The Annals of Statistics
2020, Vol. 48, No. 4, 2455–2477
https://doi.org/10.1214/19-AOS1894
© Institute of Mathematical Statistics, 2020

ESTIMATION AND INFERENCE FOR PRECISION MATRICES OF NONSTATIONARY TIME SERIES

BY XIUCAI DING[*] AND ZHOU ZHOU[**]

Department of Statistics, University of Toronto, [*]xiucai.ding@mail.utoronto.ca; [**]zhou@utstat.toronto.edu

We consider the estimation of and inference on precision matrices of a rich class of univariate locally stationary linear and nonlinear time series, assuming that only one realization of the time series is observed. Using a Cholesky decomposition technique, we show that the precision matrices can be directly estimated via a series of least squares linear regressions with smoothly time-varying coefficients. The method of sieves is utilized for the estimation and is shown to be optimally adaptive in terms of estimation accuracy and efficient in terms of computational complexity. We establish an asymptotic theory for a class of $\mathcal{L}^2$ tests based on the nonparametric sieve estimators. The latter are used for testing whether the precision matrices are diagonal or banded. A Gaussian approximation result is established for a wide class of quadratic forms of nonstationary and possibly nonlinear processes of diverging dimensions which is of interest by itself.

1. Introduction. Consider a centered univariate nonstationary time series $x_{1,n}, \ldots, x_{n,n} \in \mathbb{R}$. Denote by $\Omega_n := [\mathrm{Cov}(x_{1,n}, \ldots, x_{n,n})]^{-1}$ the precision matrix of the series. Modelling, estimation and inference of Ω_n are of fundamental importance in a wide range of problems in time series analysis. For example, the $\mathcal{L}^2$ optimal linear forecast of $x_{n+1,n}$ based on $x_{1,n}, \ldots, x_{n,n}$ is determined by Ω_n and the covariance between $x_{n+1,n}$ and $(x_{1,n}, \ldots, x_{n,n})$ [3]. In time series regression with fixed regressors, the best linear unbiased estimator of the regression coefficient is a weighted least squares estimator with weights proportional to the square root of the precision matrix of the errors [15]. Furthermore, the precision matrix is a key part in Gaussian likelihood and quasilikelihood estimation and inference of time series [3, 19]. We shall omit the subscript n in the sequel if no confusions arise.

Observe that Ω is an $n \times n$ matrix. When the time series length n is at least moderately large, it is generally not a good idea to first estimate the covariance matrix of $(x_1, \ldots, x_n)$ and then invert it to obtain an estimate of Ω. One main reason is that small errors in the covariance matrix estimation may be amplified through inversion when n is large, especially when the condition number of the covariance matrix is large. Also, matrix inversion is not computationally efficient for large n. As a result it is desirable to directly estimate Ω. In this paper we utilize a Cholesky decomposition technique to directly estimate Ω through a series of least squares linear regressions. Specifically, write

$$x_i = \sum_{j=1}^{i-1} \phi_{ij} x_{i-j} + \epsilon_i, \quad i = 2, \ldots, n, \tag{1.1}$$

where $\sum_{j=1}^{i-1} \phi_{ij} x_{i-j} := \widehat{x}_i$ is the best linear forecast of x_i based on $x_1, \ldots, x_{i-1}$ and ϵ_i is the forecast error. Let $\epsilon_1 := x_1$ and denote by σ_i^2 the variance of ϵ_i, $i = 1, 2, \ldots, n$. Observe that ϵ_i are uncorrelated random variables. As a result it is straightforward to show that [22]

$$\Omega = \Phi^* \widetilde{\mathbf{D}} \Phi, \tag{1.2}$$

Received March 2019; revised August 2019.
MSC2020 subject classifications. 62M10, 62G05, 62G10.
Key words and phrases. Nonstationary time series, precision matrices, Cholesky decomposition, sieve estimation, high-dimensional Gaussian approximation, random matrices, white noise and bandedness tests.

where the diagonal matrix $\widetilde{\mathbf{D}}$ is defined as $\widetilde{\mathbf{D}} = \operatorname{diag}\{\sigma_1^{-2}, \dots, \sigma_n^{-2}\}$, Φ is a lower triangular matrix having ones on its diagonal and $-\phi_{ij}$ at its $(i, i-j)$th element for $j < i$ and $*$ denotes matrix or vector transpose. The most significant advantage of the above Cholesky decomposition is structural simplification that transfers the difficult problem of precision matrix estimation to that of estimating a series of least squares regression coefficients and error variances.

However, the Cholesky decomposition idea is not directly applicable to precision matrix estimation of nonstationary time series. The reason is that there are in total $n(n+1)/2$ regression coefficients and error variances to be estimated in the Cholesky decomposition of Ω. Meanwhile, observe that there are also $n(n+1)/2$ parameters to be estimated for the precision matrix of a general nonstationary time series. Hence, Cholesky decomposition, although it performs structural simplification, does not reduce the dimensionality of the parameter space. On the other hand, we only observe one realization of the time series with n observations. As a result dimension reduction techniques with natural assumptions in nonstationary time series analysis are needed for the estimation of Ω.

We adopt two natural and widely used assumptions in nonstationary time series for the dimension reduction. The first such assumption is local stationarity which refers to slowly or smoothly time-varying underlying data generating mechanisms of the series. Utilizing the locally stationary framework in Zhou and Wu [39], we show that, for a wide class of locally stationary nonlinear processes, each off-diagonal element of the Φ matrix as well as the error variance series σ_i^2 can be well approximated by smooth functions on $[0,1]$. Specifically, we show that there exist smooth functions $\phi_j(\cdot)$ and $g(\cdot)$ such that $\sup_{i>b} |\phi_{ij} - \phi_j(i/n)| = o(n^{-1/2})$, $j = 1, 2, \dots, n$ and $\sup_{i>b} |\sigma_i^2 - g(i/n)| = o(n^{-1/2})$, where $b = b_n$ diverges to infinity with $b/n \to 0$ whose specific value will be determined later in the article. To our knowledge, the latter is the first result on smooth approximation to general nonstationary precision matrices. From classic approximation theory [25], a d times continuously differentiable function can be well approximated by a basis expansion with $O((n/\log n)^{1/(2d+1)})$ parameters. Thanks to the local stationarity assumption, the number of parameters needed for estimating σ_i^2 is reduced from n to $O((n/\log n)^{1/(2d+1)})$. A similar conclusion holds for each off-diagonal element of Φ.

The second assumption we adopt is short range dependence which refers to fast decay of the dependence between x_i and x_{i+j} as j diverges. Using the physical dependence measures introduced in Zhou and Wu [39], modern operator spectral theory and approximation theory [9, 27], we show, as a theoretical contribution of the paper, that the off-diagonal elements of Φ decay fast to zeros for a general class of locally stationary short range dependent processes. Specifically, we show that ϕ_{ij} can be effectively treated as 0 whenever $j > b$. Hence, the total number of parameters one needs to estimate is reduced to the order $b[b + (n/\log n)^{1/(2d+1)}]$ which is typically much smaller than the sample size n.

