum penalized likelihood estimation*, Volume II. Springer.

Grégoire, G. and Z. Hamrouni (2002). Change point estimation by local linear smoothing. *J. Multivariate Anal. 83*(1), 56–83.

Gu, C. (2013). *Smoothing Spline ANOVA Models (2nd Ed.)*. New York: Springer-Verlag.

Hariz, S. B., J. J. Wylie, and Q. Zhang (2007). Optimal rate of convergence for nonparametric change-point estimators for nonstationary sequences. *Ann. Statist. 35*(4), 1802–1826.

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

Inclán, C. and G. C. Tiao (1994). Use of cumulative sums of squares for retrospective detection of changes of variance. *J. Amer. Statist. Assoc. 89*, 913–923.

Loader, C. R. (1996). Change point estimation using nonparametric regression. *Ann. Statist. 24* (4), 1667–1678.

Matteson, D. S. and N. A. James (2014). A nonparametric approach for multiple change point analysis of multivariate data. *J. Amer. Statist. Assoc. 109* (505), 334–345.

Pan, J. and J. Chen (2006). Application of modified information criterion to multiple change point problems. *J. Multivariate Anal. 97*, 2221–2241.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist. 6*, 461–464.

Shang, Z. and G. Cheng (2013). Local and global asymptotic inference in smoothing spline models. *Ann. Statist. 41* (5), 2608–2638.

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok (Eds.), *Compressed Sensing: Theory and Applications*, Chapter 5, pp. 210–268. Cambridge: Cambridge University Press.

Wahba, G. (1990). *Spline Models for Observational Data*, Volume 59 of CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia: SIAM.

subsequences of residuals, $\{y_i - \widehat{f}^{(\iota-1)}(i/n) : i = 1, \ldots, \widehat{\tau}^{(\iota)}\}$ and $\{y_i - \widehat{f}^{(\iota-1)}(i/n) : i = \widehat{\tau}^{(\iota)} + 1, \ldots, n\}$.

(b) Now given the estimates $\widehat{\tau}^{(\iota)}$, $[\widehat{\sigma}^2]^{(\iota)}$ and $[\widehat{\delta}^2]^{(\iota)}$, we update the mean estimate by the minimizer of (2) where τ, σ^2 and δ^2 are replaced respectively by their current estimates.

3. Iterate until the algorithm converges.

2.2 Mean Estimation Given τ, σ^2, and δ^2

When τ, σ^2 and δ^2 are given, the mean function f_0 is estimated as the minimizer of the penalized weighted least squares (2) in a reproducing kernel Hilbert space $\mathcal{H}$ of functions on the domain $\mathcal{T}$. A reproducing kernel Hilbert space (RKHS) is a Hilbert space $\mathcal{H}$ where the evaluation functional $[t] : \mathcal{H} \to \mathbb{R}, f \mapsto f(t)$ is continuous for every $t \in \mathcal{T}$. The Riesz Representation Theorem then indicates that for all $t \in \mathcal{T}$ there exists a unique function $R_t \in \mathcal{H}$ with the reproducing property $\langle R_t, f \rangle = [t](f) = f(t)$, where $\langle \cdot, \cdot \rangle$ is the inner product on $\mathcal{H}$. Now the reproducing kernel R of $\mathcal{H}$ is defined as a function $R : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ such that $R(s,t) = \langle R_s, R_t \rangle$. One can show that each RKHS is uniquely associated with a reproducing kernel and vice versa.

Note that the penalty functional J in (2) is a squared semi-norm on $\mathcal{H}$. The null space of J, namely $\mathcal{N}_J = \{f : J(f) = 0\}$, induces a direct sum decomposition $\mathcal{H} = \mathcal{N}_J \oplus \mathcal{H}_J$, where $\mathcal{H}_J$ is the complement of $\mathcal{N}_J$ in $\mathcal{H}$. This then yields a decomposition of the reproducing kernel $R = R_0 + R_J$, where R_0 and R_J are respectively the reproducing kernels on the subspaces $\mathcal{N}_J$ and $\mathcal{H}_J$. See, e.g., Gu (2013, Chapter 2) for more details on RKHSs.

We now introduce an example of cubic smoothing splines to illustrate these concepts. We shall use the cubic smoothing splines in all the numerical studies of the paper.