Equivariant Variance Estimation for Multiple Change-point Model

Ning Hao[1], Yue Selena Niu[1], and Han Xiao[2]

[1] *University of Arizona*

[2] *Rutgers University*

Abstract

The variance of noise plays an important role in many change-point detection procedures and the associated inferences. Most commonly used variance estimators require strong assumptions on the true mean structure or normality of the error distribution, which may not hold in applications. More importantly, the qualities of these estimators have not been discussed systematically in the literature. In this paper, we introduce a framework of equivariant variance estimation for multiple change-point models. In particular, we characterize the set of all equivariant unbiased quadratic variance estimators for a family of change-point model classes, and develop a minimax theory for such estimators.

Keywords: Change-point detection, Inference, Minimax, Quadratic estimator, Total variation, Unbiasedness.

1 Introduction

This paper focuses on the variance estimation under the presence of change points. Our goal is to estimate the noise variance without identifying the locations of the changes, so that the variance estimator can be used in the subsequent change-point detection procedures. We characterize the finite sample minimax risk of the proposed estimator, over a broad model class with little restrictions on the change-point structure. Our estimator is equivariant over the data sequence, which greatly simplifies the calculations and leads to explicit minimax risk bounds.

Change points or structural changes have emerged from many applications, and thus been extensively studied in statistics (Fryzlewicz, 2014; Frick et al., 2014; Killick et al., 2012), biological science (Zhang & Siegmund, 2007; Niu & Zhang, 2012), econometrics (Bai & Perron, 1998; Altissimo & Corradi, 2003; Banerjee & Urga, 2005; Juhl & Xiao, 2009; Oka & Qu, 2011), engineering (Lavielle, 2005; Arlot et al., 2019) and many other fields. The literature on the change point analysis has been vast, so we only sample a small portion here. For overviews, see Perron (2006), Chen & Gupta (2012), Niu et al. (2016) and Truong et al. (2020).

A premier goal of change-point detection is to estimate and make inferences about the change-point locations. A good variance estimator is vital in many change-point detection procedures. For example, in binary segmentation and related methods (Olshen et al., 2004; Fryzlewicz, 2014), the variance is required to decide when to stop the recursive procedure. In other methods, for example, screening and ranking algorithm (SaRa) in Niu & Zhang (2012) and simultaneous multiscale change-point estimator (SMUCE) in Frick et al. (2014), the choice of tuning or thresholding parameters depends on the variance. In general, it is important to gauge the noise level, which determines the optimal detection boundary and detectability of the change-point problem (Arias-Castro et al., 2005). Moreover, an accurate and reliable estimate of the variance is necessary for constructing confidence sets of the change points. In practice, the noise variance is usually needed and estimated as the first step of a change-point analysis. However, most commonly used variance estimators, reviewed in Section 2.1, are based on some technical assumptions and can be severely biased when these assumptions fail to hold. The quality of these estimators, such as unbiasedness and efficiency, has been less studied. In fact, to our best knowledge, the exact unbiased variance estimator under a finite sample setup has not been discussed before this work. There are two main challenges to the error variance estimation for change-point models. First, the information on the mean structure such as the number of change points and jump magnitudes is unknown, while complex mean structures often makes the variance estimation more difficult. Second, the noise may not be Gaussian in practice, while many methods work well only under normality. In spite of the importance of this problem and these issues, there has been no systematic study on variance estimation for the multiple change-point model (1). This work aims to fill this gap.

Our approach is inspired by the classical difference-based variance estimation in nonparametric regression, studied in Rice (1984); Gasser et al. (1986); Müller & Stadtmüller (1987); Hall et al. (1990), among many others. In particular, Müller & Stadtmüller (1999) innovatively builds a variance estimator by regressing the lag-k Rice estimators on the lags, in the context of nonparametric regression with discontinuities. Recent developments along this direction include Tong et al. (2013); Tecuapetla-Gómez & Munk (2017); see also a recent review (Levine & Tecuapetla-Gomez, 2019). These works focused on asymptotic analysis of variance estimation for more flexible models, and hence required much stronger conditions on the number of change points or discontinuities. In contrast to the existing literature, we narrow down to change-point models, but the thrust of our study is to have exact and non-asymptotic results regarding the unbiasedness and the minimax risk of the variance estimators, under minimal conditions. To the best of our knowledge, similar results have not appeared in the literature, and are difficult to obtain without the equivariance framework introduced in this paper.

In this paper, we develop a new framework of equivariant variance estimation. Roughly speaking, we will embed the data index set $[n] = \{1, ..., n\}$ on a circle instead of the usual straight line segment so the indices n and 1 are neighbors. In other words, there is no 'head' or 'tail' in the index set, and every position plays the same role. As we will illustrate in Section 2.4, there is a natural cyclic group action on the index set, which leads to an equivariant estimation framework. Under this framework, we are able to characterize all the equivariant unbiased quadratic variance estimators for a family of change-point model classes, and establish a minimax theory on variance estimation. This family of change-point model classes, denoted by Θ_L, is indexed by a positive integer L, which is the minimal distance between change-point locations allowed for any mean structure in the class. In general, a smaller L leads to a broader model class, and hence, a higher minimax risk. In this work, we give both lower and upper bounds in nonasymptotic forms for the minimax risk of equivariant unbiased quadratic estimators for these model classes. Another advantage of the equivariant framework is that it requires minimal assumptions on the noise distribution. In fact, our theoretical analysis relies on no other assumption than the existence of the fourth moment. In particular, the performance of the proposed framework is guaranteed also for skewed or heavy-tailed distributions. We also note that the notion of equivariance has not been sufficiently explored in the literature except Olshen et al. (2004), which focuses on short segment detection rather than a framework of equivariant estimation.

To summarize the main contributions of our work, first, we introduce a new framework on equivariant variance estimation, and characterize the equivariant unbiased quadratic variance estimators for a family of change-point model classes. This framework resembles the classical theory of linear unbiased estimation, but is also technically more complicated. Second, we derive nonasymptotic lower and upper minimax risk bounds for the proposed estimators. In particular, in Corollary 2, we give a surprisingly simple and exact answer to the minimax problem with an explicit minimax risk for a broad change-point model class Θ_2. Third, our approach requires minimal model assumptions on the noise distribution and mean structure, which can hardly be weaken further. Last but not least, we suggest an equivariant variance estimator that is computationally simple and practically useful in applications. As a by-product, we show the ℓ_2 risk explicitly for the regression based estimator proposed by Müller & Stadtmüller (1999) and theoretically compare its risk with our method. Therefore, our theoretical result implies that the Müller-Stadtmüller estimator is nearly minimax. In the numerical studies, compared to an oracle variance estimator that knows the true mean, the relative efficiency of our methods is often within 1.5 across different scenarios.

2 Variance estimation

2.1 Existing variance estimators

In this paper, we focus on the problem of noise variance estimation for a multiple change-point model. In particular, consider a sequence of random variables $X_1, \cdots, X_n$ satisfying

$$X_i = \theta_i + \varepsilon_i, \qquad 1 \leq i \leq n, \quad \text{with} \tag{1}$$

$$\theta_1 = \theta_2 = \cdots = \theta_{\tau_1} \neq \theta_{\tau_1+1} = \cdots = \theta_{\tau_2} \neq \theta_{\tau_2+1} = \cdots \quad \cdots = \theta_{\tau_J} \neq \theta_{\tau_J+1} = \cdots = \theta_n, \tag{2}$$

where the mean vector $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)^\top$ is piecewise constant, and $\boldsymbol{\tau} = (\tau_1, ..., \tau_J)^\top$ is the location vector of change points. We assume that the noises $\{\varepsilon_i\}_{i=1}^n$ are independent and identically distributed (i.i.d.) with $\mathrm{E}(\varepsilon_1) = 0$ and $\mathrm{Var}(\varepsilon_1) = \sigma^2 > 0$.

Many estimators for the variance or standard deviation of the additive noise have been employed in recent works on change-point detection. For example, Fryzlewicz (2014) uses a median absolute deviation (MAD) estimator (Hampel, 1974), defined by

$$\hat{\sigma}_1 = 1.4826 * \mathrm{med}(|\boldsymbol{X} - \mathrm{med}(\boldsymbol{X})|), \tag{3}$$

where $\mathrm{med}(\boldsymbol{X})$ is the median of the vector $\boldsymbol{X} = (X_1, ..., X_n)^\top$, the constant 1.4826 the ratio between standard deviation and the third quartile of the Gaussian distribution. One advantage of this estimator is that it is robust against outliers. Obviously, the method depends on Gaussianity assumption and a sparsity assumption that $\boldsymbol{\theta}$ is a constant vector except a small number of entries.

Frick et al. (2014) suggests an estimator used in Davies & Kovac (2001),

$$\hat{\sigma}_2 = \frac{1.48}{\sqrt{2}} * \mathrm{med}(|\boldsymbol{X}_{(-1)} - \boldsymbol{X}_{(-n)}|), \tag{4}$$

where $\boldsymbol{X}_{(-1)} = (X_2, ..., X_n)^\top$ and $\boldsymbol{X}_{(-n)} = (X_1, ..., X_{n-1})^\top$. This estimator is similar to the

MAD except that it does not require $\boldsymbol{\theta}$ to be an almost constant vector. Nevertheless, it still needs the normality of the noises.

The Rice estimator, introduced in Rice (1984),

$$\hat{\sigma}_3^2 = \frac{1}{2n}\|\boldsymbol{X}_{(-1)} - \boldsymbol{X}_{(-n)}\|^2 \tag{5}$$

is another popular method. For example, Pique-Regi et al. (2008) uses it for variance estimation. It does not depend on Gaussianity of the noise. But it might be seriously biased. In fact, as an immediate consequence of Proposition 2, the bias of $\hat{\sigma}_3^2$ is $\frac{1}{n}(V(\boldsymbol{\theta})/2 - \sigma^2)$, where $V(\boldsymbol{\theta}) = \sum_{i=1}^{n-1}(\theta_i - \theta_{i+1})^2$. To eliminate the bias and improve the efficiency, (Müller & Stadtmüller, 1999) proposed a regression based estimator via lag-k Rice estimators. As we will see in Section 2.3, it is a special case of difference-based quadratic variance estimator, which has been a popular approach in nonparametric regression (Dette et al., 1998). Nevertheless, it seems that this approach has not been widely recognized and employed in change-point analysis. There are a few interesting open problems to be answered for the Müller-Stadtmüller estimator. First, can we find its risk with respect to a loss function, e.g., ℓ_2 loss? Second, the quality of any variance estimators to a change-point model highly depends on the mean structure $\boldsymbol{\theta}$. It is desirable to find optimal or nearly optimal variance estimators for certain change-point model classes. In particular, is the Müller-Stadtmüller estimator optimal? Undoubtedly, affirmative answers to these questions will promote the applications of the difference-based quadratic variance estimator including the Müller-Stadtmüller estimator in the field of change-point analysis.

In fact, direct answers to these questions are difficult, as we explain in the supplementary document. Instead, we take a detour via an equivariance framework and answer all questions above.

2.2 Model descriptions

In model (1), the data vector $\boldsymbol{X} = (X_1, X_2, \dots, X_n)^\top$ is observed and indexed by the set $[n] = \{1, ..., n\}$. We define a segment, denoted by $[k, \ell]$, as a subset of $[n]$ consisting of consecutive integers $\{k, k+1, \cdots, \ell\}$. The working model (1) is standard and widely used in the literature. Here we make and emphasize a key extension. That is, the index set is arranged on a circle, and the indices 1 and n do not play special roles as start and end points. Consequently, a segment $[k, \ell]$ with $k \geq \ell$ is also well-defined. For example, $[n-1, 3] = \{n-1, n, 1, 2, 3\}$. For the mean vector $\boldsymbol{\theta}$ with the form (2), we assume that it consists of J segments with constant means, $[\tau_1 + 1, \tau_2]$,..., $[\tau_J + 1, \tau_1]$, which are separated by change point $1 \leq \tau_1 < \tau_2 < \cdots < \tau_J \leq n$. Denote the common value of θ_i on the segment $[\tau_j + 1, \tau_{j+1}]$ by μ_j. For a mean vector $\boldsymbol{\theta}$, we denote by $L(\boldsymbol{\theta})$ the minimal length of all constant segments in $\boldsymbol{\theta}$. The magnitude of $L(\boldsymbol{\theta})$ is a complexity measure of a change-point model. We will consider a family of nested model classes $\Theta_2 \supset \Theta_3 \supset \cdots$, where

$$\Theta_L = \{\boldsymbol{\theta} \in \mathbb{R}^n : \ L(\boldsymbol{\theta}) \geq L\}. \tag{6}$$

In general, the larger L is, the easier the change-point analysis. In particular, when $L(\boldsymbol{\theta}) = 1$, each observation can have its own mean different from all others, and there is no sensible change-point problem. Therefore, we only consider the case $L(\boldsymbol{\theta}) \geq 2$ in this paper. Note that, by definition, $L(\boldsymbol{\theta}) = n$ if $\boldsymbol{\theta}$ is a constant vector, and otherwise, $L(\boldsymbol{\theta}) \leq n/2$.

Note that the classical model treats the first segment and the last segment of $\boldsymbol{\theta}$ as two separated segments. That is, the index n is treated as a known change point, no matter whether $\theta_1 = \theta_n$ or not. The classical model classes can be defined by

$$\Theta_L^c = \{\boldsymbol{\theta} \in \mathbb{R}^n : \ L(\boldsymbol{\theta}) \geq L, \ \tau_J = n\}. \tag{7}$$

In fact, $\Theta_L \supset \Theta_L^c$ by definition. For example, let $n = 8$ and $\boldsymbol{\theta} = (0, 0, 1, 1, 1, 1, 0, 0)^\top$. We have $\boldsymbol{\theta} \in \Theta_4$ but $\boldsymbol{\theta} \notin \Theta_4^c$. The larger generality of Θ_L over Θ_L^c can be negligible in real applications. However, as we will see, it is advantageous to work on the family (6) to obtain neat theoretical results.

We use i, k, h, $\ell \in [n]$ to denote the index of the data, and K and L to denote the length of segments. Occasionally, an index i in X_i or θ_i may go beyond $[n]$ in formulas. In that case, we use the convention $X_i = X_{i-nM}$ where M is the unique integer such that $i - nM \in [n]$. Similarly, we use $j \in [J]$ to denote the index of change points and use the convention $\tau_{J+1} = \tau_1$. The length of a segment $[k, \ell]$ is defined as the cardinality of the set $[k, \ell]$, which is $\ell - k + 1$ when $k \leq \ell$ and $n + \ell - k + 1$ otherwise.

We assume the following condition on the error distribution in this paper.

Condition 1. $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. with $\mathrm{E}(\varepsilon_1) = 0$, $\mathrm{Var}(\varepsilon_1) = \sigma^2$, and $\kappa_4 = \mathrm{E}(\varepsilon_1^4)/\sigma^4 < \infty$.

We view this assumption as a "minimal" one for the variance estimation problem, because there is no distributional assumption. The existence of the 4-th moment is necessary for studying the mean squared error of the variance estimator.

We define two quantities related to the mean structure

$$V(\boldsymbol{\theta}) = \sum_{i=1}^{n-1} (\theta_i - \theta_{i+1})^2$$
$$W(\boldsymbol{\theta}) = \sum_{i=1}^{n} (\theta_i - \theta_{i+1})^2 = V(\boldsymbol{\theta}) + (\theta_n - \theta_1)^2 = \sum_{j=1}^{J} (\mu_j - \mu_{j+1})^2.$$

In fact, $V(\boldsymbol{\theta})$ and $W(\boldsymbol{\theta})$ measure the total variation of the mean vector in ℓ_2-norm. There is no change point in the sequence if and only if $V(\boldsymbol{\theta}) = W(\boldsymbol{\theta}) = 0$.

With the convention that $X_i = X_{n+i}$, we define

$$T_k = \sum_{i=1}^{n} (X_i - X_{i+k})^2,$$

which plays a central role in our variance estimation framework. In fact, it can be considered as a circular version of the lag-k Rice estimator, defined as

$$S_k = \sum_{i=1}^{n-k} (X_i - X_{i+k})^2.$$

In particular, S_1 is called Rice estimator, introduced in Rice (1984).

2.3 An equivariant approach for variance estimation

The means and covariances of T_k's can be calculated as follows.

Proposition 1 *Under Condition 1, for* $1 \le k \le L(\boldsymbol{\theta})$,

$$\mathrm{E}T_k = 2n\sigma^2 + kW(\boldsymbol{\theta}).$$

Moreover, for $1 \le k \le L(\boldsymbol{\theta})/2$,

$$\mathrm{Var}(T_k) = 4n\kappa_4\sigma^4 + 8k\sigma^2 W(\boldsymbol{\theta});$$

and for $1 \le k < h \le L(\boldsymbol{\theta})/2$,

$$\mathrm{Cov}(T_k, T_h) = 4n(\kappa_4 - 1)\sigma^4 + 8k\sigma^2 W(\boldsymbol{\theta}).$$

With Proposition 1, we rescale T_k and consider a regression model

$$Y_k = \alpha + k\beta + e_k, \quad k = 1, ..., K \tag{8}$$

where $Y_k = T_k/(2n)$, $(\alpha, \beta)^\top = (\sigma^2, W(\boldsymbol{\theta})/(2n))^\top$, and e_k is the noise term with mean zero and covariance

$$\mathrm{Cov}(e_1, ..., e_K)^\top = \boldsymbol{\Sigma} = \frac{\sigma^4}{n}\left[\boldsymbol{I}_K + (\kappa_4 - 1)\mathbf{1}_K\mathbf{1}_K^\top + \frac{2W(\boldsymbol{\theta})}{n\sigma^2}\boldsymbol{H}_K\right], \tag{9}$$

where $\boldsymbol{I}_K$ is the $K \times K$ identity matrix, $\mathbf{1}_K$ is a vector of length K with all entries equal to 1, $\boldsymbol{H}_K = (H_{ij})$ is a $K \times K$ matrix with $H_{ij} = \min\{i, j\}$. As Y_k and T_k are easily calculated from the data, we can estimate the variance, i.e., the intercept α in the regression model (8), by the ordinary least squares (OLS) estimator, denoted by $\hat{\alpha}_K$. Specifically, let $\boldsymbol{Y}_K = (Y_1, \ldots, Y_k)^\top$, $\boldsymbol{\eta}_K = (1, 2, \ldots, K)^\top$, $\boldsymbol{Z}_K = (\mathbf{1}_K, \boldsymbol{\eta}_K)$, then

$$\hat{\alpha}_K = (1, 0)(\boldsymbol{Z}_K^\top \boldsymbol{Z}_K)^{-1} \boldsymbol{Z}_K^\top \boldsymbol{Y}_K. \tag{10}$$

Theorem 1 *Assume Condition 1. The OLS estimator $\hat{\alpha}_K$ is unbiased when $2 \leq K \leq L(\boldsymbol{\theta})$. Moreover, if $K \leq L(\boldsymbol{\theta})/2$, we have*

$$\mathrm{Var}(\hat{\alpha}_K) = \frac{\sigma^4}{n} \left(\kappa_4 - 1 + \frac{4K + 2}{K(K-1)} + \frac{2W(\boldsymbol{\theta})}{n\sigma^2} \frac{(K+1)(K+2)(2K+1)}{15K(K-1)} \right). \tag{11}$$

If $K \leq L(\boldsymbol{\theta})$,

$$\mathrm{Var}(\hat{\alpha}_K) \leq \frac{\sigma^4}{n} \left(\kappa_4 - 1 + \frac{4K + 2}{K(K-1)} + \frac{W(\boldsymbol{\theta})}{n\sigma^2} \frac{(K+1)(K+2)^2}{3K(K-1)} \right). \tag{12}$$

Theorem 1 gives an exact ℓ_2 risk of the variance estimator $\hat{\alpha}_K$ for $2 \leq K \leq L(\boldsymbol{\theta})/2$. Note that the risk depends on $\boldsymbol{\theta}$ only through its total variation $W(\boldsymbol{\theta})$. When $K > L(\boldsymbol{\theta})/2$, the exact risk also depends on other information of the mean, besides the total variation $W(\boldsymbol{\theta})$. See Theorem 3 for more details. In the proof of Theorem 1 in the supplementary document, we show that the equality in (12) is achieved for a specific $\boldsymbol{\theta}$ satisfying: $K = L(\boldsymbol{\theta})$, n/K is an even number, all segments are of the same length, and the segment means μ_j have the same absolute value, but with alternating signs. Therefore, the upper bound provided in (12) is tight.