Now, we utilize the method of sieves to estimate the smooth functions $\phi_j(\cdot)$ and $g(\cdot)$ mentioned above. The method of sieves refers to approximating an infinite dimensional space with a sequence of finer and finer finite dimensional subspaces. Typical examples include Fourier, wavelet and orthogonal polynomial approximations to smooth functions on compact intervals. We refer to [6] by Chen for a thorough review of the subject. There are two major advantages of the sieve method when used for precision matrix estimation. First, many sieve estimators, such as the Fourier and wavelet methods mentioned above, do not have inferior performances at the boundary of the estimating interval. This is important as inaccurate estimates at the boundary may drastically lower the accuracy of the whole precision matrix estimation even though entries are well estimated in the interior. Second, the computational complexities of many sieve methods are both adaptive (to the smoothness of the functions of interest) and efficient. When estimating one smooth function of time, local methods such as the kernel estimation perform one regression at each time point. This could be computational inefficient when n is large. On the contrary, the above mentioned three sieve methods only need to perform a single regression at the whole time interval with the number of covariates determined by the smoothness of the function of interest. In many cases this yields a much faster estimation. For instance, in the extreme case where the time series dependence is exponentially decaying and the functions are infinitely differentiable, the sieve method only needs $O(n \log^5 n)$ operations to estimate Ω. Under the same scenario the computational complexity of the kernel method is of order $O(n^2 k_n \log^2 n)$, where k_n is the bandwidth used for the regression and is typically of the order $n^{-1/5}$.

In this paper we show that the sieve estimates of the functions $\phi_j(\cdot)$ achieve, uniformly over time and j, the minimax rate for nonparametric function estimation [25]. This extends previous convergence rate results on nonparametric sieve regression to the case of a diverging number of covariates and nonstationary predictors and errors. Combining the latter result with modern random matrix theory [28], we show that the operator norm of the estimated precision matrix converges at a fast rate which is determined by the strength of time series dependence and the smoothness of the underlying data generating mechanism. In the best scenario where the dependence is exponentially decaying and $\phi_j(\cdot)$ and $g(\cdot)$ are infinitely differentiable, the convergence rate is shown to be of the order $\log^3 n/\sqrt{n}$, which is almost as fast as parametrically estimating a single parameter from i.i.d. samples.

Sieve estimators have already been used to estimate the smooth conditional mean function in various settings. For instance, in [1], the authors proved that the sieve least square estimators could achieve minimax rate in the sense of sup-norm loss for a fixed number of i.i.d. regressors and errors with a general class of sieve basis functions; later, Chen and Christensen [7] showed that the spline and wavelet sieve regression estimators attain the above global uniform convergence rate for a fixed number of weakly dependent and stationary regressors. In this article we study nonparametric sieve estimates for locally stationary time series with diverging number of covariates under physical dependence and obtain the same minimax rate for the functions $\phi_j(\cdot)$.

After estimating Ω, one may want to perform various tests on its structure. In this paper we focus on two such tests, one on whether $\{x_i\}_{i=1}^n$ is a nonstationary white noise and the other on whether Ω is banded. Two test statistics based on the $\mathcal{L}^2$ distances between the estimated and hypothesized Φ are proposed. These tests boil down to quadratic forms of the estimated sieve regression coefficients which are quadratic forms of nonstationary, dependent vectors of diverging dimensionality. To our knowledge, there have been no previous works on $\mathcal{L}^2$ inference of nonparametric sieve estimators as well as the inference of high-dimensional quadratic forms of nonstationary nonlinear time series. Here, we utilize Stein's method together with an m-dependence approximation technique and prove that the laws of a large class of quadratic forms of nonstationary nonlinear processes with diverging dimensionality can be well approximated by those of quadratic forms of diverging dimensional Gaussian processes. Consequently, asymptotic normality can be established for those high-dimensional quadratic forms. The latter Gaussian approximation result is of separate interest and may be of wider applicability in nonstationary time series analysis. In [33], Xu, Zhang and Wu derived the $\mathcal{L}^2$ asymptotics for the quadratic form $\overline{X}^*\overline{X}$, where $\overline{X}$ is the sample mean of n i.i.d. random vectors and $\overline{X}^*$ is its transpose. In the present paper we prove new and much more general $\mathcal{L}^2$ asymptotics for quadratic forms $\overline{\mathbf{Z}}^* E \overline{\mathbf{Z}}$ for any bounded positive semidefinite matrix E using Stein's method [24, 37], where $\overline{\mathbf{Z}}$ is the sample mean of a high-dimensional, nonstationary and dependent process. It is very interesting that similar ideas have been used in proving the universality of random matrix theory [11, 13, 17, 26].

We point out that the idea of Cholesky decomposition has been used in time series analysis under some different settings when multiple replicates of the vector of interest are available.

Assuming a longitudinal setup where multiple realizations can be observed, Wu and Pourahmadi [31] studied the estimation of covariance matrices using nonparametric smoothing techniques. Bickel and Levina [2] considered estimating large covariance and precision matrices by either banding or tapering the sample covariance matrix and its inverse, assuming that multiple independent samples can be observed. On the other hand, we assume that only one realization of the time series is observed which is the case in many real applications. Hence, none of the aforementioned results can be applied under this scenario.

Finally, we mention that estimating large-dimensional covariance and precision matrices has attracted much attention in the last two decades. One main research line is to assume that we can observe n i.i.d. copies of a p dimensional random vector. When p is comparable to or larger than n, it is well known that sample covariance and precision matrices are inconsistent estimators [10, 21]. To overcome the difficulty from high dimensionality, researchers usually impose two main structural assumptions in order to consistently estimate the covariance and precision matrices, sparsity structure and factor model structure. Various families of covariance matrices and regularization methods have been introduced assuming some types of sparsity; this includes the bandable covariance matrices [2, 4, 31], sparse covariance matrices [5, 18, 36] and sparse precision matrices [34, 35]. On the other hand, factor models in the high-dimensional setting have been used in a range of applications in finance and economics. For a comprehensive review on factor model based methods, we refer to [14]. Although high dimensional covariance and precision matrix estimation has witnessed unprecedented development, there have been no previous works on precision matrix estimation for nonstationary time series with only one realization to the best of our knowledge. On the other hand, there have been a small literature on time series covariance or precision matrix estimation. Under stationarity, [20, 32] consider thresholding and banding techniques for estimating the covariance matrix with only one realization of the series. Under sparsity assumptions, [8] estimates *marginal* covariance and precision matrices of high-dimensional stationary and locally stationary time series using thresholding and Lasso techniques. Note that when estimating marginal covariance or precision matrices of a p dimensional time series of length n, the series can be viewed as n dependent replicates of the vector of interest which is completely different than the situation considered in this article.

The rest of the paper is organized as follows. In Section 2 we introduce a rich class of nonstationary (locally stationary) and nonlinear time series and study the theoretical properties of its covariance and precision matrices. In Section 3 we consistently estimate the precision matrices and provide convergent rates for these estimators. In Section 4 we propose two adaptive tests using some simple statistics from our estimation procedure, where the convergence rates and powers of the tests are adaptive to the strength of the temporal dependence and the smoothness of the underlying data generating mechanism. Monte Carlo simulations, technical proofs and auxiliary lemmas are provided in the Supplementary Material [12].