There are three summands in the ℓ_2-risk of $\hat{\alpha}_K$ (11). The first summand $\frac{\sigma^4}{n}(\kappa_4 - 1)$ is equal to $\mathrm{Var}(\hat{\sigma}_O^2)$ where

$$\hat{\sigma}_O^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \theta_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 \tag{13}$$

is the oracle estimator when the true mean is known. When $K \leq L(\boldsymbol{\theta})/2$, according to Proposition 1, the generalized least squares (GLS) estimator $\tilde{\alpha}_K$ based on model (8) is obtained using the covariance matrix (9). Clearly $\tilde{\alpha}_K$ depends on $\boldsymbol{\theta}$ through $W(\boldsymbol{\theta})/\sigma^2$ in the covariance (9). In a special case when $W(\boldsymbol{\theta}) = 0$, the covariance is compound symmetric, and the OLS and GLS estimators coincide (McElroy, 1967) and equal to $\hat{\sigma}_{O,K}^2 := \frac{1}{K}\sum_{k=1}^{K} Y_k$ with ℓ_2-risk $\frac{\sigma^4}{n}\left(\kappa_4 - 1 + \frac{4K+2}{K(K-1)}\right)$. Therefore, the first two summands in (11) can not be reduced for any linear unbiased estimators based on $\{Y_k\}_{k=1}^K$. We will elaborate the related minimax theory in subsection 2.5.

We may also calculate the mean and covariance of S_k's.

Proposition 2 *Under Condition 1 with $\tau_J = n$, for $1 \leq k \leq L(\boldsymbol{\theta})$,*

$$\mathrm{E}S_k = 2n\sigma^2 + k\left[V(\boldsymbol{\theta}) - 2\sigma^2\right].$$

Moreover, if $\mathrm{E}(\varepsilon_1^3) = 0$, for $1 \leq k \leq L(\boldsymbol{\theta})/2$,

$$\mathrm{Var}(S_k) = 2(n-k)(\kappa_4 + 1)\sigma^4 + 2(n-2k)(\kappa_4 - 1)\sigma^4 + 8k\sigma^2 V(\boldsymbol{\theta});$$

and for $1 \leq k < h \leq L(\boldsymbol{\theta})/2$,

$$\mathrm{Cov}(S_k, S_h) = (4n - 4h - 2k)(\kappa_4 - 1)\sigma^4 + 8k\sigma^2 V(\boldsymbol{\theta}).$$

To our best knowledge, Müller and Stadtmüller first constructed variance estimators via a regression approach based on S_k's (Müller & Stadtmüller, 1999). They studied variance estimation and tests for jump points in nonparametric estimation under an asymptotic setting

$L(\boldsymbol{\theta})/n \to c$ as $n \to \infty$.

Remark. The condition $\tau_J = n$ in Proposition 2 means that when study the properties of S_k's, we consider the classical change-point model where the first segment is $[1, \tau_1]$, and the last segment is $[\tau_{J-1}+1, n]$.

Comparing with T_k's, the mean and covariance structure of S_k's is more complex. Moreover, Proposition 2 requires one more condition $\mathrm{E}\varepsilon_1^3 = 0$, i.e. zero skewness. The following proposition gives a precise comparison of the OLS estimators based on T_k's and S_k's.

Proposition 3 *Assume Condition 1,* $\mathrm{E}(\varepsilon_1^3) = 0$*, and* $\tau_J = n$*. Let* $\breve{\alpha}_K$ *be the OLS estimator obtained by using* S_k *in place of* T_k*. Then* $\breve{\alpha}_K$ *is unbiased when* $2 \le K \le L(\boldsymbol{\theta})$*. Moreover, if* $K \le L(\boldsymbol{\theta})/2$*, we have*

$$\mathrm{Var}(\breve{\alpha}_K) = \frac{\sigma^4}{n}\left(\kappa_4 - 1 + \frac{4K+2}{K(K-1)} + \frac{2V(\boldsymbol{\theta})}{n\sigma^2}\cdot\frac{(K+1)(K+2)(2K+1)}{15K(K-1)} + \frac{1}{n}\cdot\frac{2(K-7)(K+1)(K+2)}{K(K-1)}\right)$$

If $K \le L(\boldsymbol{\theta})$ *and* $K \le n/2$*,*

$$\mathrm{Var}(\breve{\alpha}_K) \le \frac{\sigma^4}{n}\left(\kappa_4 - 1 + \frac{4K+2}{K(K-1)} + \frac{V(\boldsymbol{\theta})}{n\sigma^2}\cdot\frac{(K+1)(K+2)^2}{3K(K-1)} + \frac{1}{n}\cdot\frac{2(K-7)(K+1)(K+2)}{K(K-1)}\right).$$

We call $\breve{\alpha}_K$ the Müller-Stadtmüller (MS) estimator. As an immediate consequence of Theorem 1 and Proposition 3, when $2 \le K \le L(\boldsymbol{\theta})/2$,

$$\mathrm{Var}(\breve{\alpha}_K) - \mathrm{Var}(\hat{\alpha}_K) = \frac{\sigma^2}{n^2}\left\{\left[\sigma^2 - 2(\theta_1-\theta_n)^2\right]\cdot\frac{(K+1)(K+2)(2K+1)}{15K(K-1)} - \sigma^2\cdot\frac{(K+1)(K+2)}{K(K-1)}\right\}.$$

It follows that $\hat{\alpha}_K$ has a smaller variance if $\theta_1 = \theta_n$ and $K \ge 7$; and $\breve{\alpha}_K$ has a smaller variance if $(\theta_1-\theta_n)^2 > \sigma^2/2$. Asymptotically, $\mathrm{Var}(\breve{\alpha}_K) - \mathrm{Var}(\hat{\alpha}_K) = o(\mathrm{Var}(\breve{\alpha}_K))$ when $K(\sigma^2 + (\theta_1 - \theta_n)^2) = o(n)$. So these two estimators often perform similarly, which is also verified by our numerical studies. In this paper, we aim to derive nonasymptotic and exact risk bounds for the variance estimators, which seems too complicated using S_k's. Therefore, we focus on T_k's subsequently and introduce the equivariant framework in the next subsection.

2.4 Equivariant unbiased estimation

Geometrically, we can embed the index set $[n] = \{1, ..., n\}$ into the unit circle $\mathcal{S}^1 \subset \mathbb{R}^2$ by the exponential map $\pi_n : i \mapsto e^{\frac{2\pi i\sqrt{-1}}{n}}$. The set $[n]$ is invariant of natural group action $\mathbb{Z}_n \hookrightarrow \mathcal{S}^1$, where $\mathbb{Z}_n$ is the cyclic group of order n, and the unit element $1 \in \mathbb{Z}_n$ maps $\mathcal{S}^1$ to itself via a rotation by an angle $\frac{2\pi}{n}$. This group action naturally induces a group action of $\mathbb{Z}_n$ on the sample space $\mathbb{R}^n$, where the unit element $1 \in \mathbb{Z}_n$ maps an n-vector $(X_1, ..., X_n)^\top$ to $(X_2, ..., X_n, X_1)^\top$. There is another way to represent this group action via $n \times n$ circulant matrices. Define $\boldsymbol{C}_k$ as a circulant matrix with its (i, j) entry

$$C_{k,ij} = \begin{cases} 1, & j - i = k \mod n \\ 0, & \text{otherwise.} \end{cases}$$

Again, we may treat the subscript k in $\boldsymbol{C}_k$ as a number modulo n. It is easy to verify that $\boldsymbol{C}_k\boldsymbol{C}_\ell = \boldsymbol{C}_{k+\ell}$ holds under standard matrix multiplication and $\boldsymbol{C}_k^\top = \boldsymbol{C}_{-k} = \boldsymbol{C}_{n-k}$, so $\mathcal{C}_n = \{\boldsymbol{C}_k\}$ is a group isomorphic to $\mathbb{Z}_n$. Under this isomorphism, the group action $\mathbb{Z}_n \hookrightarrow \mathbb{R}^n$ can be represented by matrix multiplication $\boldsymbol{X} \mapsto \boldsymbol{C}_k\boldsymbol{X}$. Note that both the parameter space of the mean vector, Θ, and the sample space, $\mathcal{X}$, are $\mathbb{R}^n$ for the change-point model. An estimator $\hat{\boldsymbol{\theta}}$ of the mean vector $\boldsymbol{\theta}$ is called *equivariant* if and only if $\boldsymbol{C}_k\hat{\boldsymbol{\theta}}(\boldsymbol{X}) = \hat{\boldsymbol{\theta}}(\boldsymbol{C}_k\boldsymbol{X})$ for all k, i.e., the estimation procedure commutes with the group action. For the problem of variance estimation, as the group action does not affect the value of variance parameter σ^2, a variance estimator $\hat{\sigma}^2$ is equivariant (or simply invariant) if $\hat{\sigma}^2(\boldsymbol{X}) = \hat{\sigma}^2(\boldsymbol{C}_k\boldsymbol{X})$.

In this sense, T_k is an equivariant version of S_k because the values of T_k's remain the same under the group action. Consequently, we have

Proposition 4 *$\hat{\alpha}_K$ is an equivariant variance estimator. Under condition 1, $\hat{\alpha}_K$ is equivariant and unbiased for $2 \le K \le L(\boldsymbol{\theta})$.*

We consider the class of quadratic estimators of the form $\sum_{i,j=1}^n a_{ij}X_iX_j$, or $\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X}$, where $\boldsymbol{A} = (a_{ij})$ is a symmetric matrix. It is straightforward to see $Y_k = \frac{1}{2n}T_k = \boldsymbol{X}^\top\boldsymbol{A}_k\boldsymbol{X}$

with $\boldsymbol{A}_k = \frac{1}{n}\left(\boldsymbol{I} - \frac{1}{2}\boldsymbol{C}_k - \frac{1}{2}\boldsymbol{C}_k^\top\right)$. That is, $\{T_k\}_{k=1}^L$ and their linear combinations are quadratic estimators. It turns out that any equivariant unbiased quadratic variance estimator for model class Θ_L must be a linear combination of T_1,..., T_L, as characterized by the following theorem.

Theorem 2 *The set of all equivariant unbiased quadratic variance estimators for the model class Θ_L is*

$$\mathcal{Q}_L = \left\{ \frac{1}{2n}\sum_{k=1}^L c_k T_k = \sum_{k=1}^L c_k Y_k \,:\, c_1, ..., c_L \in \mathbb{R}, \sum_{k=1}^L c_k = 1, \sum_{k=1}^L k c_k = 0 \right\}.$$

Interestingly, $\mathcal{Q}_2$ consists of only one estimator, i.e., $\hat{\alpha}_2 = 2Y_1 - Y_2$. As a corollary of Theorems 1 and 2, we have

Corollary 1 *The OLS estimator $\hat{\alpha}_2 = 2Y_1 - Y_2$ is the unique quadratic equivariant unbiased variance estimator for model class Θ_2. Its variance satisfies*

$$\mathrm{Var}(\hat{\alpha}_2) \le \frac{\sigma^4}{n}\left(\kappa_4 + 4 + \frac{8W(\boldsymbol{\theta})}{n\sigma^2}\right).$$

Before we conclude this subsection, we point out that it is also possible to characterize the unbiased quadratic estimators over the class of classical change-point models Θ_L^c defined in (7). It turns out this characterization is much more complicated than Theorem 2. Furthermore, the variance of an unbiased $\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X}$ over Θ_L^c also depends on the mean vector $\boldsymbol{\theta}$ in a more complicated way. These observations give us another motivation to consider the equivariant estimators over the larger class Θ_L. We discuss the unbiased estimators over Θ_L^c with more details in Supplementary 4.

2.5 Minimax risk

Theorem 2 concludes that all equivariant unbiased quadratic estimators for model class Θ_L are linear combinations of Y_1,...,Y_L, including the OLS estimator studied in subsection 2.3.

A natural question is whether the OLS estimator is optimal, and if not, how far it is from an optimal estimator. In this subsection, we will answer this question from the perspective of minimax theory.

Consider the class $\mathcal{Q}_L$ of all equivariant unbiased estimators over the model class

$$\Theta_{L,w} = \{(\boldsymbol{\theta}, \sigma^2) : \ L(\boldsymbol{\theta}) \geq L, \, W(\boldsymbol{\theta})/(n\sigma^2) \leq w, \, \sigma^2 > 0\}, \quad \text{where} \quad L \geq 2, \, w \geq 0.$$

For any estimator $\hat{\sigma}^2$, define the ℓ_2 risk up to a factor $\frac{\sigma^4}{n}$

$$r(\hat{\sigma}^2) = \frac{n}{\sigma^4} \mathrm{E}(\hat{\sigma}^2 - \sigma^2)^2.$$

This risk is scale invariant by definition. As we will show soon, for a fixed model $(\boldsymbol{\theta}, \sigma^2)$, the risk of the optimal estimator depends on the minimal segment length $L(\boldsymbol{\theta})$ and the ratio $W(\boldsymbol{\theta})/(n\sigma^2)$. Therefore, we consider the model class $\Theta_{L,w}$ in our minimax analysis, where the two parameters L and w bound these two quantities respectively. Define the minimax risk of all equivariant unbiased estimators in $\mathcal{Q}_L$ over model class $\Theta_{L,w}$ as follows.

$$r_{L,w} = \min_{\hat{\sigma}^2 \in \mathcal{Q}_L} \max_{(\boldsymbol{\theta}, \sigma^2) \in \Theta_{L,w}} r(\hat{\sigma}^2). \tag{14}$$

We can solve the minimax problem for the case $L = 2$ as a simple corollary of Theorems 1 and 2.

Corollary 2 *$\hat{\alpha}_2 = 2Y_1 - Y_2$ is the minimax estimator for model class $\Theta_{2,w}$ with minimax risk $r_{2,w} \leq \kappa_4 + 4 + 8w$ with equality holding when n is a multiple of 4.*

Corollary 2 gives an elegant minimax solution for the broadest model class considered in this paper. At the level of $L = 2$, the OLS estimator is optimal, no matter what value w takes. Intuitively, as L grows and the model class shrinks, we may borrow more information from neighbors because of the piecewise constant mean structure, and get lower minimax risk. Nevertheless, the minimax estimator and the exact risk are difficult to find for $L \geq 3$.

We will provide instead both lower and upper bounds of the minimax risk. We first calculate the risk of any equivariant unbiased estimator in $\mathcal{Q}_L$.

Theorem 3 *Let $\boldsymbol{c} = (c_1, c_2, \ldots, c_L)^\top$ such that $\sum_{k=1}^L c_k = 1$ and $\sum_{k=1}^L kc_k = 0$. For $(\boldsymbol{\theta}, \sigma^2) \in \Theta_{L,w}$, the risk of $\hat{\sigma}^2_{\boldsymbol{c}} = \sum_{k=1}^L c_k Y_k \in \mathcal{Q}_L$ is*

$$r(\hat{\sigma}^2_{\boldsymbol{c}}) = \kappa_4 - 1 + \boldsymbol{c}^\top \left(\boldsymbol{I}_L - \frac{W(\boldsymbol{\theta})}{n\sigma^2} \boldsymbol{G}(\boldsymbol{\theta}) \right) \boldsymbol{c}, \tag{15}$$

where $\boldsymbol{G}(\boldsymbol{\theta}) = (G_{k\ell})$ is a $L \times L$ matrix with

$$G_{k\ell} = |k - \ell| + \frac{1}{W(\boldsymbol{\theta})} \sum_{i=1}^n (\theta_i - \theta_{i+k+\ell})^2. \tag{16}$$

As shown in the proof of Proposition 5, the quadratic form in (15) is positive definite on the constrained linear space which $\boldsymbol{c}$ lies in. Therefore, we can minimize the risk (15) to get the optimal solution in $\mathcal{Q}_L$ for any model in $\Theta_{L,w}$, putting aside the fact that the solution may depend on unknown parameters. Because all estimators in $\mathcal{Q}_L$ are linear combinations of Y_k's, they are also linear estimators of the intercept in model (8). It is not surprising that the optimization problem (15) has the same optimal solution as the least squares problem (8). We state the result formally as below.

Proposition 5 *There is a unique solution to the optimization problem*

$$\text{minimize} \quad r(\hat{\sigma}^2_{\boldsymbol{c}}) \quad \text{subject to} \quad \sum_{k=1}^L c_k = 1, \quad \sum_{k=1}^L kc_k = 0.$$

Let $\boldsymbol{c}_{\boldsymbol{\theta},\sigma^2}$ be the minimizer for a model $(\boldsymbol{\theta}, \sigma^2) \in \Theta_{L,w}$. Then $\hat{\sigma}^2_{\boldsymbol{c}_{\boldsymbol{\theta},\sigma^2}}$ is the GLS estimator of model (8) with $K = L$. Moreover, if $(\boldsymbol{\theta}, \sigma^2) \in \Theta_{2L,w} \subset \Theta_{L,w}$, then $\boldsymbol{c}_{\boldsymbol{\theta},\sigma^2}$ depends on the model $(\boldsymbol{\theta}, \sigma^2)$ only through $W(\boldsymbol{\theta})/(n\sigma^2)$.

By minimizing (15) with linear constraints, we can easily find the optimal $\boldsymbol{c}$ and corresponding risk for an individual model $(\boldsymbol{\theta}, \sigma^2) \in \Theta_{L,w}$. Nevertheless, we see from (16) that the value of $G_{k\ell}$ depends on $\sum_{i=1}^{n}(\theta_i - \theta_{i+k+\ell})^2$, which is not a function of $W(\boldsymbol{\theta})$ when $k+\ell > L(\boldsymbol{\theta})$. Thus, there is no simple way to characterize the behavior of $\boldsymbol{G}(\boldsymbol{\theta})$ for all models in $\Theta_{L,w}$. As a result, it is a highly nontrivial problem to identify the minimax estimator and the minimax risk.

In Theorem 4, we will provide both lower and upper bounds of the minimax risk. We first introduce the main ideas and some necessary notations. We consider the OLS and GLS estimators and their risks over the model class to bound the minimax risk. For OLS, formula (12) in Theorem 1 implies an upper bound of minimax risk.

$$\min_{\hat{\sigma}^2 \in \mathcal{Q}_L} \max_{(\boldsymbol{\theta},\sigma^2)\in\Theta_{L,w}} r(\hat{\sigma}^2) \leq \max_{(\boldsymbol{\theta},\sigma^2)\in\Theta_{L,w}} r(\hat{\alpha}_L) = \kappa_4 - 1 + \frac{4L+2}{L(L-1)} + \frac{(L+1)(L+2)^2}{3L(L-1)} w. \quad (17)$$

For GLS, we consider a smaller model class $\Theta_{2L,w}$, over which the GLS estimator in $\mathcal{Q}_L$ depends on $\boldsymbol{\theta}$ only through $W(\boldsymbol{\theta})/(n\sigma^2)$. Specifically, let $\boldsymbol{\Sigma}_{L,w}$ be the covariance matrix (9) with $K = L$ and $W(\boldsymbol{\theta})/(n\sigma^2) = w$, we define $\tilde{\alpha}_{L,w}$ as the GLS estimator based on (8) and covariance matrix $\boldsymbol{\Sigma}_{L,w}$, i.e.

$$\tilde{\alpha}_{L,w} = (1,0)(\boldsymbol{Z}_L^\top \boldsymbol{\Sigma}_{L,w}^{-1} \boldsymbol{Z}_L)^{-1} \boldsymbol{Z}_L^\top \boldsymbol{\Sigma}_{L,w}^{-1} \boldsymbol{Y}_L.$$

The maximal risk of the GLS $\tilde{\alpha}_{L,w}$ over $\Theta_{2L,w}$ can be derived to offer a lower bound of the minimax risk. Finally, we study a GLS estimator based on an upper bound of the covariance structure (9) and its maximal risk over $\Theta_{L,w}$, which leads to a minimax upper bound different from (17).

Let $\{D_k\}$ be the sequence defined recursively by $D_k = (2+\lambda)D_{k-1} - D_{k-2}$ with initial values $D_0 = 1$, $D_1 = 1 + \lambda$. Define the matrix

$$\boldsymbol{V}_{L,\lambda} := \begin{pmatrix} \frac{1 - D_{L-1}/D_L}{\lambda} & \frac{D_L - 1}{\lambda D_L} \\ \frac{D_L - 1}{\lambda D_L} & \frac{D_{L-1}/D_L + \lambda L - 1}{\lambda^2} \end{pmatrix},$$

and define

$$g_L(\lambda) := \kappa_4 - 1 + \boldsymbol{V}_{L,\lambda}^{-1}[1,1], \tag{18}$$

where $\boldsymbol{V}_{L,\lambda}^{-1}[1,1]$ is the top left entry of the 2×2 matrix $\boldsymbol{V}_{L,\lambda}^{-1}$.

Theorem 4 *Let $r_{L,w}$ be the minimax risk defined in (14), and $g_L(\cdot)$ be a function defined in (18). For the subclass $\Theta_{2L,w}$, the GLS estimator $\tilde{\alpha}_{L,w} \in \mathcal{Q}_L$ is minimax with the risk*

$$\min_{\hat{\sigma}^2 \in \mathcal{Q}_L} \max_{(\boldsymbol{\theta},\sigma^2) \in \Theta_{2L,w}} r(\hat{\sigma}^2) = \max_{(\boldsymbol{\theta},\sigma^2) \in \Theta_{2L,w}} r(\tilde{\alpha}_{L,w}) = g_L(2w),$$

The minimax risk on the model class $\Theta_{L,w}$ satisfies (17) and

$$g_L(2w) \leq r_{L,w} \leq g_L(4w). \tag{19}$$

The function $g_L(\cdot)$ in (18) is defined through the sequence $\{D_k\}$. Although the explicit expression of D_k and hence $g_L(\cdot)$ can be derived, it is complicated and barely provides any additional insight, so we choose not to present it. Instead, we characterize the behavior of $g_L(\cdot)$ around 0 in the following proposition.

Proposition 6 *$g_L(\cdot)$ is a nonnegative increasing function on $[0,\infty)$ with*

$$g_L(0) = \kappa_4 - 1 + \frac{4L+2}{L(L-1)}, \quad \textit{and} \quad g_L'(0) = \frac{(L+1)(L+2)(2L+1)}{15L(L-1)}.$$

This proposition, together with (17), shows that the exceeded minimax risk of the OLS estimator is bounded by

$$\begin{aligned}
&\frac{(L+1)(L+2)^2}{3L(L-1)}w - \frac{2(L+1)(L+2)(2L+1)}{15L(L-1)}w + o(w) \\
=&\frac{(L+1)(L+2)(L+8)}{15L(L-1)}w + o(w).
\end{aligned}$$

As an immediate consequence, we have the following corollary.

Corollary 3 *The OLS estimator $\hat{\alpha}_L$ is asymptotically minimax under condition $w = o(1)$, i.e.,*

$$\lim_{n\to\infty} \max_{(\boldsymbol{\theta},\sigma^2)\in\Theta_{L,w}} r(\hat{\alpha}_K) = \lim_{n\to\infty} \min_{\hat{\sigma}^2\in\mathcal{Q}_L} \max_{(\boldsymbol{\theta},\sigma^2)\in\Theta_{L,w}} r(\hat{\sigma}^2) = \kappa_4 - 1 + \frac{4L+2}{L(L-1)}.$$

In Figure 1, we illustrate the minimax risk bounds discussed above. In particular, we plot the upper bounds given by OLS in (17) (labeled by `OLS-L`) and by GLS in (19) (labeled by `GLS-L`). (17) is tighter when w is small, and (19) gives a sharper bound when w is large. Two other lines in Figure 1, labeled by `OLS-2L` and `GLS-2L`, are for the risks of the OLS and GLS estimators over a smaller model class $\Theta_{2L,w}$, as in (11) and (19). In particular, as stated in Theorem 4, the `GLS-2L` line, corresponding to $g_L(2w)$, gives a lower bound of the minimax risk over $\Theta_{L,w}$. All the curves are plotted over a big range $0 \le w \le 0.8$. For example, a model class $\Theta_{L,w}$ with $w = 0.8$ would include a model $\boldsymbol{\theta}$ which changes mean at a level of 2 standard deviation every 5 data points, or at a level of 4 standard deviation every 20 data points. In general, a large ratio $W(\boldsymbol{\theta})/(n\sigma^2)$ indicates that either the magnitude of mean changes is large or the mean changes frequently. In the former scenario, we may detect the obvious change points first and reduce the total variation $W(\boldsymbol{\theta})$, then estimate the variance, which facilitate the detection of subtle change points. In the second scenario, it would be difficult to identify all the change points simultaneously even if we know the true variance. Therefore, it is reasonable to consider variance estimation for a class $\Theta_{L,w}$ with small or moderate w. Finally, we conclude that the OLS estimator $\hat{\alpha}_K$, defined in (10) and considered in Section 2.3, gives a simple and good solution to the variance estimation problem, especially for a model class $\Theta_{L,w}$ where w is not too big. We call $\hat{\alpha}_K$ the equivariant variance estimator (EVE), whose numerical performance will be presented next.

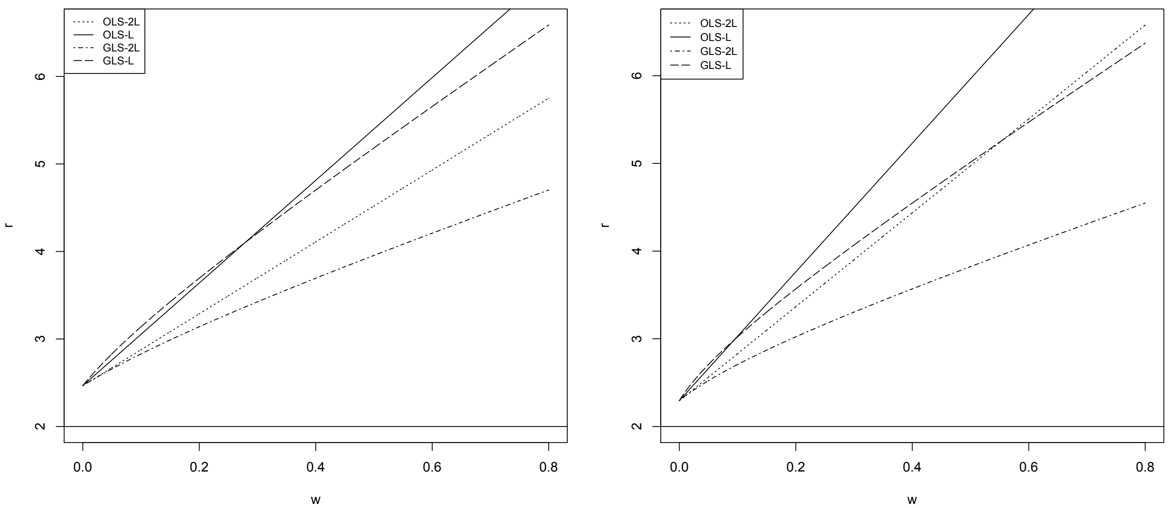

Figure 1: Lower (`GLS-2L`) and upper (`OLS-L`, `GLS-L`) bounds of the minimax risk $r_{L,w}$ with respect to w. The left and right panels correspond to $L = 10$ and $L = 15$ respectively.

3 Numerical studies

3.1 Simulated data examples

We illustrate the performance of our method using simulated data. We consider three error distributions, standard Gaussian distribution $\varepsilon_i \sim N(0, 1)$, a scaled t-distribution $\varepsilon_i \sim \sqrt{\frac{2}{3}}\, t_6$, and a translated exponential distribution $\varepsilon_i \sim Exp(1) - 1$, all of which have mean zero and variance one, with $\kappa_4 = 3,\ 6,\ 9$, respectively. Note that the exponential distribution is non-symmetric with a nonzero third moment. We fix $n = 1,000$ and consider three mean structures. Specifically, we consider a null model without any change point in scenario 1, a sparse mean model with few change points in scenario 2, and a model with frequent changes in scenario 3, as detailed below.

Scenario 1: $\boldsymbol{\theta} = \mathbf{0}$.

Scenario 2: $\theta_i = 1$ when $100m + 1 \le i \le 100m + 10$, $m \in \{1, 2, ..., 6\}$; $\theta_i = -3$ when $801 \le i \le 820$, and $\theta_i = 0$ otherwise.

Scenario 3: $\theta_i = 1$ when $20m + 1 \le i \le 20m + 10$, $m \in \{0, 1, ..., 49\}$, and $\theta_i = -1$

otherwise.

We report the simulation results for different methods by the average values and standard errors over 500 independent replicates for each scenario. Because practically it is more often to use standard deviation σ rather than the variance σ^2 in inference, we take square root to all variance estimators and report the results on standard deviation estimation. In total, there are 9 scenarios (3 mean scenarios $\times$ 3 error distributions), labeled by S1-G, S1-T,..., S3-E in tables. For example, S1-G indicates Scenario 1 with Gaussian error.

To show the sensitivity to the choice of K of our method, we compare the performance of the EVE for $K = 5$, 10, 15, and 20 in Table 1. For the null model (Scenario 1), larger K leads to a better performance, as affirmed in Theorem 1. Nevertheless, the improvement using an K larger than 10 is marginal. In contrast, in Scenario 3 when there are many change points, there is an upward bias when K is larger than 10. In Scenario 2, a larger K leads to slightly larger bias but smaller variance. In this case, our method is not sensitive to the choice of K. We observe that the standard errors of all estimators for the exponential and t distributions are larger than the Gaussian distribution because their fourth moments are larger. This is consistent with Theorem 1.

Table 1: Average values of estimators with standard errors in parenthesis over 500 replicates.

	K=5	K=10	K=15	K=20	tuned	Oracle
S1-G	0.999(0.029)	1.000(0.026)	1.000(0.025)	1.000(0.024)	0.999(0.028)	1.000(0.023)
S1-T	0.999(0.039)	0.999(0.037)	0.999(0.036)	0.999(0.035)	0.999(0.038)	1.000(0.034)
S1-E	0.998(0.048)	0.998(0.046)	0.998(0.046)	0.998(0.046)	0.998(0.047)	0.998(0.046)
S2-G	1.000(0.029)	1.000(0.026)	1.004(0.026)	1.009(0.025)	1.000(0.028)	1.000(0.023)
S2-T	0.999(0.039)	0.999(0.037)	1.003(0.036)	1.008(0.035)	1.000(0.038)	1.000(0.034)
S2-E	0.998(0.049)	0.998(0.046)	1.003(0.046)	1.007(0.046)	0.999(0.047)	0.998(0.046)
S3-G	1.000(0.034)	1.000(0.030)	1.253(0.026)	1.468(0.031)	1.001(0.030)	1.000(0.023)
S3-T	0.999(0.043)	0.999(0.040)	1.254(0.033)	1.469(0.035)	1.000(0.041)	1.000(0.034)
S3-E	0.998(0.052)	0.998(0.049)	1.252(0.041)	1.467(0.041)	0.999(0.049)	0.998(0.046)

We see that the choice of K is crucial when the mean variation is large as in scenario 3. We develop a simple method to tune K. Given a range of K, say $K_{\min} = 5 \leq K \leq K_{\max} = 20$, we calculate Y_1,..., $Y_{K_{\max}+1}$ and use Y_1,..., Y_K to predict Y_{K+1} based on the linear model (8). We calculate a score defined by $SC(K) = |\hat{Y}_{K+1} - Y_{K+1}|/\hat{\sigma}_e$, where $\hat{\sigma}_e$ is estimated based on the RSS. An K is selected by

$$\hat{K} = \underset{\{K_{\min} \leq K \leq K_{\max}\}}{\text{argmax}} SC(K).$$

This tuning process chooses $K = 10$ with high probability (96.8%, 96.0%, and 95.2%) in S3-G, S3-T, and S3-E, respectively. In the first two scenarios, the choice of K is not crucial. Overall, the tuning method works well. In practice, we suggest that one should plot the first few Y_k's, e.g., Y_1,...,Y_{20}, and see whether there is an obvious change on the slope. If not, $K = 10$ seems a safe choice and can be used as a rule of thumb. Otherwise, the tuning method can be used.

We compare the variance estimators introduced in Section 2.1 with the EVE. The simulation results are summarized in Table 2. The regression based estimators EVE and MS with $K = 10$ are labeled by MS(K=10) and EVE(K=10), respectively. The EVE with tuned K is labeled by EVE. The estimators defined in (3), (4), (5), and the oracle estimator (13) are labeled by MAD, DK, Rice, and Oracle, respectively. We also report the relative efficiency of each estimator to the oracle one (13)in Table 3. It is clear from the results that the regression based methods MS and EVE perform best among all except the oracle one in all scenarios. The relative efficiency of the EVE and MS to the oracle is constantly low. The tuning method works well. All of the MAD, DK and Rice estimators are seriously biased in some scenarios. In general, MAD and DK estimators tend to be biased upward when the mean structure is complex, e.g., in S2-G and S3-G, and to be biased downward when the noise distribution is t or exponential, e.g., in S1-T and S1-E. The Rice estimator is immune to the error distribution, but is biased upward when the mean structure is complex, e.g., in Scenario 3. As illustrated in our theoretical result, the EVE and MS estimator perform similarly. The EVE is slightly better when $\theta_1 = \theta_n$, and the MS estimator is better in Scenario 3 when $|\theta_1 - \theta_n|$ is large.

Table 2: Average values of estimators with standard errors in parenthesis over 500 replicates.

	EVE	EVE(K=10)	MS(K=10)	MAD	DK	Rice	Oracle
S1-G	0.999(0.027)	1.000(0.026)	1.000(0.026)	1.001(0.040)	1.001(0.041)	0.999(0.028)	1.000(0.023)
S1-T	0.999(0.038)	0.999(0.037)	0.999(0.037)	0.867(0.036)	0.916(0.038)	0.999(0.039)	1.000(0.034)
S1-E	0.998(0.047)	0.998(0.046)	0.998(0.046)	0.714(0.033)	0.727(0.038)	0.998(0.048)	0.998(0.046)
S2-G	1.001(0.028)	1.000(0.026)	1.000(0.026)	1.049(0.042)	1.005(0.041)	1.007(0.028)	1.000(0.023)
S2-T	1.000(0.038)	0.999(0.037)	0.999(0.037)	0.921(0.036)	0.921(0.039)	1.006(0.039)	1.000(0.034)
S2-E	1.000(0.047)	0.998(0.046)	0.998(0.046)	0.781(0.034)	0.735(0.038)	1.005(0.048)	0.998(0.046)
S3-G	1.001(0.030)	1.000(0.030)	1.000(0.030)	1.557(0.052)	1.071(0.043)	1.094(0.028)	1.000(0.023)
S3-T	1.000(0.041)	0.999(0.040)	0.999(0.040)	1.556(0.046)	0.994(0.041)	1.094(0.038)	1.000(0.034)
S3-E	0.999(0.049)	0.998(0.049)	0.998(0.049)	1.575(0.066)	0.821(0.043)	1.093(0.046)	0.998(0.046)

Table 3: Estimated relative efficiency of each method to the oracle estimator based on 500 replicates.

	EVE	EVE(K=10)	MS(K=10)	MAD	DK	Rice
S1-G	1.39	1.21	1.22	2.87	3.06	1.44
S1-T	1.21	1.13	1.13	15.85	7.25	1.30
S1-E	1.06	1.02	1.03	39.84	36.45	1.12
S2-G	1.47	1.25	1.25	7.74	3.12	1.52
S2-T	1.24	1.14	1.14	6.40	6.54	1.33
S2-E	1.05	1.02	1.03	23.54	34.54	1.12
S3-G	1.70	1.61	1.60	575.87	12.72	17.63
S3-T	1.39	1.33	1.32	262.07	1.43	8.75
S3-E	1.17	1.14	1.14	161.15	16.26	5.16

3.2 Error from real data

In real applications, the noise distributions are unknown and often far from being Gaussian, which makes the variance estimation even more challenging. To illustrate the performances of different variance estimators, we use a SNP genotying data set produced by Illumina 550K platform, available in `http://penncnv.openbioinformatics.org/`. The log R ratio (LRR) sequence of the data set has mean zero except a few short segments, called copy number variations (CNVs). We pick the LRR sequence of Chromosome 11 of the subject father with 27272 data points. As the CNVs are few and short in this data set, we treat all data points as random noise. We standardize the data to have mean zero and variance one. We use the same mean structures as before and draw the errors randomly from the standardized sequence. The results are shown in Tables 4 and 5. We observe that the performance of the EVE and MS estimator is similar to the oracle estimator and better than other estimators.