2. Locally stationary time series.

Consider a locally stationary time series [38–40]

$$x_i = G\left(\frac{i}{n}, \mathcal{F}_i\right), \tag{2.1}$$

where $\mathcal{F}_i = (\dots, \eta_{i-1}, \eta_i)$ and $\eta_i, i \in \mathbb{Z}$ are i.i.d. random variables and $G : [0,1] \times \mathbb{R}^\infty \to \mathbb{R}$ is a measurable function such that $\xi_i(t) := G(t, \mathcal{F}_i)$ is a properly defined random variable for all $t \in [0,1]$. The above represents a wide class of locally stationary linear and nonlinear processes. We refer to Zhou and Wu [30, 39, 40] for detailed discussions and examples. And following [30, 39, 40], we introduce the following dependence measure to quantify the temporal dependence of (2.1).

DEFINITION 2.1. Let $\{\eta_i'\}$ be an i.i.d. copy of $\{\eta_i\}$. We assume that for some $q > 0$, $\|x_i\|_q < \infty$, where $\|\cdot\|_q = [\mathbb{E}|\cdot|^q]^{1/q}$ is the $\mathcal{L}_q$ norm of a random variable. For $j \geq 0$, we define the physical dependence measure by

$$\delta(j, q) := \sup_{t \in [0,1]} \max_i \big\| G(t, \mathcal{F}_i) - G(t, \mathcal{F}_{i,j}) \big\|_q, \tag{2.2}$$

where $\mathcal{F}_{i,j} := (\mathcal{F}_{i-j-1}, \eta_{i-j}', \eta_{i-j+1}, \dots, \eta_i)$.

The measure $\delta(j, q)$ quantifies the changes in the system's output when the input of the system j steps ahead is changed to an i.i.d. copy. If the change is small, then we have short-range dependence. It is notable that $\delta(j, q)$ is related to the data generating mechanism and can be easily computed. We refer the readers to [39], Section 4, for examples of such computation. In the present paper we impose the following assumptions on (2.1) and the physical dependence measure to control the temporal dependence of the nonstationary time series.

ASSUMPTION 2.2. There exist constants $\tau > 10$ and $q > 4$; for some universal constant $C > 0$, we have that

$$\delta(j, q) \leq C j^{-\tau}, \quad j \geq 1. \tag{2.3}$$

Furthermore, G defined in (2.1) satisfies the property of stochastic Lipschitz continuity, for any $t_1, t_2 \in [0, 1]$, we have

$$\big\| G(t_1, \mathcal{F}_i) - G(t_2, \mathcal{F}_i) \big\|_q \leq C_1 |t_1 - t_2|, \tag{2.4}$$

where C_1 is some universal constant independent of i, t_1 and t_2. We also assume that

$$\sup_t \max_i \big\| G(t, \mathcal{F}_i) \big\|_q < \infty. \tag{2.5}$$

Equation (2.3) indicates that the time series has short-range dependence. Further, (2.4) implies that $G(\cdot, \cdot)$ changes smoothly over time and ensures local stationarity. Furthermore, for each fixed $t \in [0, 1]$, denote

$$\gamma(t, j) = \operatorname{Cov}\big(G(t, \mathcal{F}_0), G(t, \mathcal{F}_j)\big). \tag{2.6}$$

Observe that $\gamma(t, j)$ is the jth order autocovariance of the time series $\{x_i\}_{i=1}^n$ at time t. Equations (2.4) and (2.5) imply that $\gamma(t, j)$ is Lipschitz continuous in t. Further, we need the following assumption on the smoothness of $\gamma(t, j)$.

ASSUMPTION 2.3. There exists integer $d \geq 1$ such that for all $j \geq 0$, we have $\gamma(t, j) \in C^d([0, 1])$, where $C^d([0, 1])$ is the function space on $[0, 1]$ of continuous functions that have continuous first d derivatives.

Finally, in this paper, we assume that for any $n \in \mathbb{N}$, there exists some universal constant $\varsigma > 0$ such that the smallest eigenvalue of the covariance matrix of $(x_1, \dots, x_n)$, denoted by $\lambda_n(\Sigma_n)$, satisfies

$$\lambda_n(\Sigma_n) \geq \varsigma. \tag{2.7}$$

The above assumption is commonly used in the literature on precision matrix estimation [5, 8, 34].

2.1. *Examples.* In this subsection we list a few examples of locally stationary processes satisfying Assumptions 2.2 and 2.3.

EXAMPLE 2.4 (Nonstationary linear processes). Let $\{\epsilon_i\}$ be i.i.d. random variables; let $a_j(\cdot)$, $j = 0, 1, \dots$ be $C^d([0,1])$ functions such that

$$G(t, \mathcal{F}_i) = \sum_{k=0}^{\infty} a_j(t)\epsilon_{i-k}.$$

The above model is studied in [39], Section 4.1. By [39], Proposition 2, we find that Assumption 2.2 will be satisfied if

$$\begin{aligned}&\sup_{t\in[0,1]} \left|a_j(t)\right|^{\min(2,q)} \le Cj^{-\tau}, \quad j \ge 1;\\ &\sum_{j=0}^{\infty} \sup_{t\in[0,1]} \left|a_j'(t)\right|^{\min(2,q)} < \infty,\end{aligned} \tag{2.8}$$

for some constant $C > 0$. Furthermore, by the assumption (2.8) and the rule of term by term differentiation [29], Theorem 7.14, Assumption 2.3 will be satisfied if

$$\sup_{t\in[0,1]} \left|a_j^{(d)}(t)\right|^{\min(2,q)} \le Cj^{-\tau}, \quad j \ge 1.$$

EXAMPLE 2.5 (Nonstationary nonlinear process). Let $\{\epsilon_i\}$ be i.i.d. random variables. We now consider a process of the following form:

$$\xi_i(t) = R\bigl(t, \xi_{i-1}(t), \epsilon_i\bigr), \tag{2.9}$$

where R is some (possibly nonlinear) measurable function. This process has been studied in [39], Section 4.2. Suppose that for some x_0, we have $\sup_{t\in[0,1]} \|R(t, x_0, \epsilon_i)\|_q < \infty$. Denote

$$\chi := \sup_{t\in[0,1]} L(t), \quad \text{where } L(t) = \sup_{x \ne y} \frac{\|R(t,x,\epsilon_0) - R(t,y,\epsilon_0)\|_q}{|x-y|}.$$

It is known from [39], Theorem 6, that if $\chi < 1$, then (2.9) admits a unique locally stationary solution with $\xi_i(t) = G(t, \mathcal{F}_i)$ and the physical dependence measure satisfies that $\delta(j,q) \le C\chi^j$. Hence, the temporal dependence is of exponential decay (see equation (2.15)), which is much faster than (2.2). Furthermore, we conclude from [39] that (2.4, Proposition 4) holds true if

$$\begin{aligned}&\sup_{t\in[0,1]} \bigl\|M\bigl(G(t, \mathcal{F}_0)\bigr)\bigr\|_q < \infty,\\ &\quad\text{where } M(x) = \sup_{0\le t<s\le 1} \frac{\|R(t,x,\epsilon_0) - R(s,x,\epsilon_0)\|_q}{|t-s|}.\end{aligned}$$

To verify Assumption 2.3, we assume that $G(t, \mathcal{F}_i)$ admits the following Volterra expansion [30]:

$$G(t, \mathcal{F}_i) = \sum_{k=1}^{\infty} \sum_{u_1,\dots,u_k=0}^{\infty} g_k(t, u_1, \dots, u_k)\epsilon_{i-u_1} \dots \epsilon_{i-u_k},$$

where g_k's are the Volterra kernels. Suppose $g_k \in C^d[0,1]$ for t and

$$\sup_{t\in[0,1]} \sum_{k=1}^{\infty} \sum_{u_1,\dots,u_k=0}^{\infty} \bigl(g_k^{(d)}(t, u_1, \dots, u_k)\bigr)^2 < \infty.$$

Then, we can use term by term differentiation to see that Assumption 2.3 holds.