Table 4: Average values of estimators with standard errors in parenthesis over 500 replicates.

	EVE	EVE(K=10)	MS(K=10)	MAD	DK	Rice	Oracle
S1	1.000(0.034)	1.000(0.033)	1.000(0.033)	0.886(0.034)	0.930(0.041)	1.001(0.036)	1.000(0.031)
S2	1.002(0.035)	1.001(0.034)	1.001(0.034)	0.939(0.033)	0.935(0.041)	1.008(0.036)	1.000(0.031)
S3	1.001(0.036)	1.001(0.035)	1.001(0.035)	1.555(0.046)	1.005(0.042)	1.096(0.035)	1.000(0.031)

Table 5: Estimated relative efficiency of each method to the oracle estimator based on 500 replicates.

	EVE	EVE(K=10)	MS(K=10)	MAD	DK	Rice
S1	1.19	1.11	1.12	14.45	6.72	1.32
S2	1.28	1.17	1.18	4.95	5.99	1.41
S3	1.31	1.27	1.27	314.77	1.82	10.67

3.3 Labor productivity

This example is motivated by Hansen (2001). We consider the variance estimation of the U.S. labor productivity of major sectors: manufacturing/durable (DUR), manufacturing/nondurable (NDUR), business (BUS), nonfarm business (NFBUS), and nonfinancial corporations (NFC). All the series range from 1987 Q1 to 2019 Q4. The data is obtained from U.S. Bureau of Labor Statistics (`https://www.bls.gov/lpc/`). We report the estimated standard deviations of the quarterly growth rates in percentages in Table 6. Besides the estimators introduced earlier, the sample standard deviation (SD) is also included for comparison. We find that SD might overestimate σ as it ignores the potential change points. DK often underestimates σ possibly due to non-Gaussian noise distribution. The MAD estimator seems to be unstable, with larger biases. The Rice estimator is similar to the proposed EVE estimator (with data-driven choice of K), which provides most reliable estimates. These patterns are also consistent with the simulation.

	EVE	MAD	DK	Rice	SD
DUR	3.61	5.49	3.40	3.80	5.20
NDUR	3.49	3.71	3.30	3.39	3.81
BUS	2.49	2.37	2.41	2.50	2.59
NFBUS	2.54	2.37	2.62	2.55	2.60
NFC	3.60	3.11	3.40	3.76	3.62

Table 6: Variance estimation for the US labor productivity indices.

4 Discussion

The detection or segmentation procedures for change-point models often require the prior knowledge of the variance, and it is a common practice to estimate the variance as the first step of the analysis. We find that the regression based quadratic variance estimators, such as MS estimator (Müller & Stadtmüller, 1999) and the EVE proposed in this work, perform better than other popular approaches. We show the ℓ_2 risk explicitly for both the EVE and MS estimator. These two estimators are based on leg-k Rice estimators S_k and a circular version T_k, respectively. Practically, the EVE is slightly preferred when the noises are skewed as it does not require vanished third moment. Theoretically, it is easier to work with T_k because of the symmetric set-up, and all unbiased equivariance quadratic variance estimators are linear combinations of T_k, as shown in Theorem 2. It is much more difficult to characterize all unbiased quadratic variance estimators, which are not necessarily linear combinations of S_k. As a conclusion, we recommend both the EVE and MS estimator for variance estimation in change-point analysis.

As a next step, it is natural to consider the change-point model where the observations are serially correlated. In this time series context, not only the variance, but also the long run variance is of critical importance in change point analysis. It is desirable to construct easy-to-do yet accurate long run variance estimators as well. The framework and idea introduced in this paper will be indispensable for this direction of future research.

Funding

The authors are partially supported by National Science Foundation, Simons Foundation, and University of Arizona faculty seed grant.

References

Altissimo, F. & Corradi, V. (2003). Strong rules for detecting the number of breaks in a time series. *Journal of Econometrics* **117**, 207–244.

Arias-Castro, E., Donoho, D. L. & Huo, X. (2005). Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Transactions on Information Theory* **51**, 2402–2425.

Arlot, S., Celisse, A. & Harchaoui, Z. (2019). A kernel multiple change-point algorithm via model selection. *Journal of machine learning research* **20**.

Bai, J. & Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica* , 47–78.

Banerjee, A. & Urga, G. (2005). Modelling structural breaks, long memory and stock market volatility: an overview. *Journal of Econometrics* **129**, 1–34.

Chen, J. & Gupta, A. K. (2012). *Parametric statistical change point analysis: with applications to genetics, medicine, and finance.* Birkhäuser.

Davies, P. L. & Kovac, A. (2001). Local extremes, runs, strings and multiresolution. *Annals of Statistics* , 1–48.

Dette, H., Munk, A. & Wagner, T. (1998). Estimating the variance in nonparametric regressionwhat is a reasonable choice? *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**, 751–764.

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

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics* **42**, 2243–2281.

Fuhrmann, P. A. (2011). *A polynomial approach to linear algebra.* Springer Science & Business Media.

Gasser, T., Sroka, L. & Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73**, 625–633.

Hall, P., Kay, J. & Titterinton, D. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77**, 521–528.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association* **69**, 383–393.

Hansen, B. E. (2001). The new econometrics of structural change: dating breaks in us labour productivity. *Journal of Economic perspectives* **15**, 117–128.

Juhl, T. & Xiao, Z. (2009). Tests for changing mean with monotonic power. *Journal of Econometrics* **148**, 14–24.

Killick, R., Fearnhead, P. & Eckley, I. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association* **107**, 1590–1598.

Lavielle, M. (2005). Using penalized contrasts for the change-point problem. *Signal processing* **85**, 1501–1510.

Levine, M. & Tecuapetla-Gomez, I. (2019). Acf estimation via difference schemes for a semiparametric model with m-dependent errors. *arXiv preprint arXiv:1905.04578* .

McElroy, F. (1967). A necessary and sufficient condition that ordinary least-squares estimators be best linear unbiased. *Journal of the American Statistical Association* **62**, 1302–1304.

Müller, H.-G. & Stadtmüller, U. (1987). Estimation of heteroscedasticity in regression analysis. *The Annals of Statistics* **15**, 610–625.

Müller, H.-G. & Stadtmüller, U. (1999). Discontinuous versus smooth regression. *The Annals of Statistics* **27**, 299–337.

Niu, Y. S., Hao, N. & Zhang, H. (2016). Multiple change-point detection: A selective overview. *Statist. Sci.* **31**, 611–623.

Niu, Y. S. & Zhang, H. (2012). The screening and ranking algorithm to detect DNA copy number variations. *The Annals of Applied Statistics* **6**, 1306–1326.

Oka, T. & Qu, Z. (2011). Estimating structural changes in regression quantiles. *Journal of Econometrics* **162**, 248–267.

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

Perron, P. (2006). Dealing with structural breaks. *Palgrave handbook of econometrics* **1**, 278–352.

Pique-Regi, R., Monso-Varona, J., Ortega, A., Seeger, R. C., Triche, T. J. & Asgharzadeh, S. (2008). Sparse representation and bayesian detection of genome copy number alterations from microarray data. *Bioinformatics* **24**, 309–318.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *The Annals of Statistics* , 1215–1230.

Tecuapetla-Gómez, I. & Munk, A. (2017). Autocovariance estimation in regression with a discontinuous signal and m-dependent errors: A difference-based approach. *Scandinavian Journal of Statistics* **44**, 346–368.

Tong, T., Ma, Y. & Wang, Y. (2013). Optimal variance estimation without estimating the mean function. *Bernoulli* **19**, 1839–1854.

Truong, C., Oudre, L. & Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing* **167**, 107299.

Zhang, N. R. & Siegmund, D. O. (2007). A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* **63**, 22–32.

Supplementary Material to Equivariant Variance Estimation for Multiple Change-point Model

A Proof of Theorem 1

We start this section with a lemma which facilitates our proof of Propositions 1 and 2, and conclude with the proof of Theorem 1.

Lemma 1 *Let $i \in [n]$, $j \in [J]$ and $\theta_i = \mu_j$ for a model $\boldsymbol{\theta}$. For $k \leq L(\boldsymbol{\theta})$, $\theta_i - \theta_{i+k}$ is either 0 or $\mu_j - \mu_{j+1}$. For $k \leq L(\boldsymbol{\theta})/2$, $(\theta_i - \theta_{i+k})(\theta_{i+k} - \theta_{i+2k}) = 0$.*

Proof of Lemma 1. For $k \leq L(\boldsymbol{\theta})$, there is at most one change point between i and $i+k$. Therefore,

$$\theta_i - \theta_{i+k} = \begin{cases} \mu_j - \mu_{j+1}, & \text{when } \tau_j < i \leq \tau_{j+1} < i+k; \\ 0, & \text{when } \tau_j < i < i+k \leq \tau_{j+1}. \end{cases}$$

For $k \leq L(\boldsymbol{\theta})/2$, there is at most one change point between i and $i+2k$. At least one of $\theta_i - \theta_{i+k}$ and $\theta_{i+k} - \theta_{i+2k}$ is zero, so is the product.

Proof of Propositions 1 and 2. Within this proof, i, i', $i'' \in [n]$ are three different indices, and j, j', $j'' \in [J]$ such that $\theta_i = \mu_j$, $\theta_{i'} = \mu_{j'}$ and $\theta_{i''} = \mu_{j''}$.

Under Condition 1, it is straightforward to obtain

$$\mathrm{E}(\varepsilon_i - \varepsilon_{i'})^2 = 2\sigma^2, \tag{20}$$

$$\mathrm{E}(X_i - X_{i'})^2 = (\theta_i - \theta_{i'})^2 + 2\sigma^2 = (\mu_j - \mu_{j'})^2 + 2\sigma^2. \tag{21}$$

It follows Lemma 1

$$\sum_{i=1}^{n}(\theta_i - \theta_{i+k})^2 = k\sum_{j=1}^{J}(\mu_j - \mu_{j+1})^2.$$

So we have

$$\begin{aligned}
\mathrm{E}T_k &= \sum_{i=1}^{n}(X_i - X_{i+k})^2 \\
&= \sum_{i=1}^{n}(\theta_i - \theta_{i+k})^2 + 2\sigma^2 \\
&= k\sum_{j=1}^{J}(\mu_j - \mu_{j+1})^2 + 2n\sigma^2 \\
&= 2n\sigma^2 + kW(\boldsymbol{\theta}).
\end{aligned}$$

Similarly, we have

$$\mathrm{E}S_k = 2n\sigma^2 + k\left(V(\boldsymbol{\theta}) - 2\sigma^2\right).$$

For the covariance part, we start with

$$\begin{aligned}
\mathrm{Var}(\varepsilon_i - \varepsilon_{i'})^2 &= \mathrm{E}(\varepsilon_i - \varepsilon_{i'})^4 - \left[\mathrm{E}(\varepsilon_i - \varepsilon_{i'})^2\right]^2 \\
&= \mathrm{E}(\varepsilon_i^4 - 4\varepsilon_i^3\varepsilon_{i'} + 6\varepsilon_i^2\varepsilon_{i'}^2 - 4\varepsilon_i\varepsilon_{i'}^3 + \varepsilon_{i'}^4) - (2\sigma^2)^2 \\
&= 2\kappa_4\sigma^4 + 6\sigma^4 - 4\sigma^4 \\
&= 2(\kappa_4 + 1)\sigma^4,
\end{aligned}$$

$$
\begin{aligned}
\mathrm{Cov}[(\varepsilon_i - \varepsilon_{i'})^2, (\varepsilon_{i'} - \varepsilon_{i''})^2] &= \mathrm{E}(\varepsilon_i - \varepsilon_{i'})^2(\varepsilon_{i'} - \varepsilon_{i''})^2 - \mathrm{E}(\varepsilon_i - \varepsilon_{i'})^2\mathrm{E}(\varepsilon_{i'} - \varepsilon_{i''})^2 \\
&= \mathrm{E}(\varepsilon_i^2\varepsilon_{i'}^2 + \varepsilon_i^2\varepsilon_{i''}^2 + \varepsilon_{i'}^4 + \varepsilon_{i'}^2\varepsilon_{i''}^2 + \text{terms with odd degrees}) - (2\sigma^2)^2 \\
&= 3\sigma^4 + \kappa_4\sigma^4 - 4\sigma^2 \\
&= (\kappa_4 - 1)\sigma^4.
\end{aligned}
$$

Recall our convention that $\theta_i = \mu_j$, $\theta_{i'} = \mu_{j'}$, $\theta_{i''} = \mu_{j''}$.

$$
\begin{aligned}
\mathrm{Var}(X_i - X_{i'})^2 &= \mathrm{Var}(\varepsilon_i - \varepsilon_{i'} + \mu_j - \mu_{j'})^2 \\
&= \mathrm{Var}\left[(\varepsilon_i - \varepsilon_{i'})^2 + 2(\varepsilon_i - \varepsilon_{i'})(\mu_j - \mu_{j'}) + (\mu_j - \mu_{j'})^2\right] \\
&= \mathrm{Var}[(\varepsilon_i - \varepsilon_{i'})^2] + 4(\mu_j - \mu_{j'})^2\mathrm{Var}(\varepsilon_i - \varepsilon_{i'}) + 2(\mu_j - \mu_{j'})\mathrm{Cov}\left[(\varepsilon_i - \varepsilon_{i'})^2, (\varepsilon_i - \varepsilon_{i'})\right] \\
&= 2(\kappa_4 + 1)\sigma^4 + 4(\mu_j - \mu_{j'})^2 2\sigma^2 + 0 \\
&= 2(\kappa_4 + 1)\sigma^4 + 8\sigma^2(\mu_j - \mu_{j'})^2.
\end{aligned}
$$

The second to last equality follows the fact $\mathrm{Cov}\left[(\varepsilon_i - \varepsilon_{i'})^2, (\varepsilon_i - \varepsilon_{i'})\right] = \mathrm{E}(\varepsilon_i - \varepsilon_{i'})^3 = 0$. It follows directly, for $k \leq L(\boldsymbol{\theta})$,

$$
\sum_{i=1}^{n} \mathrm{Var}(X_i - X_{i+k})^2 = n[2(\kappa_4 + 1)\sigma^4] + k\sum_{j=1}^{J} 8\sigma^2(\mu_j - \mu_{j+1})^2. \tag{22}
$$

$$
\begin{aligned}
&\mathrm{Cov}[(X_i - X_{i'})^2, (X_{i'} - X_{i''})^2] \\
=&\mathrm{Cov}[(\varepsilon_i - \varepsilon_{i'})^2 + 2(\varepsilon_i - \varepsilon_{i'})(\mu_j - \mu_{j'}) + (\mu_j - \mu_{j'})^2, (\varepsilon_{i'} - \varepsilon_{i''})^2 + 2(\varepsilon_{i'} - \varepsilon_{i''})(\mu_{j'} - \mu_{j''}) + (\mu_{j'} - \mu_{j''})^2] \\
=&\mathrm{Cov}[(\varepsilon_i - \varepsilon_{i'})^2 + 2(\varepsilon_i - \varepsilon_{i'})(\mu_j - \mu_{j'}), (\varepsilon_{i'} - \varepsilon_{i''})^2 + 2(\varepsilon_{i'} - \varepsilon_{i''})(\mu_{j'} - \mu_{j''})] \\
=&(\kappa_4 - 1)\sigma^4 - 2(\mu_j - 2\mu_{j'} + \mu_{j''})\mathrm{E}\varepsilon_{i'}^3 + 4(\mu_j - \mu_{j'})(\mu_{j'} - \mu_{j''})\mathrm{Cov}[\varepsilon_i - \varepsilon_{i'}, \varepsilon_{i'} - \varepsilon_{i''}] \\
=&(\kappa_4 - 1)\sigma^4 - 2(\mu_j - 2\mu_{j'} + \mu_{j''})\mathrm{E}\varepsilon_{i'}^3 - 4(\mu_j - \mu_{j'})(\mu_{j'} - \mu_{j''})\sigma^2 \\
=&(\kappa_4 - 1)\sigma^4 - 2(\theta_i - 2\theta_{i'} + \theta_{i''})\mathrm{E}\varepsilon_{i'}^3 - 4(\theta_i - \theta_{i'})(\theta_{i'} - \theta_{i''})\sigma^2.
\end{aligned}
$$

As we will see in the next a few lines, the second summand above involving the third moment will be canceled out in calculating the covariance structure of T_k's because of equivariance. For S_k's, we will need an additional condition $\mathrm{E}\varepsilon_{i'}^3 = 0$ in order to get a neat formula.