2.2. *Properties of the time-varying best linear predictors.* Many important consequences can be derived due to Assumptions 2.2 and 2.3. We list the most useful ones in this section and put their proofs into the Supplementary Material [12]. The following lemma controls the uniform decay rate of $\gamma(t, j)$ as j increases.

LEMMA 2.6. *Under Assumptions 2.2 and 2.3 there exists some constant $C > 0$, such that*

$$\sup_t \big|\gamma(t, j)\big| \le C j^{-\tau}, \quad j \ge 1.$$

Our first important conclusion is that the coefficients defined in (1.1) are of polynomial decay. Hence, when $i > b$ is large, where $b = O(n^{2/\tau})$, we only need to focus on autoregressive fit of order b instead of $i - 1$. Recall (1.1). Denote $\boldsymbol{\phi}_i = (\phi_{i1}, \dots, \phi_{i,i-1})^*$. Then, we have

$$\boldsymbol{\phi}_i = \Omega_i \boldsymbol{\gamma}_i, \tag{2.10}$$

where Ω_i and $\boldsymbol{\gamma}_i$ are defined as $\Omega_i = [\mathrm{Cov}(\mathbf{x}_{i-1}^i, \mathbf{x}_{i-1}^i)]^{-1}$, $\boldsymbol{\gamma}_i = \mathrm{Cov}(\mathbf{x}_{i-1}^i, x_i)$, with $\mathbf{x}_{i-1}^i = (x_{i-1}, \dots, x_1)^*$. The above claims are formally summarized in the following proposition.

PROPOSITION 2.7. *Under Assumption 2.2 and (2.7) and letting $b = O(n^{2/\tau})$, there exists some constant $C > 0$, such that*

$$|\phi_{ij}| \le \begin{cases} \max\{C n^{-4+5/\tau}, C j^{-\tau}\} & i \ge b^2; \\ \max\{C n^{-2+3/\tau}, C j^{-\tau}\} & b < i < b^2. \end{cases} \tag{2.11}$$

Furthermore, when $i > b$, denote $\boldsymbol{\phi}_i^b = (\phi_{i1}, \dots, \phi_{ib})$ and $\widetilde{\boldsymbol{\phi}}_i^b = \Omega_i^b \gamma_i^b$ with entries $(\widetilde{\phi}_{i1}, \dots, \widetilde{\phi}_{ib})$, where $\Omega_i^b = [\mathrm{Cov}(\mathbf{x}_i, \mathbf{x}_i)]^{-1}$, $\gamma_i^b = \mathbb{E}(\mathbf{x}_i x_i)$, $\mathbf{x}_i = (x_{i-1}, \dots, x_{i-b})$, we have

$$\sup_i \big\|\boldsymbol{\phi}_i^b - \widetilde{\boldsymbol{\phi}}_i^b\big\| \le C n^{-2+1/\tau}.$$

To our knowledge, Proposition 2.7 is the first result on the decay rate of the best linear forecast coefficients under nonstationarity. It serves the first dimension reduction for our parameter space. It states that we can treat $\phi_{ij} = 0$ when $j > b$. Hence, the number of coefficients needed for the Cholesky decomposition reduces from $O(n^2)$ to $O(nb)$. Finally, denote $\boldsymbol{\phi}^b(\frac{i}{n}) := (\phi_1(\frac{i}{n}), \dots, \phi_b(\frac{i}{n}))$ by

$$\boldsymbol{\phi}^b\left(\frac{i}{n}\right) = \widetilde{\Omega}_i^b \widetilde{\boldsymbol{\gamma}}_i^b, \tag{2.12}$$

where $\widetilde{\Omega}_i^b$ and $\widetilde{\boldsymbol{\gamma}}_i^b$ are defined as

$$\widetilde{\Omega}_i^b = \big[\mathrm{Cov}(\widetilde{\mathbf{x}}_i, \widetilde{\mathbf{x}}_i)\big]^{-1}, \qquad \widetilde{\boldsymbol{\gamma}}_i = \mathrm{Cov}(\widetilde{\mathbf{x}}_i, x_i),$$

with $\widetilde{\mathbf{x}}_{i,k} = G(\frac{i}{n}, \mathcal{F}_{i-k})$, $k = 1, 2, \dots, b$. The following lemma shows that $\boldsymbol{\phi}_i^b$ can be well approximated by $\boldsymbol{\phi}^b(\frac{i}{n})$ when $i > b$.

LEMMA 2.8. *Under Assumption 2.2 and the assumption (2.7), there exists some constant $C > 0$, such that for all $j \le b$,*

$$\sup_{i > b} \left|\phi_{ij} - \phi_j\left(\frac{i}{n}\right)\right| \le C n^{-1+2/\tau}.$$

Lemma 2.8 claims that each off-diagonal element $\{\phi_{ij}\}_{i=b}^n$ can be well approximated by a smooth function $\phi_j(\cdot)$. It provides the second dimension reduction. Due to the smoothness of $\phi_j(\cdot)$, it can be well approximated by a sieve expansion of order c, where $c \ll n$. This will further reduce the dimension of the parameter space from $O(nb)$ to $O(bc)$. Throughout of the rest of the paper, unless otherwise specified, we will always assume $b = O(n^{2/\tau})$. Recall that ϵ_i is the prediction error with variance σ_i^2,

$$\epsilon_i = x_i - \widehat{x}_i. \tag{2.13}$$

We define $\widetilde{\epsilon}_i := x_i - \sum_{j=1}^{\min(b,i-1)} \phi_{ij} x_{i-j}$. First of all, we deduce from Proposition 2.7 and Assumption 2.2 that

$$\max_{1 \leq i \leq n} |\epsilon_i - \widetilde{\epsilon}_i| = o(n^{-3}) \quad \text{in probability.} \tag{2.14}$$

Denote $\widetilde{\sigma}_i^2$ as the variance of $\{\widetilde{\epsilon}_i\}$. Then, we have

LEMMA 2.9. *Suppose Assumption 2.2 and (2.7) hold true. We have* $\sup_i \widetilde{\sigma}_i^2 < \infty$. *Furthermore, denote the physical dependence measure of* $\{\widetilde{\epsilon}_i\}$ *as* $\delta^{\epsilon}(j,q)$. *Then, there exists some constant* $C > 0$, *such that for* $\delta^{\epsilon}(j,q) \leq Cj^{-\tau}$.