It follows last equation that, for $k \le L(\boldsymbol{\theta})/2$,

$$
\begin{aligned}
&\sum_{1\le i,i'\le n, i\ne i'} \mathrm{Cov}[(X_i - X_{i+k})^2, (X_{i'} - X_{i'+k})^2] \\
=&2\sum_{i=1}^n \mathrm{Cov}[(X_i - X_{i+k})^2, (X_{i+k} - X_{i+2k})^2] \\
=&2\sum_{i=1}^n (\kappa_4 - 1)\sigma^4 - 2(\theta_i - 2\theta_{i+k} + \theta_{i+2k})\mathrm{E}\varepsilon_{i+k}^3 - 4(\theta_i - \theta_{i+k})(\theta_{i+k} - \theta_{i+2k})\sigma^2 \\
=&2n(\kappa_4 - 1)\sigma^4,
\end{aligned}
$$

where the last equality is implied by two facts,

$$
\sum_{i=1}^n \theta_i - 2\theta_{i+k} + \theta_{i+2k} = 0 \tag{23}
$$

and

$$
(\theta_i - \theta_{i+k})(\theta_{i+k} - \theta_{i+2k}) = 0. \tag{24}
$$

In particular, (23) holds because of the equivariant formulation of T_k; (24) follows Lemma 1. To summarize, we have

$$\sum_{1\le i,i'\le n, i\neq i'} \mathrm{Cov}[(X_i - X_{i+k})^2, (X_{i'} - X_{i'+k})^2] = 2n(\kappa_4 - 1)\sigma^4. \tag{25}$$

For $k \le L(\boldsymbol{\theta})/2$, by (22) and (25), we have

$$\begin{aligned}
\mathrm{Var}(T_k) &= \mathrm{Var} \sum_{i=1}^{n} (X_i - X_{i+k})^2 \\
&= \sum_{i=1}^{n} \mathrm{Var}(X_i - X_{i+k})^2 + \sum_{i\neq i'} \mathrm{Cov}[(X_i - X_{i+k})^2, (X_{i'} - X_{i'+k})^2] \\
&= n[2(\kappa_4 + 1)\sigma^4] + k \sum_{j=1}^{J} 8\sigma^2 (\mu_j - \mu_{j+1})^2 + 2n(\kappa_4 - 1)\sigma^4 \\
&= 4n\kappa_4\sigma^4 + 8k\sigma^2 \sum_{j=1}^{J} (\mu_j - \mu_{j+1})^2 \\
&= 4n\kappa_4\sigma^4 + 8k\sigma^2 W(\boldsymbol{\theta})
\end{aligned}$$

For $k < h \le L(\boldsymbol{\theta})/2$,

$$\begin{aligned}
\mathrm{Cov}(T_k, T_h) &= \mathrm{Cov}\left(\sum_{i=1}^{n} (X_i - X_{i+k})^2, \sum_{i=1}^{n} (X_i - X_{i+h})^2 \right) \\
&= \sum_{i=1}^{n} \sum_{i'=1}^{n} \mathrm{Cov}\left((X_i - X_{i+k})^2, (X_{i'} - X_{i'+h})^2 \right),
\end{aligned}$$

where the summands are not zero only when $i = i'$, $i = i' + h$, $i + k = i'$ or $i + k = i' + h$. For the case $i = i'$, we have

$$
\begin{aligned}
&\sum_{i=1}^{n} \mathrm{Cov}\left((X_i - X_{i+k})^2, (X_i - X_{i+h})^2\right) \\
=&\sum_{i=1}^{n} \left((\kappa_4 - 1)\sigma^4 - 2(\theta_{i+k} - 2\theta_i + \theta_{i+h})\mathrm{E}\varepsilon_i^3 - 4(\theta_{i+k} - \theta_i)(\theta_i - \theta_{i+h})\sigma^2\right) \\
=&n(\kappa_4 - 1)\sigma^4 + 0 + \sum_{i=1}^{n} 4(\theta_i - \theta_{i+k})(\theta_i - \theta_{i+h})\sigma^2,
\end{aligned}
$$

where the last summand is not zero only when $\tau_j < i \leq \tau_{j+1} < i + k < i + h < \tau_{j+2}$. So it equals to

$$
4k\sigma^2 \sum_{j=1}^{J} (\mu_j - \mu_{j+1})^2 = 4k\sigma^2 W(\boldsymbol{\theta}).
$$

It is straightforward to verify that the sum is the same when $i + k = i' + h$, and the sum is $n(\kappa_4 - 1)\sigma^4$ when $i = i' + h$ or $i + k = i'$. Overall, we have

$$
\mathrm{Cov}(T_k, T_h) = 4n(\kappa_4 - 1)\sigma^4 + 8k\sigma^2 W(\boldsymbol{\theta}).
$$

The computation for covariance among S_k's is similar except that it requires vanished third moment condition as they are not equivariant.

We need the following lemma to prove Theorem 1.

Lemma 2 *Let $\vartheta^2 = W(\boldsymbol{\theta})/\sigma^2$ for simple notation. The variance of least squares estimator $(\hat{\alpha}, \hat{\beta})^\top$ is*

$$
\begin{aligned}
&\frac{\sigma^4}{n}\left[\frac{2}{K(K-1)}\begin{pmatrix} 2K+1 & -3 \\ -3 & \frac{6}{K+1} \end{pmatrix} + (\kappa_4 - 1)\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \right. \\
&\left. \qquad \frac{2\vartheta^2}{n}\frac{1}{K(K-1)}\begin{pmatrix} \frac{1}{15}(K+1)(K+2)(2K+1) & -\frac{1}{10}(K+2)(K+3) \\ -\frac{1}{10}(K+2)(K+3) & \frac{6}{5}\frac{K^2+1}{K+1} \end{pmatrix}\right].
\end{aligned}
$$

Proof of Lemma 2. Denote by a $K \times 2$ matrix $\boldsymbol{Z}$ the design matrix of the regression model

(8), i.e.,

$$\boldsymbol{Z} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & K \end{pmatrix}^{\top}. \tag{26}$$

The covariance matrix of OLS is $(\boldsymbol{Z}^{\top}\boldsymbol{Z})^{-1}\boldsymbol{Z}^{\top}\boldsymbol{\Sigma}\boldsymbol{Z}(\boldsymbol{Z}^{\top}\boldsymbol{Z})^{-1}$.

$$\begin{aligned}\boldsymbol{Z}^{\top}\boldsymbol{\Sigma}\boldsymbol{Z} &= \frac{\sigma^4}{n}\boldsymbol{Z}^{\top}(\boldsymbol{I} + (\kappa_4 - 1)\mathbf{1}\mathbf{1}^{\top} + \frac{2\vartheta^2}{n}\boldsymbol{H})\boldsymbol{Z} \\ &= \frac{\sigma^4}{n}\left[\boldsymbol{Z}^{\top}\boldsymbol{Z} + (\kappa_4 - 1)\boldsymbol{Z}^{\top}\mathbf{1}\mathbf{1}^{\top}\boldsymbol{Z} + \frac{2\vartheta^2}{n}\boldsymbol{Z}^{\top}\boldsymbol{H}\boldsymbol{Z}\right].\end{aligned}$$

$$\begin{aligned}&(\boldsymbol{Z}^{\top}\boldsymbol{Z})^{-1}\boldsymbol{Z}^{\top}\boldsymbol{\Sigma}\boldsymbol{Z}(\boldsymbol{Z}^{\top}\boldsymbol{Z})^{-1} \\ =&\frac{\sigma^4}{n}\left[(\boldsymbol{Z}^{\top}\boldsymbol{Z})^{-1} + (\kappa_4 - 1)(\boldsymbol{Z}^{\top}\boldsymbol{Z})^{-1}\boldsymbol{Z}^{\top}\mathbf{1}\mathbf{1}^{\top}\boldsymbol{Z}(\boldsymbol{Z}^{\top}\boldsymbol{Z})^{-1} + \frac{2\vartheta^2}{n}(\boldsymbol{Z}^{\top}\boldsymbol{Z})^{-1}\boldsymbol{Z}^{\top}\boldsymbol{H}\boldsymbol{Z}(\boldsymbol{Z}^{\top}\boldsymbol{Z})^{-1}\right] \\ =&\frac{\sigma^4}{n}\left[S_1 + S_2 + S_3\right].\end{aligned}$$

It is straightforward to calculate

$$\boldsymbol{Z}^{\top}\boldsymbol{Z} = \begin{pmatrix} K & \frac{1}{2}K(K+1) \\ \frac{1}{2}K(K+1) & \frac{1}{6}K(K+1)(2K+1) \end{pmatrix}.$$

$$S_1 = (\boldsymbol{Z}^{\top}\boldsymbol{Z})^{-1} = \frac{2}{K(K-1)}\begin{pmatrix} 2K+1 & -3 \\ -3 & \frac{6}{K+1} \end{pmatrix}.$$

$$
\begin{aligned}
S_2 =&(\kappa_4 - 1)(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \mathbf{1}\mathbf{1}^\top \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \\
=&(\kappa_4 - 1)\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.
\end{aligned}
$$

The above equation follows the fact that $\mathbf{1}$ is the first column of $\boldsymbol{Z}$ and $(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \mathbf{1}$ is the first column of $(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{Z} = \boldsymbol{I}$.

To calculate S_3, we rewrite $\boldsymbol{H}$ as

$$
\boldsymbol{H} = \mathbf{1}\mathbf{1}^\top + \sum_{k=1}^{K-1} \boldsymbol{\eta}_k \boldsymbol{\eta}_k^\top = \sum_{k=1}^{K} \boldsymbol{\eta}_k \boldsymbol{\eta}_k^\top,
$$

where $\boldsymbol{\eta}_K = \mathbf{1}$, and for $k < K$, $\boldsymbol{\eta}_k$ is a vector $(0, ..., 0, 1, ..., 1)^\top$ with first $K - k$ entries 0 and last k entries 1.

$$
\begin{aligned}
S_3 =& \frac{2\vartheta^2}{n}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{H}\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1} \\
=& \frac{2\vartheta^2}{n}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\left(\sum_{k=1}^{K}\boldsymbol{\eta}_k\boldsymbol{\eta}_k^\top\right)\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1} \\
=& \frac{2\vartheta^2}{n}\sum_{k=1}^{K}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{\eta}_k\boldsymbol{\eta}_k^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1} \\
=& \frac{2\vartheta^2}{n}\sum_{k=1}^{K}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\begin{pmatrix} k \\ \frac{k(2K+1-k)}{2}\end{pmatrix}\boldsymbol{\eta}_k^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1} \\
=& \frac{2\vartheta^2}{n}\sum_{k=1}^{K}\frac{2}{K(K-1)}\begin{pmatrix} 2K+1 & -3 \\ -3 & \frac{6}{K+1}\end{pmatrix}\begin{pmatrix} k \\ \frac{k(2K+1-k)}{2}\end{pmatrix}\boldsymbol{\eta}_k^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1} \\
=& \frac{2\vartheta^2}{n}\sum_{k=1}^{K}\frac{k}{K(K-1)}\begin{pmatrix} 3k-2K-1 \\ 6\frac{K-k}{K+1}\end{pmatrix}\boldsymbol{\eta}_k^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1} \\
=& \frac{2\vartheta^2}{n}\sum_{k=1}^{K}\left(\frac{k}{K(K-1)}\right)^2\begin{pmatrix} 3k-2K-1 \\ 6\frac{K-k}{K+1}\end{pmatrix}\begin{pmatrix} 3k-2K-1 \\ 6\frac{K-k}{K+1}\end{pmatrix}^\top \\
=& \frac{2\vartheta^2}{n}\left[\sum_{k=1}^{K}\frac{k^2}{K^2(K-1)^2}\begin{pmatrix} (3k-2K-1)^2 & 6(3k-2K-1)\frac{K-k}{K+1} \\ 6(3k-2K-1)\frac{K-k}{K+1} & 36(\frac{K-k}{K+1})^2\end{pmatrix}\right]
\end{aligned}
$$

With the help of equations

$$
\begin{aligned}
\sum_{k=1}^{K}k^2(K-k) &= \frac{1}{12}K^2(K-1)(K+1), \\
\sum_{k=1}^{K}k^2(K-k)^2 &= \frac{1}{30}K(K-1)(K+1)(K^2+1),
\end{aligned}
$$

we can calculate

$$
\begin{aligned}
&\sum_{k=1}^{K} k^2[3k-2K-1]^2\\
=&\sum_{k=1}^{K} k^2[3(k-K)+K-1]^2\\
=&\sum_{k=1}^{K} k^2[9(K-k)^2-6(K-k)(K-1)+(K-1)^2]\\
=&9\sum_{k=1}^{K} k^2(K-k)^2-6(K-1)\sum_{k=1}^{K} k^2(K-k)+(K-1)^2\sum_{k=1}^{K} k^2\\
=&\frac{9}{30}K(K-1)(K+1)(K^2+1)-6(K-1)\frac{1}{12}K^2(K-1)(K+1)+(K-1)^2\frac{1}{6}K(K+1)(2K+1)\\
=&\frac{1}{15}K(K-1)(K+1)(K+2)(2K+1),
\end{aligned}
$$

$$
\begin{aligned}
&\sum_{k=1}^{K} k^2 6(3k-2K-1)\frac{K-k}{K+1}\\
=&\frac{6}{K+1}\sum_{k=1}^{K} k^2[3(k-K)+K-1](K-k)\\
=&\frac{6}{K+1}\sum_{k=1}^{K}[-3k^2(K-k)^2+(K-1)k^2(K-k)\\
=&\frac{6}{K+1}\left(-3\frac{1}{30}K(K-1)(K+1)(K^2+1)+(K-1)\frac{1}{12}K^2(K-1)(K+1)\right)\\
=&-\frac{1}{10}K(K-1)(K+2)(K+3),
\end{aligned}
$$

$$
\begin{aligned}
&\sum_{k=1}^{K} k^2 36(\frac{K-k}{K+1})^2 \\
=&\frac{36}{(K+1)^2}\sum_{k=1}^{K} k^2(K-k)^2 \\
=&\frac{36}{(K+1)^2}\frac{1}{30}K(K-1)(K+1)(K^2+1) \\
=&\frac{6}{5}\frac{K(K-1)(K^2+1)}{K+1}.
\end{aligned}
$$

Finally, we get

$$
S_3 = \frac{2\vartheta^2}{n}\frac{1}{K(K-1)}\begin{pmatrix} \frac{1}{15}(K+1)(K+2)(2K+1) & -\frac{1}{10}(K+2)(K+3) \\ -\frac{1}{10}(K+2)(K+3) & \frac{6}{5}\frac{K^2+1}{K+1} \end{pmatrix}.
$$

Taking sum of S_1, S_2 and S_3, we can get the conclusion of the lemma.

Proof of Theorem 1. The first conclusion (11) of Theorem 1 follows Lemma 2 immediately.

Now we prove (12). Denote $\boldsymbol{d}_K^\top = (d_1, \dots, d_K) = (1, 0)(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top$, i.e. $d_1, \dots, d_K$ are the coefficients of the OLS $\hat{\alpha}_K$. Define $\boldsymbol{B}_1 = \frac{1}{2n}\sum_{k=1}^{K} d_k(\boldsymbol{I} - \boldsymbol{C}_k)$, then the OLS $\hat{\alpha}_K$ can be equivalently represented as $\hat{\alpha}_K = \boldsymbol{X}^\top\boldsymbol{B}\boldsymbol{X}$, where $\boldsymbol{B} = \boldsymbol{B}_1 + \boldsymbol{B}_1^\top$. By Lemma 5, the variance of $\hat{\alpha}_K$ can be expressed as

$$
\mathrm{Var}(\hat{\alpha}_K) = \frac{\sigma^4}{n}\left(\kappa_4 - 1 + \boldsymbol{d}_K^\top\boldsymbol{d}_K\right) + 4\sigma^2\|\boldsymbol{B}\boldsymbol{\theta}\|^2 = \frac{\sigma^4}{n}\left(\kappa_4 - 1 + \frac{4K+2}{K(K-1)}\right) + 4\sigma^2\|\boldsymbol{B}\boldsymbol{\theta}\|^2.
$$

Let $\boldsymbol{U}$ be the K dimensional upper triangular matrix with 1 on and above the diagonal, and

0 below the diagonal. Let $l_j = \tau_{j+1} - \tau_j$, and define the l_j-dimensional vector

$$\boldsymbol{s}_j := \begin{pmatrix} \boldsymbol{U}\boldsymbol{d}_K \\ \boldsymbol{0} \end{pmatrix},$$

where the last $l_j - K$ entries are zero. The elements of $\boldsymbol{B}_1^\top \boldsymbol{\theta}$ at the locations $\tau_j+1, \ldots, \tau_{j+1}$ is $(\mu_j - \mu_{j-1})\boldsymbol{s}_j/(2n)$. Define the operation $\overleftarrow{\cdot}$ as arranging the rows of a matrix upside-down. In particular, $\overleftarrow{\boldsymbol{s}_j}$ is the upside-down version of the vector $\boldsymbol{s}_j$. The elements of $\boldsymbol{B}_1\boldsymbol{\theta}$ at the same locations is $(\mu_j - \mu_{j+1})\overleftarrow{\boldsymbol{s}_j}/(2n)$. Note that the supports of $\boldsymbol{s}_j$ and $\overleftarrow{\boldsymbol{s}_j}$ do not overlap if $l_j \geq 2K$, and overlap completely if $l_j = K$, so the value of the inner product $\boldsymbol{s}_j^\top \overleftarrow{\boldsymbol{s}_j}$ varies according to the segment length l_j. It can be shown that the absolute value of the inner product is maximized when $l_j = K$, and the value $\boldsymbol{s}_j^\top \overleftarrow{\boldsymbol{s}_j} = \boldsymbol{d}_K^\top \boldsymbol{U}^\top \overleftarrow{\boldsymbol{U}} \boldsymbol{d}_K < 0$ when $l_j = K$. Therefore, it holds that

$$\begin{aligned}
&\|(\mu_j - \mu_{j+1})\boldsymbol{s}_j + (\mu_j - \mu_{j-1})\overleftarrow{\boldsymbol{s}_j}\|^2 \\
&\qquad \leq (\mu_j - \mu_{j+1})^2\|\boldsymbol{s}_j\|^2 + (\mu_j - \mu_{j-1})^2\|\overleftarrow{\boldsymbol{s}_j}\|^2 + 2|(\mu_j - \mu_{j-1})(\mu_j - \mu_{j+1})\boldsymbol{d}_K^\top \boldsymbol{U}^\top \overleftarrow{\boldsymbol{U}} \boldsymbol{d}_K| \\
&\qquad \leq [(\mu_j - \mu_{j+1})^2 + (\mu_j - \mu_{j-1})^2]\boldsymbol{d}_K^\top(\boldsymbol{U}^\top \boldsymbol{U} - \boldsymbol{U}^\top \overleftarrow{\boldsymbol{U}})\boldsymbol{d}_K.
\end{aligned} \tag{27}$$

Taking the sum over all segments,

$$\|\boldsymbol{B}\boldsymbol{\theta}\|^2 \leq \frac{1}{2n^2} \cdot \boldsymbol{d}_K^\top(\boldsymbol{U}^\top\boldsymbol{U} - \boldsymbol{U}^\top\overleftarrow{\boldsymbol{U}})\boldsymbol{d}_K \cdot \sum_{j=1}^{J}(\mu_j - \mu_{j+1})^2 = \frac{W(\boldsymbol{\theta})}{2n^2} \cdot \boldsymbol{d}_K^\top \left(\boldsymbol{U}^\top\boldsymbol{U} - \boldsymbol{U}^\top\overleftarrow{\boldsymbol{U}}\right)\boldsymbol{d}_K.$$

Therefore, the variance of the OLS $\hat{\alpha}_K$ is bounded from above by

$$\frac{\sigma^4}{n} \cdot \left[\kappa_4 - 1 + \frac{4K+2}{K(K-1)} + \frac{2W(\boldsymbol{\theta})}{n\sigma^2}\boldsymbol{d}_K^\top \left(\boldsymbol{U}^\top\boldsymbol{U} - \boldsymbol{U}^\top\overleftarrow{\boldsymbol{U}}\right)\boldsymbol{d}_K\right].$$

The calculation of the quadratic term $\boldsymbol{d}_K^\top \left(\boldsymbol{U}^\top\boldsymbol{U} - \boldsymbol{U}^\top\overleftarrow{\boldsymbol{U}}\right)\boldsymbol{d}_K$ is very similar with the proof of Lemma 2, so we omit the details, and directly give the result as the upper bound in (12).

Finally, we argue that the upper bound in (12) can be achieved. Suppose in model (2),

$K = L(\boldsymbol{\theta})$, $J = n/K$ is an even number, all segments are of the same length, and the segments means μ_j have the same absolute value, but with alternating signs. Then in (27), the two inequalities become identities with $|\mu_j - \mu_{j+1}| = \sqrt{(W(\boldsymbol{\theta})/J}$, and so is the one in (12).

B Proof of Theorem 2

Let $\hat{\sigma}^2_{\boldsymbol{A}} = \boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X}$. The following Lemmas are helpful to prove Theorem 2.

Lemma 3 *$\hat{\sigma}^2_{\boldsymbol{A}}$ is equivariant if and only if $\boldsymbol{A}$ is circulant.*

Lemma 4 *Define*

$$\mathcal{I} = \{\Lambda \subset [n] : \Lambda \text{ consists of consecutive integers modulo } n\ \}$$
$$\mathcal{I}_L = \{\Lambda \in \mathcal{I} :\ L \leq |\Lambda| \leq n - L \ \text{ or } |\Lambda| = n\}$$

The variance estimate $\hat{\sigma}^2_{\boldsymbol{A}}$ is unbiased over model class Θ_L if and only if

$$\mathrm{tr}\boldsymbol{A} = 1, \quad \text{and} \quad \sum_{i,j\in\Lambda} a_{ij} = 0,\ \forall\,\Lambda \in \mathcal{I}_L.$$

Proof of Lemma 3. $\hat{\sigma}^2_{\boldsymbol{A}}$ is equivariant if and only if $\hat{\sigma}^2_{\boldsymbol{A}}(\boldsymbol{X}) = \hat{\sigma}^2_{\boldsymbol{A}}(\boldsymbol{C}_k\boldsymbol{X})$ for all $\boldsymbol{C}_k \in \mathcal{C}_n$ and $\boldsymbol{X} \in \mathbb{R}^n$, where $\boldsymbol{C}_k$ is a circulant matrix defined in Section 2.4. Directly calculation shows $\hat{\sigma}^2_{\boldsymbol{A}}(\boldsymbol{C}_k\boldsymbol{X}) = (\boldsymbol{C}_k\boldsymbol{X})^\top\boldsymbol{A}(\boldsymbol{C}_k\boldsymbol{X}) = \boldsymbol{X}^\top(\boldsymbol{C}_k^\top\boldsymbol{A}\boldsymbol{C}_k)\boldsymbol{X}$. Therefore, $\hat{\sigma}^2_{\boldsymbol{A}}$ is equivariant if and only if $\boldsymbol{A} = \boldsymbol{C}_k^\top\boldsymbol{A}\boldsymbol{C}_k$ for all $\boldsymbol{C}_k$, which implies that $\boldsymbol{A}$ is a circulant matrix by classic result in linear algebra, e.g., Theorem 5.20 in Fuhrmann (2011).

Proof of Lemma 4. It is straightforward to show

$$\mathrm{E}\hat{\sigma}^2_{\boldsymbol{A}} = \mathrm{E}(\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{\theta}^\top\boldsymbol{A}\boldsymbol{\theta} + \sigma^2\mathrm{tr}\boldsymbol{A}.$$

Therefore, $\mathrm{E}\hat{\sigma}^2_{\boldsymbol{A}} = \sigma^2$ for all $\boldsymbol{\theta} \in \Theta_L$ if and only if $\mathrm{tr}\boldsymbol{A} = 1$ and $\boldsymbol{\theta}^\top \boldsymbol{A}\boldsymbol{\theta} = 0$ for all $\boldsymbol{\theta} \in \Theta_L$.

Now we show that $\boldsymbol{\theta}^\top \boldsymbol{A}\boldsymbol{\theta} = 0$ for all $\boldsymbol{\theta} \in \Theta_L$ if and only if $\sum_{i,j\in\Lambda} a_{ij} = 0$, for all $\Lambda \in \mathcal{I}_L$. Let $\mathbf{1}_\Lambda \in \mathbb{R}^n$ be a vector with entries equal to 1 in index set Λ, and equal to 0 otherwise. Note that $\mathbf{1}_\Lambda \in \Theta_L$ when $\Lambda \in \mathcal{I}_L$, and $\mathbf{1}_\Lambda^\top \boldsymbol{A}\mathbf{1}_\Lambda = \sum_{i,j\in\Lambda} a_{ij}$. This implies the "only if" part.

For the other direction, we first show that $\sum_{i,j\in\Lambda} a_{ij} = 0$ for all $\Lambda \in \mathcal{I}_L$ implies two facts: $a_{ij} = 0$ when $L < |i-j| < n-L$; $\sum_{i\in\Lambda; j\in\Lambda'} a_{ij} = 0$ for connected Λ and Λ'. Here we call that Λ, $\Lambda' \in \mathcal{I}_L$ are connected if Λ and Λ' are disjoint and $\Lambda \cup \Lambda' \in \mathcal{I}_L$.

For fact 1, let us start with showing $a_{1,L+2} = 0$. Consider four index set $\Lambda_1 = \{1, ..., L+1\}$, $\Lambda_2 = \{2, ..., L+1\}$, $\Lambda_3 = \{2, ..., L+2\}$ and $\Lambda_4 = \{1, ..., L+2\}$. Because Λ_1, Λ_2, Λ_3, $\Lambda_4 \in \mathcal{I}_L$, we have $\sum_{i,j\in\Lambda_k} a_{ij} = 0$ for $1 \le k \le 4$, which implies

$$a_{1,L+2} = a_{L+2,1} = \frac{1}{2}\left(\sum_{i,j\in\Lambda_2} a_{ij} + \sum_{i,j\in\Lambda_4} a_{ij} - \sum_{i,j\in\Lambda_1} a_{ij} - \sum_{i,j\in\Lambda_3} a_{ij}\right) = 0.$$

Similar arguments show $a_{ij} = 0$ for all pairs (i,j) with $L < |i-j| < n-L$.

Fact 2 directly follows

$$\sum_{i\in\Lambda; j\in\Lambda'} a_{ij} = \frac{1}{2}\left(\sum_{i,j\in\Lambda\cup\Lambda'} a_{ij} - \sum_{i,j\in\Lambda} a_{ij} - \sum_{i,j\in\Lambda'} a_{ij}\right).$$

Now for $\boldsymbol{\theta} \in \Theta_L$, either $\boldsymbol{\theta}$ is a constant vector (trivial case) or we have a sequence of disjoined index sets Λ_1,...,$\Lambda_M \in \mathcal{I}_L$ such that the pairs (Λ_1, Λ_2),..., (Λ_M, Λ_1) are connected, and $\boldsymbol{\theta} = \sum_{m=1}^{M} \mu_m 1_{\Lambda_m}$ for some μ_m's. Therefore,

$$
\begin{aligned}
&\boldsymbol{\theta}^\top \boldsymbol{A} \boldsymbol{\theta} \\
&= \left(\sum_{m=1}^{M} \mu_m 1_{\Lambda_m} \right)^\top \boldsymbol{A} \sum_{m=1}^{M} \mu_m 1_{\Lambda_m} \\
&= \sum_{m=1}^{M} \sum_{t=1}^{M} \mu_m \mu_t 1_{\Lambda_m}^\top \boldsymbol{A} 1_{\Lambda_t} \\
&= \sum_{m=1}^{M} \sum_{t=1}^{M} \left(\mu_m \mu_t \sum_{i \in \Lambda_m; j \in \Lambda_t} a_{ij} \right) \\
&= 0
\end{aligned}
$$

The last equality follows the fact that $\sum_{i \in \Lambda_m; j \in \Lambda_t} a_{ij} = 0$ for all m, t. We have to show the equation for only two cases: Λ_m and Λ_t are connected, and they are not connected. The connected case follows fact 2 directly. If Λ_m and Λ_t are not connected, then any $i \in \Lambda_m$ and $j \in \Lambda_t$ satisfy $L < |i - j| < n - L$, so $a_{ij} = 0$ by fact 1.

Proof of Theorem 2. By definition we have

$$
T_k = \sum_{i=1}^{n} (X_i - X_{i+k})^2 = 2 \sum_{i=1}^{n} X_i^2 - 2 \sum_{i \neq j} X_i X_j = 2n \boldsymbol{X}^\top \boldsymbol{A}_k \boldsymbol{X},
$$

where $\boldsymbol{A}_k = \frac{1}{n} \left(\boldsymbol{I} - \frac{1}{2} \boldsymbol{C}_k - \frac{1}{2} \boldsymbol{C}_k^\top \right)$. For any estimator $\sum_{k=1}^{L} c_k Y_k \in \mathcal{Q}_L$, we can write it as $\hat{\sigma}^2_{\boldsymbol{A}} = \boldsymbol{X}^\top \boldsymbol{A} \boldsymbol{X}$ where $\boldsymbol{A} = \sum_{k=1}^{L} c_k \boldsymbol{A}_k$ with $\sum_{k=1}^{L} c_k = 1$ and $\sum_{k=1}^{L} k c_k = 0$. $\boldsymbol{A}_k$ is circulant, so is $\boldsymbol{A}$. Therefore, $\hat{\sigma}^2_{\boldsymbol{A}}$ is equivariant by Lemma 3. Moreover,

$$
\mathrm{tr} \boldsymbol{A} = \mathrm{tr} \sum_{k=1}^{L} c_k \boldsymbol{A}_k = \sum_{k=1}^{L} c_k \mathrm{tr} \boldsymbol{A}_k = \sum_{k=1}^{L} c_k = 1.
$$

It is easy to check the sum of all entries in a principal submatrix of $\boldsymbol{A}_k$ over the index set $\Lambda \in \mathcal{I}_L$ is $\frac{k}{n}$. Then for $\boldsymbol{A} = \sum_{k=1}^{L} c_k \boldsymbol{A}_k$, we have

$$\sum_{i,j\in\Lambda} a_{ij} = \sum_{k=1}^{L} c_k \frac{k}{n} = \frac{1}{n}\sum_{k=1}^{L} k c_k = 0.$$

We conclude that $\hat{\sigma}^2_{\boldsymbol{A}}$ is also unbiased by Lemma 4, and hence, all estimators in $\mathcal{Q}_L$ are equivariant and unbiased.

Now we have any unbiased and equivariant quadratic estimator $\hat{\sigma}^2_{\boldsymbol{A}} = \boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X}$ is in $\mathcal{Q}_L$. If $\hat{\sigma}^2_{\boldsymbol{A}}$ is equivariant, then $\boldsymbol{A}$ is circulant by Lemma 3. In the proof of Lemma 4, we show that $a_{ij} = 0$ for all $L < |i-j| < n-L$ if $\hat{\sigma}^2_{\boldsymbol{A}}$ is unbiased for model class Θ_L. Therefore, $\boldsymbol{A}$ is in the linear space spanned by symmetric circulant matrices $\{\boldsymbol{I}, \boldsymbol{C}_k + \boldsymbol{C}_{-k},\ k = 1, ..., L\}$. We may write $\boldsymbol{A}$ as an element in this linear space with $b_0\boldsymbol{I} + \sum_{k=1}^{L} b_k(\boldsymbol{C}_k + \boldsymbol{C}_{-k})$. By Lemma 4, we have $\sum_{1\le i,j\le n} a_{ij} = 0$, which implies $b_0 = -2\sum_{k=1}^{L} b_k$. That is, $\boldsymbol{A} = \sum_{k=1}^{L} -2b_k(\boldsymbol{I} - \frac{1}{2}(\boldsymbol{C}_k + \boldsymbol{C}_{-k}))$ that is in a subspace spanned by $\{A_1, ..., A_L\}$. Therefore, we may write $\boldsymbol{A} = \sum_{k=1}^{L} c_k \boldsymbol{A}_k$. Again by Lemma 4, unbiasedness implies $\mathrm{tr}\boldsymbol{A} = 1$ and $\sum_{1\le i,j\le L} a_{ij} = 0$, which further imply the constraints $\sum_{k=1}^{L} c_k = 1$ and $\sum_{k=1}^{L} k c_k = 0$. Thus, we give a complete description of all unbiased equivariant quadratic variance estimators.

Proof of Corollary 1. By Theorem 2, $\mathcal{Q}_2$ consists of $c_1 Y_1 + c_2 Y_2$ with $c_1 + c_2 = 1$, $c_1 + 2c_2 = 0$, which determine a unique estimator $2Y_1 - Y_2$. The upper bound for the variance directly follows formula (12) in Theorem 1 with $K = 2$.

C Proof of Theorems 3 and 4

Proof of Corollary 2. As $\hat{\alpha}_2$ is the unique element in $\mathcal{Q}_2$, it is the minimax estimator. By

Theorem 1,

$$\begin{aligned}
&\max_{(\boldsymbol{\theta},\sigma^2)\in\Theta_{2,w}} r(\hat{\alpha}_2)\\
&\le \max_{(\boldsymbol{\theta},\sigma^2)\in\Theta_{2,w}} \kappa_4 - 1 + 5 + 8\frac{W(\boldsymbol{\theta})}{n\sigma^2}\\
&= \kappa_4 + 4 + 8w.
\end{aligned}$$

In the proof of Theorem 1, we show that the minimax risk is achieved when n is a multiple of $2L = 4$.

We need a lemma before proving Theorems 3 and 4.

Lemma 5 *Assume the same conditions of Theorem 3. Write the unbiased and equivariant estimator* $\hat{\sigma}^2_{\boldsymbol{c}}$ *as* $\hat{\sigma}^2_{\boldsymbol{c}} = \boldsymbol{X}^\top \boldsymbol{A}_{\boldsymbol{c}} \boldsymbol{X}$, *where* $\boldsymbol{A}_{\boldsymbol{c}} = \sum_{k=1}^{L} c_k \boldsymbol{A}_k$ *with* $\boldsymbol{A}_k = \frac{1}{n}\left(\boldsymbol{I} - \frac{1}{2}\boldsymbol{C}_k - \frac{1}{2}\boldsymbol{C}_k^\top\right)$. *Then its risk can be expressed as*

$$r(\hat{\sigma}^2_{\boldsymbol{c}}) = \kappa_4 - 1 + \boldsymbol{c}^\top\boldsymbol{c} + \frac{4n}{\sigma^2}\|\boldsymbol{A}_{\boldsymbol{c}}\boldsymbol{\theta}\|^2.$$

Proof of Lemma 5. By Theorem 2 and its proof, we consider an estimator of the form $\hat{\sigma}^2_{\boldsymbol{c}} = \boldsymbol{X}^\top \boldsymbol{A}_{\boldsymbol{c}} \boldsymbol{X} = \sum_{k=1}^{L} c_k Y_k \in \mathcal{Q}_L$, where $\boldsymbol{A}_{\boldsymbol{c}} = \sum_{k=1}^{L} c_k \boldsymbol{A}_k$, and $\boldsymbol{A}_k = \frac{1}{n}\left(\boldsymbol{I} - \frac{1}{2}\boldsymbol{C}_k - \frac{1}{2}\boldsymbol{C}_k^\top\right)$. As $\boldsymbol{c}$ is a fixed vector in this proof, we use $\boldsymbol{A}$ to denote $\boldsymbol{A}_{\boldsymbol{c}}$ for simple notation. Note that all entries in the diagonal of $\boldsymbol{A}$ are $\frac{1}{n}$, and $\boldsymbol{A}\boldsymbol{1} = \boldsymbol{0}$.

We calculate the variance of a general estimator in $\mathcal{Q}_L$. $\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X}$ is unbiased for σ^2, so

$$\mathrm{Var}(\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X}) = \mathrm{E}[(\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X})^2] - \sigma^4. \tag{28}$$

We calculate the second moment

$$
\begin{aligned}
&\mathrm{E}[(\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X})^2] \\
=&\mathrm{E}[(\boldsymbol{\theta}^\top \boldsymbol{A}\boldsymbol{\theta} + 2\boldsymbol{\varepsilon}^\top \boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon}^\top \boldsymbol{A}\boldsymbol{\varepsilon})^2] \\
=&\mathrm{E}[(0 + 2\boldsymbol{\varepsilon}^\top \boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon}^\top \boldsymbol{A}\boldsymbol{\varepsilon})^2] \\
=&\mathrm{E}[4(\boldsymbol{\varepsilon}^\top \boldsymbol{A}\boldsymbol{\theta})^2 + (\boldsymbol{\varepsilon}^\top \boldsymbol{A}\boldsymbol{\varepsilon})^2 + 4\boldsymbol{\varepsilon}^\top \boldsymbol{A}\boldsymbol{\theta}\boldsymbol{\varepsilon}^\top \boldsymbol{A}\boldsymbol{\varepsilon}] \\
=&\mathrm{E}[4(\boldsymbol{\varepsilon}^\top \boldsymbol{A}\boldsymbol{\theta})^2 + (\boldsymbol{\varepsilon}^\top \boldsymbol{A}\boldsymbol{\varepsilon})^2], \qquad (29)
\end{aligned}
$$

where the last equation follows the fact $\mathrm{E}[4\boldsymbol{\varepsilon}^\top \boldsymbol{A}\boldsymbol{\theta}\boldsymbol{\varepsilon}^\top \boldsymbol{A}\boldsymbol{\varepsilon}] = 0$. Note that $\boldsymbol{\varepsilon}^\top \boldsymbol{A}\boldsymbol{\theta}\boldsymbol{\varepsilon}^\top \boldsymbol{A}\boldsymbol{\varepsilon}$ is a homogeneous cubic polynomial on ε_i's. Thus, all terms in this polynomial have expectation zero except the ones involving ε_i^3's. Moreover, by the fact $a_{ii} = \frac{1}{n}$, we have

$$
\mathrm{E}[4\boldsymbol{\varepsilon}^\top \boldsymbol{A}\boldsymbol{\theta}\boldsymbol{\varepsilon}^\top \boldsymbol{A}\boldsymbol{\varepsilon}] = \mathrm{E}\left[\frac{4}{n}\boldsymbol{\theta}^\top \boldsymbol{A}\boldsymbol{\varepsilon}^{\circ 3}\right] = \frac{4\mathrm{E}\varepsilon_1^3}{n}\boldsymbol{\theta}^\top \boldsymbol{A}\boldsymbol{1} = 0,
$$

where $\boldsymbol{\varepsilon}^{\circ 3}$ denotes entry-wise cube of the vector $\boldsymbol{\varepsilon}$.