REMARK 2.10. In this paper we focus on the discussion when the physical dependence measure is of polynomial decay, that is, (2.3) holds true. However, all our results can be extended to the case when it is of exponential decay

$$\delta(j,q) \leq Ca^j, \quad 0 < a < 1. \tag{2.15}$$

In detail, Lemma 2.6 can be changed to $\sup_t |\gamma(t,j)| < Ca^j$, $j \geq 1$. Therefore, we only need to choose $b = O(\log n)$. As a consequence, Proposition 2.7 can be updated to

$$\sup_{i>b} |\phi_{ij}| \leq \max\{Cn^{-C}, Ca^j\}, \qquad \sup_i \|\boldsymbol{\phi}_i^b - \widetilde{\boldsymbol{\phi}}_i^b\| \leq Cn^{-C},$$

where $C > 1$ is some constant. Similarly, Lemma 2.8 can be modified to

$$\sup_{i>b} \left| \phi_{ij} - \phi_j\left(\frac{i}{n}\right) \right| \leq \frac{C \log n}{n}, \quad \text{for all } j \leq b.$$

Finally, the analog of Lemma 2.9 is $\delta^{\epsilon}(j,q) \leq Ca^j$.

3. Estimation of precision matrices.

As shown in (1.2), Proposition 2.7 and Lemma 2.8, in order to estimate Ω it suffices to estimate ϕ_{ij}, $i \leq b$, $\phi_j(\frac{i}{n}), i > b \geq j$ and the variances of the residuals. When $i > b$, by (2.5) and Proposition 2.7, it is easy to see that

$$\sup_i \left| \sum_{j=b+1}^{i-1} \phi_{ij} x_{i-j} \right| = o(n^{-1}) \quad \text{in probability.}$$

Therefore, we can simply write

$$x_i = \sum_{j=1}^{b} \phi_{ij} x_{i-j} + \epsilon_i + o_u(n^{-1}), \quad i = b+1, \ldots, n, \tag{3.1}$$

where $X_i = o_u(n^{-1})$ means nX_i converges to zero in probability uniformly.

3.1. *Estimating ϕ_{ij} for $i > b$.* We first estimate the time-varying coefficients $\phi_j(\frac{i}{n})$ using the method of sieves [1, 6, 7] when $i > b$.

LEMMA 3.1. *Under Assumptions 2.2 and 2.3, for any $j \le b$, we have that $\phi_j(t) \in C^d([0,1])$.*

Based on the above lemma, we use

$$\theta_j\left(\frac{i}{n}\right) := \sum_{k=1}^{c} a_{jk}\alpha_k\left(\frac{i}{n}\right), \quad j \le b, \tag{3.2}$$

to approximate $\phi_j(\frac{i}{n})$, where $\{\alpha_k(t)\}$ is a set of prechosen orthogonal basis functions on $[0,1]$ and $c \equiv c(n)$ stands for the number of basis functions. In the present paper, unless otherwise specified, we always set $c = O(n^{\alpha_1})$. The estimation of $\theta_j(t)$ boils down to that of the a_{jk}'s. Next, the results of the convergent rate on the approximation (3.2) can be found in [6], Section 2.3. We summarize it in the following lemma.

LEMMA 3.2. *Denote the sup-norm with respect to Lebesgue measure as*

$$\mathcal{L}_\infty := \sup_{t\in[0,1]} \left|\phi_j(t) - \theta_j(t)\right|.$$

We have that $\mathcal{L}_\infty = O(c^{-d})$ for the orthogonal polynomials, trigonometric polynomials, spline series with order r when $r \ge d+1$ and orthogonal wavelets with degree m when $m > d$.

Then, we impose the following regularity condition on the basis functions.

ASSUMPTION 3.3. Let $\otimes$ be the Kronecker product. For any $k = 1, 2, \ldots, b$, denote $\Sigma^k(t) \in \mathbb{R}^{k\times k}$ via $\Sigma^k_{ij}(t) = \gamma(t, |i-j|)$, we assume the eigenvalues of

$$\Sigma^k := \int_0^1 \Sigma^k(t) \otimes \big(\mathbf{b}(t)\mathbf{b}^*(t)\big)\,dt \tag{3.3}$$

are bounded above and also away from zero by a constant $\kappa > 0$, where $\mathbf{b}(t) = (\alpha_1(t), \ldots, \alpha_c(t))^* \in \mathbb{R}^c$. Further, we assume that (2.7) holds.

Since $\Sigma^k(t) \otimes (\mathbf{b}(t)\mathbf{b}^*(t))$ is positive semidefinite for any $t \in [0,1]$, the above integral is always positive semidefinite. Assumption 3.3 is mild. It is clear that when x_i is a stationary process, the assumption will hold immediately due to the orthonormality of the basis functions. We next provide one specific nonstationary example. Consider a locally stationary MA(q) process of the form

$$G(t, \mathcal{F}_i) = \sum_{j=1}^{q} a_j(t)\epsilon_{i-j} + \epsilon_i, \quad 1 \le q < \infty, \tag{3.4}$$

where ϵ_i are i.i.d. centered random variables with variance 1. The following lemma shows, under suitable conditions, Assumption 3.3 holds true for (3.4).

LEMMA 3.4. *Suppose the assumptions of Example 2.4 hold for (3.4) and*

$$\sup_t \sum_{i=1}^{q} \left|a_j(t)\right| < 1. \tag{3.5}$$

Then, Assumption 3.3 holds for (3.4) and any orthonormal basis functions.

Next, we impose the following mild assumption on the parameters.

ASSUMPTION 3.5. We assume that for τ defined in (2.3), d defined in Assumption 2.3 and α_1, there exists a constant $C > 4$, such that

$$\frac{C}{\tau} + \alpha_1 < 1, \qquad d\alpha_1 > 2.$$

Note that the above assumption can be easily satisfied by choosing $C < \tau$ and α_1, accordingly. When the physical dependence is of exponential decay, we only need $\alpha_1 < 1$ and $d\alpha_1 > 2$.

We now estimate ϕ_{ij}. Under Assumption 3.5, by (3.1), (3.2) and Lemma 3.2, we can write

$$x_i = \sum_{j=1}^{b} \sum_{k=1}^{c} a_{jk} z_{kj}\left(\frac{i}{n}\right) + \epsilon_i + o_u(n^{-1}), \quad i = b+1, \dots, n, \tag{3.6}$$

where $z_{kj}(\frac{i}{n}) := \alpha_k(\frac{i}{n}) x_{i-j}$. In view of (3.6), we can use the ordinary least square (OLS) method to estimate the coefficients a_{jk}. Denote the vector $\boldsymbol{\beta} \in \mathbb{R}^{bc}$ by $\boldsymbol{\beta}_s = a_{j_s,k_s}$, where $j_s = \lfloor \frac{s}{c} \rfloor + 1$, $k_s = s - \lfloor \frac{s}{c} \rfloor \times c$. Similarly, we define $\mathbf{y}_i \in \mathbb{R}^{bc}$ by letting $\mathbf{y}_{is} = z_{k_s,j_s}(\frac{i}{n})$. Furthermore, we denote Y^* as the $bc \times (n-b)$ design matrix of (3.6) whose columns are $\mathbf{y}_i$, $i = b+1, \dots, n$. We also denote by $\mathbf{x} \in \mathbb{R}^{n-b}$ the vector of $x_{b+1}, \dots, x_n$. Hence, the OLS estimator for $\boldsymbol{\beta}$ can be written as $\widehat{\boldsymbol{\beta}} = (Y^*Y)^{-1}Y^*\mathbf{x}$.