Now we calculate the two summands in (29).

$$
\mathrm{E}[4(\boldsymbol{\varepsilon}^\top \boldsymbol{A}\boldsymbol{\theta})^2] = 4\mathrm{Var}[(\boldsymbol{A}\boldsymbol{\theta})^\top \boldsymbol{\varepsilon}] = 4\sigma^2(\boldsymbol{A}\boldsymbol{\theta})^\top \boldsymbol{A}\boldsymbol{\theta} = 4\sigma^2\boldsymbol{\theta}^\top \boldsymbol{A}^2\boldsymbol{\theta}. \qquad (30)
$$

$$
\begin{aligned}
&(\boldsymbol{\varepsilon}^\top \boldsymbol{A}\boldsymbol{\varepsilon})^2 \\
=&\left(\sum_{1\le i,j\le n} \varepsilon_i a_{ij}\varepsilon_j\right)^2 \\
=&\sum_{1\le i,j,i',j'\le n} \varepsilon_i\varepsilon_j\varepsilon_{i'}\varepsilon_{j'}a_{ij}a_{i'j'} \\
=&2\sum_{1\le i<j\le n} \varepsilon_i^2\varepsilon_j^2 a_{ii}a_{jj} + 4\sum_{1\le i<j\le n} \varepsilon_i^2\varepsilon_j^2 a_{ij}^2 + \sum_{i=1}^n \varepsilon_i^4 a_{ii}^2 + \cdots
\end{aligned}
$$

where the omitted part has zero expectation. Therefore,

$$
\begin{aligned}
&\mathrm{E}[(\boldsymbol{\varepsilon}^\top \boldsymbol{A}\boldsymbol{\varepsilon})^2] \\
=&2\sum_{1\leq i<j\leq n}\sigma^4 a_{ii}a_{jj} + 4\sum_{1\leq i<j\leq n}\sigma^4 a_{ij}^2 + \sum_{i=1}^n \sigma^4\kappa_4 a_{ii}^2 \\
=&\sigma^4\left(2\sum_{1\leq i<j\leq n} a_{ii}a_{jj} + 4\sum_{1\leq i<j\leq n} a_{ij}^2 + \sum_{i=1}^n \kappa_4 a_{ii}^2\right) \\
=&\sigma^4\left((\sum_{i=1}^n a_{ii})^2 - \sum_{i=1}^n a_{ii}^2 + 2\sum_{i=1}^n\sum_{j=1}^n a_{ij}^2 - 2\sum_{i=1}^n a_{ii}^2 + \sum_{i=1}^n \kappa_4 a_{ii}^2\right) \\
=&\sigma^4\left((\mathrm{tr}\boldsymbol{A})^2 + 2\mathrm{tr}(\boldsymbol{A}^2) + (\kappa_4-3)\sum_{i=1}^n a_{ii}^2\right) \\
=&\sigma^4\left(1 + 2\mathrm{tr}(\boldsymbol{A}^2) + \frac{1}{n}(\kappa_4-3)\right)
\end{aligned}
$$

Finally, we have

$$
\begin{aligned}
&\mathrm{Var}(\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X}) \\
=&4\sigma^2\boldsymbol{\theta}^\top\boldsymbol{A}^2\boldsymbol{\theta} + \sigma^4\left(1 + 2\mathrm{tr}(\boldsymbol{A}^2) + \frac{1}{n}(\kappa_4-3)\right) - \sigma^4 \\
=&4\sigma^2\boldsymbol{\theta}^\top\boldsymbol{A}^2\boldsymbol{\theta} + \sigma^4\left(2\mathrm{tr}(\boldsymbol{A}^2) + \frac{1}{n}(\kappa_4-3)\right). \qquad (31)
\end{aligned}
$$

It is easy to find $\mathrm{tr}(\boldsymbol{A}^2) = \frac{1}{n}\left(1 + \frac{1}{2}\sum_{k=1}^L c_k^2\right) = \frac{1}{n}\left(1 + \frac{1}{2}\boldsymbol{c}^\top\boldsymbol{c}\right)$ as $\boldsymbol{A}$ is circulant. By (31),

$$
\begin{aligned}
r(\hat{\sigma}_{\boldsymbol{c}}^2) &= \frac{n}{\sigma^4}\mathrm{Var}(\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X}) \\
&= \frac{4n}{\sigma^2}\|\boldsymbol{A}\boldsymbol{\theta}\|^2 + \kappa_4 - 1 + \boldsymbol{c}^\top\boldsymbol{c}.
\end{aligned}
$$

We complete the proof of the lemma.

Proof of Theorem 3. We will follow the notation of Lemma 5 and write the risk as a quadratic function of c_i's with coefficients depending on the mean $\boldsymbol{\theta}$. The only nontrivial part is $\|\boldsymbol{A}\boldsymbol{\theta}\|^2 = \boldsymbol{\theta}^\top \boldsymbol{A}^2\boldsymbol{\theta}$. Now we calculate $\boldsymbol{A}^2$.

$$
\begin{aligned}
\boldsymbol{A}^2 &= \left(\sum_{k=1}^{L} c_k \boldsymbol{A}_k\right)^2 \\
&= \left(\sum_{k=1}^{L} c_k \frac{1}{n}(\boldsymbol{I} - \frac{1}{2}\boldsymbol{C}_k - \frac{1}{2}\boldsymbol{C}_{-k})\right)^2 \\
&=\frac{1}{n^2}\left(\boldsymbol{I} - \frac{1}{2}\sum_{k=1}^{L} c_k(\boldsymbol{C}_k + \boldsymbol{C}_{-k})\right)^2 \\
&=\frac{1}{n^2}\left(\boldsymbol{I} - \sum_{k=1}^{L} c_k(\boldsymbol{C}_k + \boldsymbol{C}_{-k}) + \frac{1}{4}\left(\sum_{k=1}^{L} c_k(\boldsymbol{C}_k + \boldsymbol{C}_{-k})\right)^2\right) \\
&=\frac{1}{n^2}\left(\boldsymbol{I} - \sum_{k=1}^{L} c_k(\boldsymbol{C}_k + \boldsymbol{C}_{-k}) + \frac{1}{4}\sum_{k=1}^{L}\sum_{\ell=1}^{L} c_k c_\ell(\boldsymbol{C}_{k+\ell} + \boldsymbol{C}_{-k-\ell} + \boldsymbol{C}_{k-\ell} + \boldsymbol{C}_{\ell-k})\right). \qquad (32)
\end{aligned}
$$

Note that when $0 < k \le L$, we have

$$
\begin{aligned}
\boldsymbol{\theta}^\top \boldsymbol{C}_k \boldsymbol{\theta} &= \sum_{i=1}^{n} \theta_i \theta_{i+k} \\
&= \sum_{i=1}^{n} \frac{1}{2}[\theta_i^2 + \theta_{i+k}^2 - (\theta_i - \theta_{i+k})^2] \\
&=\|\boldsymbol{\theta}\|_2^2 - \frac{1}{2}\sum_{i=1}(\theta_i - \theta_{i+k})^2 \\
&=\|\boldsymbol{\theta}\|_2^2 - \frac{1}{2}kW(\boldsymbol{\theta}). \qquad (33)
\end{aligned}
$$

Combining (32) and (33),

$$
\begin{aligned}
&\boldsymbol{\theta}^\top \boldsymbol{A}^2\boldsymbol{\theta}\\
=&\frac{1}{n^2}\left(\|\boldsymbol{\theta}\|^2-\sum_{k=1}^L c_k 2(\boldsymbol{\theta}^\top \boldsymbol{C}_k\boldsymbol{\theta})+\frac{1}{2}\sum_{k,\ell+1}^L c_k c_\ell(\boldsymbol{\theta}^\top \boldsymbol{C}_{k+\ell}\boldsymbol{\theta}+\boldsymbol{\theta}^\top \boldsymbol{C}_{k-\ell}\boldsymbol{\theta})\right)\\
=&\frac{1}{n^2}\left(\|\boldsymbol{\theta}\|^2-\sum_{k=1}^L c_k(2\|\boldsymbol{\theta}\|^2-kW(\boldsymbol{\theta}))+\frac{1}{2}\sum_{k,\ell=1}^L c_k c_\ell(2\|\boldsymbol{\theta}\|^2-\frac{1}{2}\sum_{i=1}^n(\theta_i-\theta_{i+k+\ell})^2-\frac{1}{2}|k-\ell|W(\boldsymbol{\theta}))\right)\\
=&\frac{1}{n^2}\left(\|\boldsymbol{\theta}\|^2-\sum_{k=1}^L c_k 2\|\boldsymbol{\theta}\|^2+\sum_{k=1}^L c_k kW(\boldsymbol{\theta})+\frac{1}{2}\sum_{k,\ell=1}^L c_k c_\ell 2\|\boldsymbol{\theta}\|^2-\frac{1}{4}\sum_{k,\ell=1}^L c_k c_\ell(\sum_{i=1}^n(\theta_i-\theta_{i+k+\ell})^2+|k-\ell|W(\boldsymbol{\theta})\right.
\end{aligned}
$$

As $\sum c_k = 1$ and $\sum kc_k = 0$, the first four terms in last line are canceled, and we have

$$
\boldsymbol{\theta}^\top \boldsymbol{A}^2\boldsymbol{\theta} = -\frac{1}{4n^2}\sum_{k,\ell=1}^L c_k c_\ell\left(|k-\ell|W(\boldsymbol{\theta})+\sum_{i=1}^n(\theta_i-\theta_{i+k+\ell})^2\right). \tag{34}
$$

Note that (34) is a quadratic from of c_k's, so we can write it by $-\frac{W(\boldsymbol{\theta})}{4n^2}\boldsymbol{c}^\top \boldsymbol{G}(\boldsymbol{\theta})\boldsymbol{c}$, where $\boldsymbol{c} = (c_1, ..., c_L)^\top$, $\boldsymbol{G} = (G_{k\ell})$ with

$$
G_{k\ell} = |k-\ell| + \frac{1}{W(\boldsymbol{\theta})}\sum_{i=1}^n(\theta_i-\theta_{i+k+\ell})^2. \tag{35}
$$

Putting all terms together, we have

$$
\begin{aligned}
&\mathrm{Var}(\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X})\\
=&-\frac{W(\boldsymbol{\theta})\sigma^2}{n^2}\boldsymbol{c}^\top \boldsymbol{G}(\boldsymbol{\theta})\boldsymbol{c}+\frac{\sigma^4}{n}(\kappa_4-1+\|\boldsymbol{c}\|^2)\\
=&\frac{\sigma^4}{n}\left(\kappa_4-1+\boldsymbol{c}^\top\left(\boldsymbol{I}_L-\frac{W(\boldsymbol{\theta})}{n\sigma^2}\boldsymbol{G}(\boldsymbol{\theta})\right)\boldsymbol{c}\right).
\end{aligned} \tag{36}
$$

It follows (36) that

$$r(\hat{\sigma}^2_{\boldsymbol{c}}) = \kappa_4 - 1 + \boldsymbol{c}^\top \left(\boldsymbol{I}_L - \frac{W(\boldsymbol{\theta})}{n\sigma^2} \boldsymbol{G}(\boldsymbol{\theta}) \right) \boldsymbol{c}, \tag{37}$$

where the vector $\boldsymbol{c}$ satisfies linear constraints

$$\sum_{k=1}^{L} c_k = 1, \qquad \sum_{k=1}^{L} k c_k = 0. \tag{38}$$

Proof of Proposition 5. First of all, the set $\mathcal{Q}_L$ is all linear unbiased estimators to the intercept in model (8) with $K = L$, and the two linear constraints are sufficient and necessary conditions for a linear estimator to be unbiased. Secondly, the risk (15) is, up to a constant $\frac{\sigma^4}{n}$, the variance of a linear unbiased estimator. By Gauss-Markov theorem, the GLS estimator is the best linear unbiased estimator, and hence, the minimizer of (15). Here is a remark on the quadratic form (15). Although the quadratic form in (15) is not positive definite over $\mathbb{R}^L$, it is positive definite on the constrained linear subspace which $\boldsymbol{c}$ lies in. The positive definiteness can be seen from (34), where the left hand side is always positive and the right hand side is $-\boldsymbol{c}^\top \boldsymbol{G}\boldsymbol{c}$ up to a positive constant.

By (35), if $L(\boldsymbol{\theta}) \geq 2L$, then $G_{k\ell} = |k - \ell| + (k + \ell) = 2\max\{k, \ell\}$, which implies that $\boldsymbol{G}$ is a $L \times L$ matrix independent of $\boldsymbol{\theta}$. Therefore, the quadratic form (15) depends on only $W(\boldsymbol{\theta})/(n\sigma^2)$.

Proof of Theorem 4. In the first part of the proof, we work on the minimax risk of estimator class $\mathcal{Q}_L$ over model class $\Theta_{2L,w}$, which is a subset of $\Theta_{L,w}$. This will give a lower bound of the minimax risk.

For any estimator in $\mathcal{Q}_L$, its risk over $\Theta_{2L,w}$ is an increasing function of $W(\boldsymbol{\theta})/(n\sigma^2)$ because of two facts shown in proof of Proposition 5. First, $\boldsymbol{G}(\boldsymbol{\theta})$ is a constant matrix for $\boldsymbol{\theta} \in \Theta_{2L}$. Second, $-\boldsymbol{c}^\top \boldsymbol{G}\boldsymbol{c} > 0$ by positive definiteness. Therefore, for all estimators in $\mathcal{Q}_L$, the worst scenario (maximum risk) happens when $W(\boldsymbol{\theta})/(n\sigma^2) = w$. It is sufficient to consider models with $W(\boldsymbol{\theta})/(n\sigma^2) = w$ for minimax estimation. Obviously, the GLS estimator, denoted by $\tilde{\alpha}_{L,w}$, minimizes (15) and is the minimax estimator in this case.

Let $\boldsymbol{U}_L$ be the upper triangular matrix with one on and above the diagonal, $\boldsymbol{Z}_L$ the $L \times 2$ matrix defined in (26) with $K = L$. For any model in $\Theta_{2L,w}$ with $W(\boldsymbol{\theta})/(n\sigma^2) = w$, the covariance matrix (9) of $(Y_1, \ldots, Y_L)^\top$ is

$$\boldsymbol{\Sigma}_{L,w} := \frac{\sigma^4}{n} \left[\boldsymbol{I}_L + (\kappa_4 - 1)\boldsymbol{1}_L\boldsymbol{1}_L^\top + 2w\boldsymbol{U}_L^\top \boldsymbol{U}_L \right],$$

Write the GLS estimator $\tilde{\alpha}_{L,w}$ as $\tilde{\alpha}_{L,w} = (Y_1, \ldots, Y_L)\tilde{\boldsymbol{d}}_L$. By Proposition 5,

$$\begin{aligned} \tilde{\boldsymbol{d}}_L &:= \underset{\boldsymbol{d}^\top \boldsymbol{Z}_L = (1,0)}{\operatorname{argmin}} \ \boldsymbol{d}^\top \left[\boldsymbol{I}_L + (\kappa_4 - 1)\boldsymbol{1}_L\boldsymbol{1}_L^\top + 2w\boldsymbol{U}_L^\top \boldsymbol{U}_L \right] \boldsymbol{d} \\ &= \underset{\boldsymbol{d}^\top \boldsymbol{Z}_L = (1,0)}{\operatorname{argmin}} \ \boldsymbol{d}^\top \left[\boldsymbol{I}_L + 2w\boldsymbol{U}_L^\top \boldsymbol{U}_L \right] \boldsymbol{d}. \end{aligned}$$

Therefore, the maximum risk of $\tilde{\alpha}_{L,w}$ over $\Theta_{2L,w}$ is given by

$$g_L(2w) = \max_{(\boldsymbol{\theta},\sigma^2)\in\Theta_{2L,w}} r(\tilde{\alpha}_{L,w}) = (1,0) \left[\boldsymbol{Z}_L^\top \left(\boldsymbol{I}_L + 2w\boldsymbol{U}_L^\top \boldsymbol{U}_L \right)^{-1} \boldsymbol{Z}_L \right]^{-1} \binom{1}{0} + \kappa_4 - 1. \quad (39)$$

For notational simplicity, denote $\lambda = 2w$. We proceed to calculate the elements of the matrix $\boldsymbol{Z}_L^\top(\boldsymbol{I}_L + \lambda\boldsymbol{U}_L^\top \boldsymbol{U}_L)^{-1}\boldsymbol{Z}_L$. By the Woodbury matrix identity

$$(\boldsymbol{I}_L + \lambda\boldsymbol{U}_L\boldsymbol{U}_L^\top)^{-1} = \boldsymbol{I}_L - \lambda\boldsymbol{U}_L(\boldsymbol{I}_L + \lambda\boldsymbol{U}_L^\top\boldsymbol{U}_L)^{-1}\boldsymbol{U}_L^\top.$$

Denote $(\sigma^{k\ell})_{1\le k,\ell\le L} := (\boldsymbol{I}_L + \lambda\boldsymbol{U}_L^\top\boldsymbol{U}_L)^{-1}$. Let $\boldsymbol{e}_k$ be the L-dimensional coordinate vector whose only nonzero entry is at the location k, with value 1. Note that each of the matrices $\boldsymbol{I}_L + \lambda\boldsymbol{U}_L\boldsymbol{U}_L^\top$ and $\boldsymbol{I}_L + \lambda\boldsymbol{U}_L^\top\boldsymbol{U}_L$ can be obtained form the other by reverting its columns and rows, and hence

$$
\begin{aligned}
\sigma^{LL} = \boldsymbol{e}_1^\top (\boldsymbol{I}_L + \lambda \boldsymbol{U}_L \boldsymbol{U}_L^\top)^{-1} \boldsymbol{e}_1 &= 1 - \lambda \boldsymbol{e}_1^\top \lambda \boldsymbol{U}_L (\boldsymbol{I}_L + \lambda \boldsymbol{U}_L^\top \boldsymbol{U}_L)^{-1} \boldsymbol{U}_L^\top \boldsymbol{e}_1 \\
&= 1 - \lambda \boldsymbol{1}_L^\top (\boldsymbol{I}_L + \lambda \boldsymbol{U}_L^\top \boldsymbol{U}_L)^{-1} \boldsymbol{1}_L,
\end{aligned}
$$

which implies that

$$
\left[\boldsymbol{Z}_L^\top (\boldsymbol{I}_L + \lambda \boldsymbol{U}_L^\top \boldsymbol{U}_L)^{-1} \boldsymbol{Z}_L\right][1,1] = \frac{1 - \sigma^{LL}}{\lambda}. \tag{40}
$$

Applying the Woodbury identity twice, we have

$$
(\boldsymbol{I}_L + \lambda \boldsymbol{U}_L^\top \boldsymbol{U}_L)^{-1} = \boldsymbol{I}_L - \lambda \boldsymbol{U}_L^\top \boldsymbol{U}_L + \lambda^2 \boldsymbol{U}_L^\top \boldsymbol{U}_L (\boldsymbol{I}_L + \lambda \boldsymbol{U}_L^\top \boldsymbol{U}_L)^{-1} \boldsymbol{U}_L^\top \boldsymbol{U}_L.
$$

From the identity

$$
\begin{aligned}
\sigma^{1L} = \boldsymbol{e}_1^\top (\boldsymbol{I}_L + \lambda \boldsymbol{U}_L^\top \boldsymbol{U}_L)^{-1} \boldsymbol{e}_L &= -\lambda + \lambda^2 \boldsymbol{e}_1^\top \lambda \boldsymbol{U}_L^\top \boldsymbol{U}_L^{(} \boldsymbol{I}_L + \lambda \boldsymbol{U}_L^\top \boldsymbol{U}_L)^{-1} \boldsymbol{U}_L^\top \boldsymbol{U}_L \boldsymbol{e}_L \\
&= -\lambda + \lambda^2 \boldsymbol{1}_L^\top (\boldsymbol{I}_L + \lambda \boldsymbol{U}_L^\top \boldsymbol{U}_L)^{-1} (1, 2, \ldots, L)^\top,
\end{aligned}
$$

we have

$$
\left[\boldsymbol{Z}_L^\top (\boldsymbol{I}_L + \lambda \boldsymbol{U}_L^\top \boldsymbol{U}_L)^{-1} \boldsymbol{Z}_L\right][1,2] = \frac{\sigma^{1L} + \lambda}{\lambda^2}. \tag{41}
$$

Furthermore, from the identity

$$
\begin{aligned}
\sigma^{LL} = \boldsymbol{e}_L^\top (\boldsymbol{I}_L + \lambda \boldsymbol{U}_L^\top \boldsymbol{U}_L)^{-1} \boldsymbol{e}_L &= 1 - \lambda L + \lambda^2 \boldsymbol{e}_L^\top \lambda \boldsymbol{U}_L^\top \boldsymbol{U}_L^{(} \boldsymbol{I}_L + \lambda \boldsymbol{U}_L^\top \boldsymbol{U}_L)^{-1} \boldsymbol{U}_L^\top \boldsymbol{U}_L \boldsymbol{e}_L \\
&= 1 - \lambda L + \lambda^2 (1, \ldots, L) (\boldsymbol{I}_L + \lambda \boldsymbol{U}_L^\top \boldsymbol{U}_L)^{-1} (1, \ldots, L)^\top,
\end{aligned}
$$

we have

$$
\left[\boldsymbol{Z}_L^\top (\boldsymbol{I}_L + \lambda \boldsymbol{U}_L^\top \boldsymbol{U}_L)^{-1} \boldsymbol{Z}_L\right][2,2] = \frac{\sigma^{LL} + \lambda L - 1}{\lambda^2}. \tag{42}
$$

By Lemma 6, we know that $\sigma^{LL} = D_{L-1}/D_L$ and $\sigma^{1L} = -\lambda/D_L$. Combining (40), (41) and

(42), we see that the matrix $\boldsymbol{Z}_L^\top(\boldsymbol{I}_L+\lambda\boldsymbol{U}_L^\top\boldsymbol{U}_L)^{-1}\boldsymbol{Z}_L$ equals

$$\boldsymbol{V}_{L,\lambda} := \boldsymbol{Z}_L^\top(\boldsymbol{I}_L+\lambda\boldsymbol{U}_L^\top\boldsymbol{U}_L)^{-1}\boldsymbol{Z}_L = \begin{pmatrix} \frac{1-D_{L-1}/D_L}{\lambda} & \frac{D_L-1}{\lambda D_L} \\ \frac{D_L-1}{\lambda D_L} & \frac{D_{L-1}/D_L+\lambda L-1}{\lambda^2} \end{pmatrix}.$$

The proof of the first part of Theorem 4 is complete in view of (39).