Recall $\mathbf{x}_i = (x_{i-1}, \dots, x_{i-b})^* \in \mathbb{R}^b$. Denote $X = (\mathbf{x}_{b+1}, \dots, \mathbf{x}_n) \in \mathbb{R}^{b \times (n-b)}$ and the matrices $E_i \in \mathbb{R}^{(n-b)\times(n-b)}$ such that $(E_i)_{st} = 1$, when $s = t = i - b$ and $(E_i)_{st} = 0$ otherwise. As a consequence we can write

$$Y^* = \sum_{i=b+1}^{n} \left(X \otimes \mathbf{b}\left(\frac{i}{n}\right)\right) E_i. \tag{3.7}$$

Observe that

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\frac{Y^*Y}{n}\right)^{-1} \frac{Y^*\boldsymbol{\epsilon}}{n} + o_{\mathbb{P}}(n^{-1}), \tag{3.8}$$

where $\boldsymbol{\epsilon} \in \mathbb{R}^{n-b}$ consists of $\epsilon_{b+1}, \dots, \epsilon_n$ and the error is entrywise. We decompose $\boldsymbol{\beta}$ into b blocks by denoting $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^*, \dots, \boldsymbol{\beta}_b^*)^*$, where each $\boldsymbol{\beta}_i \in \mathbb{R}^c$. Similarly, we can decompose $\widehat{\boldsymbol{\beta}}$. Therefore, our sieve estimator can be written as $\widehat{\phi}_j(\frac{i}{n}) = \widehat{\boldsymbol{\beta}}_j^* \mathbf{b}(\frac{i}{n})$, and it satisfies that

$$\phi_j\left(\frac{i}{n}\right) - \widehat{\phi}_j\left(\frac{i}{n}\right) = (\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j)^* \mathbf{b}\left(\frac{i}{n}\right). \tag{3.9}$$

We impose the following assumption on the derivative of the basis functions which is also used in [7], Assumption 4.

ASSUMPTION 3.6. There exist constants $\omega_1, \omega_2 \geq 0$ such that

$$\sup_t \|\nabla \mathbf{b}(t)\| \leq C n^{\omega_1} c^{\omega_2}, \qquad C > 0 \text{ is some constant.}$$

The above assumption is satisfied by many of the widely used basis functions. For instance, we can choose $\omega_1 = 0$, $\omega_2 = \frac{1}{2}$ for trigonometric polynomials, spline series, orthogonal wavelets and weighted Chebyshev polynomials. For more examples satisfying this assumption, we refer to [7], Section 2.1.

THEOREM 3.7. *Under Assumptions* 2.2, 2.3, 3.3, 3.5 *and* 3.6, *we have*

$$\sup_{i>b, j\le b}\left|\phi_j\left(\frac{i}{n}\right)-\widehat{\phi}_j\left(\frac{i}{n}\right)\right| = O_{\mathbb{P}}\left(\zeta_c\sqrt{\frac{\log n}{n}}+n^{-d\alpha_1}\right).$$

By carefully choosing $c = O(n^{\alpha_1})$, we show that $\widehat{\phi}_j(\frac{i}{n})$ are consistent estimators for $\phi_j(\frac{i}{n})$ uniformly in i for all $j \le b$ in Theorem 3.7. Denote $\zeta_c := \sup_i \|\mathbf{b}(\frac{i}{n})\|$, as discussed in [1], Section 3; we can write $\zeta_c = O(n^{\alpha_1^*})$, where $\alpha_1^* = \frac{1}{2}\alpha_1$ for trigonometric polynomials, spline series, orthogonal wavelets and weighted orthogonal Chebyshev polynomials. And $\alpha_1^* = \alpha_1$ for Legendre orthogonal polynomials. Further, by choosing $\alpha_1^* = \frac{1}{2}\alpha_1$, we can show our estimators attain the optimal minimax convergent rate $(n/(\log n))^{-d/(2d+1)}$ for nonparametric regression established by Stone in [25].

COROLLARY 3.8. *Under Assumptions* 2.2, 2.3, 3.3, 3.5 *and* 3.6, *using the trigonometric polynomials, spline series, orthogonal wavelets and weighted orthogonal Chebyshev polynomials, when* $c = O((n/(\log n))^{1/(2d+1)})$, *we have*

$$\sup_{i>b, j\le b}\left|\phi_j\left(\frac{i}{n}\right)-\widehat{\phi}_j\left(\frac{i}{n}\right)\right| = O_{\mathbb{P}}\left((n/(\log n))^{-d/(2d+1)}\right).$$

3.2. *Estimating ϕ_{ij} for $i \le b$.* It is notable that by Lemma 2.8, when i, j are less or equal to b, we cannot use the estimators derived from Section 3.1. Instead, a different series of least squares linear regressions should be used. For instance, to estimate ϕ_{21}, we use the following regression equations:

$$x_k = \phi_{k1}x_{k-1} + \xi_{k,2}, \quad k = 2, 3, \dots, n.$$

Note that $\xi_{2,2} = \epsilon_2$. Due to the local stationarity assumption, there exists a smooth function f_{21}, such that $\phi_{k1} \approx f_{21}(\frac{k}{n})$, $k = 2, 3, \dots, n$. Here, f_{21} can be estimated using the sieve method as described by the previous discussions and ϕ_{21} can be estimated by $\widehat{f}_{21}(2/n)$. Generally, for each fixed $i \le b$, to estimate $\boldsymbol{\phi}_i$ we make use of the following predictions:

$$x_k = \sum_{j=1}^{i-1}\lambda_{ij}^k x_{k-j} + \xi_{k,i}, \quad k = i, i+1, \dots, n, \tag{3.10}$$

where $\boldsymbol{\lambda}_i^k = (\lambda_{i1}^k, \dots, \lambda_{i,i-1}^k)$ are the coefficients of the best linear prediction, using the $i-1$ predecessors. Note that $\boldsymbol{\lambda}_i^i = \boldsymbol{\phi}_i$. Using the Yule–Walker equation, we find $\boldsymbol{\lambda}_i^k = \Omega_i^k\boldsymbol{\gamma}_i^k$, where $\Omega_i^k = [\mathrm{Cov}(\mathbf{x}_i^k, \mathbf{x}_i^k)]^{-1}$, $\boldsymbol{\gamma}_i^k = \mathrm{Cov}(\mathbf{x}_i^k, x_k)$ and $\mathbf{x}_i^k = (x_{k-1}, \dots, x_{k-i+1})$. Due to Assumption 2.3, we define $\mathbf{f}_i^k = (f_1^i(\frac{k}{n}), \dots, f_{i-1}^i(\frac{k}{n}))$ by

$$\mathbf{f}_i^k = \widetilde{\Omega}_i^k\widetilde{\gamma}_i^k, \tag{3.11}$$

with $\widetilde{\Omega}_i^k$, $\widetilde{\boldsymbol{\gamma}}_i^k$ defined by $\widetilde{\Omega}_i^k = [\mathrm{Cov}(\widetilde{\mathbf{x}}_i^k, \widetilde{\mathbf{x}}_i^k)]^{-1}$, $\widetilde{\boldsymbol{\gamma}}_i^k = \mathrm{Cov}(\widetilde{\mathbf{x}}_i^k, \widetilde{x}_k)$, where $\widetilde{\mathbf{x}}_{i,j}^k = G(\frac{k}{n}, \mathcal{F}_{i-j})$. The following lemma shows that $\lambda_{i,j}^k$ can be approximated by a smooth function $f_j^i(t)$.

LEMMA 3.9. *Under Assumptions* 2.2, 2.3 *and the assumption* (2.7), *for each fixed $i \le b$ and for any $j \le i-1$, $f_j^i(t)$ are C^d functions on* $[0, 1]$. *Furthermore, for some constant* $C > 0$, *we have*

$$\sup_{k\ge i}\left|\lambda_{ij}^k - f_j^i\left(\frac{k}{n}\right)\right| \le C\left(n^{-1+2/\tau}+n^{-d\alpha_1}\right).$$

In particular, when $k = i$, we have

$$\left|\phi_{ij} - f_j^i\left(\frac{i}{n}\right)\right| \le C\left(n^{-1+2/\tau} + n^{-d\alpha_1}\right), \quad j < i \le b. \tag{3.12}$$

Therefore, the rest of the work leaves to estimate the functions $f_j^i(t)$, $j < i \le b$, using sieve approximation by denoting $f_j^i(t) = \sum_{k=1}^c d_{jk}\alpha_k(t) + O(c^{-d})$, where we recall Lemma 3.2. Then, the above sieve expansion is plugged into (3.10). An OLS regression is then used to estimate the d_{jk}. We denote the OLS estimator of $f_j^i(\frac{i}{n})$ as $\widehat{f}_j^i(\frac{i}{n}) = \sum_{k=1}^c \widehat{d}_{jk}\alpha_k(\frac{i}{n})$.

THEOREM 3.10. *Under Assumptions 2.2, 2.3, 3.3, 3.5 and 3.6, we have*

$$\sup_{i \le b, j < i}\left|f_j^i\left(\frac{i}{n}\right) - \widehat{f}_j^i\left(\frac{i}{n}\right)\right| = O_{\mathbb{P}}\left(\zeta_c\sqrt{\frac{\log n}{n}} + n^{-d\alpha_1}\right).$$

Similar to the discussion of Corollary 3.8, using the trigonometric polynomials, spline series, orthogonal wavelets and weighted orthogonal Chebyshev polynomials and setting $c = O((n/(\log n))^{1/(2d+1)})$, we can obtain the optimal minimax convergent rate from Theorem 3.10.

3.3. *Sieve estimation for noise variances.* This subsection is devoted to the estimation of $\{\sigma_i^2\}_{i=1}^n$. We discuss the cases for $i > b$ and $i \le b$ separately. For $i > b$, denote $\epsilon_i^b = x_i - \sum_{j=1}^b \phi_{ij}x_{i-j}$ and $(\sigma_i^b)^2 = \mathbb{E}(\epsilon_i^b)^2$. σ_i can be well approximated using σ_i^b by the following lemma.

LEMMA 3.11. *Under Assumptions 2.2 and 2.3 and the assumption (2.7), for $i > b$ and some constant $C > 0$, we have*

$$\sup_{i>b}\left|\sigma_i^2 - (\sigma_i^b)^2\right| \le Cn^{-2+2/\tau}.$$

Furthermore, denote $g(\frac{i}{n}) = \mathbb{E}(x_i - \sum_{j=1}^b \phi_{ij}G(\frac{i}{n}, \mathcal{F}_{i-j}))^2$, we have

$$\sup_{i>b}\left|(\sigma_i^b)^2 - g\left(\frac{i}{n}\right)\right| \le Cn^{-1+4/\tau}.$$

Finally, $g(\frac{i}{n}) \in C^d([0, 1])$.

Lemma 3.11 indicates that $\{\sigma_i^2\}_{i \ge b}$ can be well approximated by a C^d function $g(\cdot)$. Denote $r_i^b = (\epsilon_i^b)^2$; it is notable that r_i^b cannot be observed directly. Instead, we use $\widehat{r}_i^b = \widehat{\epsilon}_i^2$, where

$$\widehat{\epsilon}_i = x_i - \sum_{j=1}^b\sum_{k=1}^c \widehat{a}_{jk}z_{kj}\left(\frac{i}{n}\right), \quad i = b+1, \dots, n. \tag{3.13}$$

By Theorem 3.7 and Assumption 3.5 we conclude that

$$\sup_{i>b}\left|r_i^b - \widehat{r}_i^b\right| = O_{\mathbb{P}}\left(n^{2/\tau}\left(\zeta_c\sqrt{\frac{\log n}{n}} + n^{-d\alpha_1}\right)\right). \tag{3.14}$$

Invoking Lemma 3.2 and Assumption 3.5, for $i > b$, we can therefore utilize the method of sieves and write

$$\widehat{r}_i^b = \sum_{k=1}^c d_k\alpha_k\left(\frac{i}{n}\right) + \omega_i^b + O_{\mathbb{P}}\left(n^{2/\tau}\left(\zeta_c\sqrt{\frac{\log n}{n}} + n^{-d\alpha_1}\right)\right). \tag{3.15}$$

The coefficients d_k's are then estimated via OLS. Similar to Lemma 2.9, we can show that the physical dependence measure of ω_i^b is also of polynomial decay. Therefore, the OLS estimator for $\boldsymbol{\alpha} = (d_1, \dots, d_c)^*$ can be written as $\widehat{\boldsymbol{\alpha}} = (W^* W)^{-1} W^* \widehat{\mathbf{r}}$, where W^* is an $c \times (n-b)$ matrix whose ith column is $(\alpha_1(\frac{i+b}{n}), \dots, \alpha_c(\frac{i+b}{n}))^*$, $i = 1, 2, \dots, n-b$ and $\widehat{\mathbf{r}}$ is an $\mathbb{R}^{n-b}$ containing $\widehat{r}_{b+1}^b, \dots, \widehat{r}_n^b$. We have the following consistency result. Denote $\widehat{\sigma}_i^2 = \widehat{g}(i/n) = \sum_{k=1}^c \widehat{d}_k \alpha_k(i/n)$, $i > b$.