We now prove the second part of Theorem 4, i.e., the upper bound. Same as the proof of Theorem 3, we consider an arbitrary $\hat{\sigma}^2_{\boldsymbol{c}} = \boldsymbol{X}^\top\boldsymbol{A}_{\boldsymbol{c}}\boldsymbol{X} = \sum_{k=1}^L c_kY_k \in \mathcal{Q}_L$. But this time we write $\boldsymbol{A}_{\boldsymbol{c}} = \boldsymbol{B}_1+\boldsymbol{B}_1^\top$, where $\boldsymbol{B}_1 = \frac{1}{2n}\sum_{k=1}^L c_k(\boldsymbol{I}-\boldsymbol{C}_k)$. As given in the proof of Theorem 1, it holds that

$$\|\boldsymbol{B}_1\boldsymbol{\theta}\|^2 = \|\boldsymbol{B}_1^\top\boldsymbol{\theta}\|^2 = W(\boldsymbol{\theta})\cdot\boldsymbol{c}^\top\boldsymbol{U}_L^\top\boldsymbol{U}_L\boldsymbol{c},$$

and hence

$$\|\boldsymbol{B\theta}\|^2 = \|\boldsymbol{B}_1\boldsymbol{\theta}+\boldsymbol{B}_1^\top\boldsymbol{\theta}\|^2 \le 4\|\boldsymbol{B}_1\boldsymbol{\theta}\|^2 = 4W(\boldsymbol{\theta})\cdot\boldsymbol{c}^\top\boldsymbol{U}_L^\top\boldsymbol{U}_L\boldsymbol{c}.$$

Therefore, by Lemma 5, on the class $\Theta_{L,w}$, the risk of $\hat{\sigma}^2_{\boldsymbol{c}}$ is bounded by

$$r(\hat{\sigma}^2_{\boldsymbol{c}}) \le \kappa_4 - 1 + \boldsymbol{c}^\top\left[\boldsymbol{I}_L + 4w\boldsymbol{U}_L^\top\boldsymbol{U}_L\right]\boldsymbol{c}.$$

According to the proof of the first part of Theorem 4, we have

$$\min_{\hat{\sigma}^2_{\boldsymbol{c}}\in\mathcal{Q}_L}\ \max_{(\boldsymbol{\theta},\sigma^2)\in\Theta_{L,w}} r(\hat{\sigma}^2_{\boldsymbol{c}}) \le g_L(4w).$$

So the upper bound has been derived.

Lemma 6 *For any integer $k\ge 1$, let $\boldsymbol{U}_k$ be the upper triangular matrix with 1 on and above the diagonal. Assume $\lambda \ge 0$.*

(i) Let D_k be the determinant of the matrix $\boldsymbol{I}_k+\lambda\boldsymbol{U}_k^\top\boldsymbol{U}_k$, then D_k satisfies the recursion $D_k = (2+\lambda)D_{k-1} - D_{k-2}$ with initial values $D_0 = 1$ and $D_1 = 1+\lambda$.

(ii) The cofactor of the $(1,k)$-th element of $\boldsymbol{I}_k+\lambda\boldsymbol{U}_k^\top\boldsymbol{U}_k$ is always $-\lambda$.

Proof of Lemma 6. Performing two operations on $\boldsymbol{I}_k + \lambda \boldsymbol{U}_k^\top \boldsymbol{U}_k$: subtracting the $(k-1)$-th row from the last row, and subtracting the $(k-1)$-th column from the last one, we have

$$\boldsymbol{I}_k + \lambda \boldsymbol{U}_k^\top \boldsymbol{U}_k \longrightarrow \begin{pmatrix} \boldsymbol{I}_{k-1} + \lambda \boldsymbol{U}_{k-1}^\top \boldsymbol{U}_{k-1} & -\boldsymbol{e}_{k-1} \\ -\boldsymbol{e}_{k-1}' & \lambda + 2, \end{pmatrix}$$

where $\boldsymbol{e}_{k-1}$ is a $(k-1)$-dimensional vector whose only nonzero element is the last one, with value 1. Therefore, it immediately follows that $D_k = (\lambda+2)D_{k-1} - D_{k-2}$. It is straightforward to verify that the initial values $D_0 = 1$ and $D_1 = 1 + \lambda$.

For the second part of the corollary, let $\boldsymbol{M}_{k1}$ be the $(k-1) \times (k-1)$ matrix obtained by deleting the first column and the last row from $\boldsymbol{I}_k + \lambda \boldsymbol{U}_k^\top \boldsymbol{U}_k$. Denote the rows of $\boldsymbol{M}_{k1}$ by $\boldsymbol{r}_i$, $1 \le i \le k-1$. Performing the row operations $\boldsymbol{r}_i - i/(i+1) \cdot \boldsymbol{r}_{i+1}$ successively for $i = 1, \ldots, k-2$, we end up with a lower triangular matrix with diagonal entries $\{-1/2, -2/3, \ldots, -(k-2)/(k-1), (k-1)\lambda\}$. Therefore, the cofactor is $(-1)^{k+1} \left[\prod_{i=1}^{k-2} -i/(i+1)\right] \cdot \lambda = -\lambda$. The proof is complete.

Proof of Proposition 6. From (39), it is straightforward to verify the value $g_L(0)$. For the derivative, we have

$$
\begin{aligned}
g_L'(0) =&(1,0)\frac{d}{dw}\left(\boldsymbol{Z}_L^\top(\boldsymbol{I}+w\boldsymbol{U}_L^\top\boldsymbol{U}_L)^{-1}\boldsymbol{Z}_L\right)^{-1}|_{w=0}\binom{1}{0}\\
=&(1,0)\left(\boldsymbol{Z}_L^\top(\boldsymbol{I}+w\boldsymbol{U}_L^\top\boldsymbol{U}_L)^{-1}\boldsymbol{Z}_L\right)^{-1}|_{w=0}\cdot\frac{d}{dw}\left(\boldsymbol{Z}_L^\top(\boldsymbol{I}+w\boldsymbol{U}_L^\top\boldsymbol{U}_L)^{-1}\boldsymbol{Z}_L\right)|_{w=0}\\
&\cdot\left(\boldsymbol{Z}_L^\top(\boldsymbol{I}+w\boldsymbol{U}_L^\top\boldsymbol{U}_L)^{-1}\boldsymbol{Z}_L\right)^{-1}|_{w=0}\binom{1}{0}\\
=&(1,0)\left(\boldsymbol{Z}_L^\top\boldsymbol{Z}_L\right)^{-1}\cdot\frac{d}{dw}\left(\boldsymbol{Z}_L^\top(\boldsymbol{I}+w\boldsymbol{U}_L^\top\boldsymbol{U}_L)^{-1}\boldsymbol{Z}_L\right)|_{w=0}\cdot\left(\boldsymbol{Z}_L^\top\boldsymbol{Z}_L\right)^{-1}\binom{1}{0}\\
=&(1,0)\left(\boldsymbol{Z}_L^\top\boldsymbol{Z}_L\right)^{-1}\boldsymbol{Z}_L^\top(\boldsymbol{I}+w\boldsymbol{U}_L^\top\boldsymbol{U}_L)^{-1}|_{w=0}\cdot\frac{d}{dw}(\boldsymbol{I}+w\boldsymbol{U}_L^\top\boldsymbol{U}_L)|_{w=0}\\
&\cdot(\boldsymbol{I}+w\boldsymbol{U}_L^\top\boldsymbol{U}_L)^{-1}|_{w=0}\boldsymbol{Z}_L\left(\boldsymbol{Z}_L^\top\boldsymbol{Z}_L\right)^{-1}\binom{1}{0}\\
=&(1,0)\left(\boldsymbol{Z}_L^\top\boldsymbol{Z}_L\right)^{-1}\boldsymbol{Z}_L^\top(\boldsymbol{U}_L^\top\boldsymbol{U}_L)\boldsymbol{Z}_L\left(\boldsymbol{Z}_L^\top\boldsymbol{Z}_L\right)^{-1}\binom{1}{0}\\
=&\frac{2(L+1)(L+2)(2L+1)}{15L(L-1)}.
\end{aligned}
$$

D Unbiased Estimators over the Class Θ_L^c

In this section we characterize the unbiased quadratic estimators over Θ_L^c, defined in (7). Recall that any quadratic estimator of σ^2 can be expressed as $\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X}$, where $\boldsymbol{A}=(a_{ij})$ is a $n\times n$ symmetric matrix. Let

$$
\begin{aligned}
\mathcal{I}_L := \{I\subset[n] : \ & I \text{ is a set of consecutive integers; } |I|\geq K;\\
& \text{either } [L]\in J, \text{ or } I\cap[L]=\emptyset;\\
& \text{either } [(n-L+1),n]\in I, \text{ or } [(n-L+1),n]\cap I=\emptyset.\}
\end{aligned}
$$

Proposition 7 *Assume $n\geq 2K$. The variance estimate $\hat\sigma_A^2$ is unbiased over Θ_L^c if and only* *if*

$$\sum_{j=1}^{n} a_{ii} = 1, \quad \textit{and} \quad \sum_{i,j \in I} a_{ij} = 0, \; \forall\, I \in \mathcal{I}_L.$$

The set of conditions given in Proposition 7 includes redundant ones. We provide an alternative set of conditions when $n > 3L$.

(C1) $\sum_{j=1}^{n} a_{ii} = 1$.

(C2) For each $2L + 1 \leq i \leq n - L$, $\sum_{j=1}^{L} a_{ij} = 0$.

(C3) For each $L + 1 \leq i \leq n - 2L$, $\sum_{j=n-L+1}^{n} a_{ij} = 0$.

(C4) $\sum_{i=1}^{L} \sum_{j=n-L+1}^{n} a_{ij} = 0$.

(C5) For each pair of i, j such that $L < i, j \leq n - L$ and $|i - j| > L$, $a_{ij} = 0$.

(C6) $\sum_{j_1,j_2=i}^{i+L-1} a_{j_1,j_2} = 0$, for all $i = 1$, $i = n - L + 1$, and $L + 1 \leq i \leq n - 2L + 1$.

(C7) $\sum_{j=i+1}^{i+L} a_{ij} + \frac{1}{2} a_{ii} = 0$, for $L + 1 \leq i \leq n - 2L$.

(C8) $\sum_{j=i+1}^{n} a_{ij} + \frac{1}{2} a_{ii} = 0$, for $n - 2L + 1 \leq i \leq n - L$.

(C9) $\sum_{j=1}^{i-1} a_{ij} + \frac{1}{2} a_{ii} = 0$, for $L + 1 \leq i \leq 2L$.

(C1)$\sim$(C9) form a minimal set of conditions to guarantee the unbiasedness of $\hat{\sigma}_A^2$ on the parameter space Θ_L^c.

E Additional Proofs

We collect the Proofs of Proposition 3 and Proposition 7 in this section. They are both regarding the model class Θ_L^c.

Proof of Proposition 3. The proof is based on comparing the variances of $\hat{\alpha}_K$ and $\check{\alpha}_K$ through (31). Recall from the proof of Theorem 1 that $\boldsymbol{d}_K = (d_1, \ldots, d_K)^\top$ is the coefficient vector of the OLS $\hat{\alpha}_K$. It also holds that $\check{\alpha}_K = (d_1 S_1 + \cdots d_K S_K)/2n$. The estimators $\hat{\alpha}_K$ and $\check{\alpha}_K$ can both be expressed in the quadratic form:

$$\hat{\alpha}_K = \tfrac{1}{2n}\boldsymbol{X}^\top \boldsymbol{A}_1 \boldsymbol{X}, \quad \check{\alpha}_K = \tfrac{1}{2n}\boldsymbol{X}^\top \boldsymbol{A}_2 \boldsymbol{X},$$

where $\boldsymbol{A}_1$ is a circulant matrix with entries d_k at locations (i,j) such that $(i-j) \mod n = \pm k$, and 2 on the diagonal. The matrix $\boldsymbol{A}_2$ is obtained from $\boldsymbol{A}_1$ by setting its upper-right and bottom-left $K \times K$ blocks as zero, top-left $K \times K$ block as

$$\operatorname{diag}\{1, 1+d_1, 1+d_1+d_2, \ldots, 1+d_1+\cdots+d_{K-1}\},$$

and bottom-right $K \times K$ block as

$$\operatorname{diag}\{1+d_1+\cdots+d_{K-1}, \ldots, 1+d_1, 1\}.$$

Let us repeat (31) here for easy reference, which says that when $\mathrm{E}\varepsilon_1^3 = 0$, the variance of any unbiased quadratic estimator $\boldsymbol{X}^\top \boldsymbol{A} \boldsymbol{X}$ equals

$$\operatorname{Var}(\boldsymbol{X}^\top \boldsymbol{A} \boldsymbol{X}) = 4\sigma^2 \boldsymbol{\theta}^\top \boldsymbol{A}^2 \boldsymbol{\theta} + \sigma^4 \left(2\operatorname{tr}(\boldsymbol{A}^2) + \frac{1}{n}(\kappa_4 - 3)\right).$$

We first calculate

$$\operatorname{tr}(\boldsymbol{A}_2^2) - \operatorname{tr}(\boldsymbol{A}_1^2) = 2\left(\sum_{k=1}^{K}(1+d_1+\cdots+d_{k-1})^2 - 4K\right) - 2\sum_{k=1}^{K} k \cdot d_k^2,$$

where the first term is due to the difference in the upper-left and bottom-right $K \times K$ blocks, and the second term comes from the upper-right and bottom-left blocks. The first term can be further calculated as

$$\begin{aligned}\sum_{k=1}^{K}(1+d_1+\cdots+d_{k-1})^2-4K &= \sum_{k=1}^{K}(2-d_k-\cdots-d_K)^2-4K\\ &= \sum_{k=1}^{K}(d_k+\cdots+d_K)^2-4\sum_{k=1}^{K}(d_k+\cdots+d_K)\\ &= \sum_{k=1}^{K}(d_k+\cdots+d_K)^2,\end{aligned}$$

where in the first and last identities we have used the fact $\sum_{k=1}^{K} d_k = 1$ and $\sum_{k=1}^{K} k\cdot d_k = 0$ respectively. Now we calculate

$$\boldsymbol{\theta}^\top \boldsymbol{A}_2^2\boldsymbol{\theta} - \boldsymbol{\theta}^\top \boldsymbol{A}_1^2\boldsymbol{\theta} = 2(\theta_n-\theta_1)^2\cdot\sum_{k=1}^{K}(d_k+\cdots+d_K)^2.$$

Similar calculations to Lemma 2 give that

$$\begin{aligned}\sum_{k=1}^{K}(d_k+\cdots+d_K)^2 &= \frac{(K+1)(K+2)(2K=1)}{15K(K-1)},\\ \sum_{k=1}^{K}k\cdot d_k^2 &= \frac{(K+1)(K+2)}{K(K-1)}.\end{aligned}$$

Combining the preceding results, we have

$$\begin{aligned}&\mathrm{Var}(\boldsymbol{X}^\top\boldsymbol{A}_2\boldsymbol{X})-\mathrm{Var}(\boldsymbol{X}^\top\boldsymbol{A}_1\boldsymbol{X})\\ &\qquad= 4\sigma^2\left[\sigma^2-2(\theta_n-\theta_1)^2\right]\cdot\frac{(K+1)(K+2)(2K-1)}{15K(K-1)}-4\sigma^4\cdot\frac{(K+1)(K+2)}{K(K-1)}\\ &\qquad= 4\sigma^2\left[-2(\theta_n-\theta_1)^2\cdot\frac{(K+1)(K+2)(2K-1)}{15K(K-1)}+\sigma^2\cdot\frac{2(K-7)(K+1)(K+2)}{15K(K-1)}\right].\end{aligned}$$

This completes the proof of Proposition 3 when $K\le L(\boldsymbol{\theta})/2$.

We now consider the case $K\le L(\boldsymbol{\theta})$. By examining the proof of Theorem 1, we see that on the model class Θ_L^c,

$$\boldsymbol{\theta}^\top \boldsymbol{A}_2^2 \boldsymbol{\theta} \le V(\boldsymbol{\theta}) \frac{(K+1)(K+2)^2}{3K(K-1)}.$$

On the other hand, the difference between $\mathrm{tr}(\boldsymbol{A}_2^2)$ and $\mathrm{tr}(\boldsymbol{A}_1^2)$ remains the same as the previous case. Combining these facts completes the proof.

Proof of Proposition 7. The proof of Proposition 7 is very similar to that of Lemma 4, adapting it to the model class Θ_L^c. We omit the details. The conditions (C1)–(C9) form a minimal set of conditions which will imply the condition in Proposition 7. The proof of its sufficiency is self evident, and will be skipped as well.