THEOREM 3.12. *Suppose Assumptions 2.2, 2.3, 3.3, 3.5 and 3.6 hold true. Then, we have*

$$\sup_{i>b} \left|\widehat{\sigma}_i^2 - \sigma_i^2\right| = O_{\mathbb{P}}\left(n^{2/\tau}\left(\zeta_c \sqrt{\frac{\log n}{n}} + n^{-d\alpha_1}\right)\right). \tag{3.16}$$

Finally, we study the estimation of σ_i^2, $i = 1, 2, \dots, b$ which enjoys the same discussion as in Section 3.2. Recall $\xi_{k,i}$ defined in (3.10). Denote $(\sigma_{k,i}(\xi))^2 = \mathbb{E}(\xi_{k,i})^2$; using a discussion similar to Lemma 3.11, we can find a smooth function g^i, such that $\sup_k \sup_{i \leq b} |(\sigma_{k,i}(\xi))^2 - g^i(\frac{k}{n})| \leq O(n^{-1+4\tau})$. In particular, we can use $g^i(\frac{i}{n})$ to estimate σ_i^2. When $i = 1$, we need to estimate the variance function of x_1.

The rest of the work leaves to estimate $g^i(t)$ using the sieve method similar to (3.15) for $i \leq b$, where we replace the errors with $\widehat{r}_k^i$, $k = i, \dots, n$. Here, $\widehat{r}_k^i$ is defined as

$$\widehat{r}_k^i := \left(x_i - \sum_{j=1}^{i-1} \widehat{f}_j^i\left(\frac{k}{n}\right) x_{i-j}\right)^2, \quad k = i, i+1, \dots, n. \tag{3.17}$$

Then, for $i \leq b$, we can estimate $\widehat{g}^i(\frac{i}{n})$ using the method of sieves similarly, except that the dimension of W^* is $c \times (n+1-i)$. The results are summarized in the following theorem. Denote $\widehat{\sigma}_i^2 = \widehat{g}^i(i/n)$.

THEOREM 3.13. *Suppose Assumptions 2.2, 2.3, 3.3, 3.5 and 3.6 hold true. Then, we have*

$$\sup_{i\leq b} \left|\widehat{\sigma}_i^2 - \sigma_i^2\right| = O_{\mathbb{P}}\left(n^{2/\tau}\left(\zeta_c \sqrt{\frac{\log n}{n}} + n^{-d\alpha_1}\right)\right). \tag{3.18}$$

In the finite sample case, for positiveness, we suggest simply choose

$$(\widehat{\sigma}_i^*)^2 = \begin{cases} \widehat{\sigma}_i^2 & \text{if } \widehat{\sigma}_i^2 > 0; \\ \dfrac{1}{n} & \text{if } \widehat{\sigma}_i^2 \leq 0. \end{cases} \tag{3.19}$$

Since n^{-1} is much smaller than the right-hand side of (3.16) and (3.18), (3.19) will not influence the results in Theorems 3.12 and 3.13.

3.4. *Precision matrix estimation.* From the Cholesky decomposition of (1.2), it is natural to choose $\widehat{\Omega} := \widehat{\Phi}^* \widehat{\widehat{\mathbf{D}}} \widehat{\Phi}$ as our estimator for the precision matrix. As we discussed in the previous sections, here $\widehat{\Phi}$ is a lower triangular matrix whose diagonal entries are all ones. For the off-diagonal entries, when $i > b$ and $j \leq b$, its $(i, i-j)$th entry is $-\widehat{\phi}_j(\frac{i}{n})$ defined in Section 3.1. And when $i \leq b$, $\Phi_{i,i-j}$ is estimated using $-\widehat{f}_j^i(\frac{i}{n})$ from Section 3.2. All other entries of $\widehat{\Phi}$ are set to be zeros. Finally, $\widehat{\widehat{\mathbf{D}}}$ is a diagonal matrix with entries $\{(\widehat{\sigma}_i^*)^{-2}\}$ estimated from (3.19). Observe that $\widehat{\Omega}$ is always positive definite.

We now discuss the computational complexity of estimating Ω. It is easy to see that when $i > b$, the number of regressors is bc and length of observation is $n - b$. Hence, the computational complexity of the least squares regression is $O(n(bc)^2)$. Similar discussion can be applied for $i \leq b$, and we hence conclude that the computational complexity for estimating $\widehat{\Omega}$ is of the order $O(nb^3c^2)$. As a result the computational complexity of our estimation is adaptive to the smoothness of the underlying data generating mechanism and the decay rate of temporal dependence. In the best scenario, when assumption (2.15) holds and $\gamma(t, j) \in C^{\infty}([0, 1])$, our procedure only requires $O(n \log^5 n)$ operations.

In the following we shall control the estimation error between Ω and $\widehat{\Omega}$. We first observe that, as $\det(\Phi\Phi^*) = \det(\widehat{\Phi}\widehat{\Phi}^*) = 1$, combining with Assumption 2.2, there exist some constants $C_1, C_2 > 0$, such that

$$C_1 \leq \lambda_{\min}(\Phi\Phi^*) \leq \lambda_{\max}(\Phi\Phi^*) \leq C_2.$$

Similar results hold for $\widehat{\Phi}\widehat{\Phi}^*$.

THEOREM 3.14. *Under Assumptions* 2.2, 2.3, 3.3, 3.5 *and* 3.6, *we have*

$$\|\Omega - \widehat{\Omega}\| = O_{\mathbb{P}}\left(n^{4/\tau}\left(n^{-d\alpha_1} + \zeta_c\sqrt{\frac{\log n}{n}}\right)\right). \tag{3.20}$$

Recall that $\|\cdot\|$ denotes the operator norm of a matrix. It can be seen from the above theorem that the estimation accuracy of precision matrices depends on the decay rate of the dependence and the smoothness of the covariance functions. The estimation accuracy gets higher for time series with smoother covariance functions and faster decay speed of dependence.

REMARK 3.15. Under the assumption (2.15), when we apply Gershgorin circle theorem to our proof, we only need $O(\log n)$ matrix entries to bound the error terms. Hence, we can change (3.20) to

$$\|\Omega - \widehat{\Omega}\| = O_{\mathbb{P}}\left(\log^2 n\left(n^{-d\alpha_1} + \zeta_c\sqrt{\frac{\log n}{n}}\right)\right).$$

In the best scenario, where the dependence is exponentially decaying and $\phi_j(\cdot)$ and $g(\cdot)$ are infinitely differentiable, following the same arguments as those in the proof of Theorem 3.14, it is easy to show the convergence rate of $\widehat{\Omega}$ is of the order $\log^3 n/\sqrt{n}$ which is almost as fast as parametrically estimating a single parameter from i.i.d. samples.

4. Testing the structure of the precision matrices.

An important advantage of our methodology is that we can test many structural assumptions of the precision matrices using some simple statistics in terms of the entries of $\widehat{\Phi}$.

4.1. *Test statistics.*

In this subsection we focus on discussing two fundamental tests in nonstationary time series analysis. One of those is to test whether the observed samples are from a nonstationary white noise process in the sense that $\text{Cov}(x_i, x_j) = \delta_{ij}\sigma_i^2$, where δ_{ij} is the Dirac delta function such that $\delta_{ij} = 1$ when $i = j$ and $\delta_{ij} = 0$ otherwise. Note that we allow heteroscedasticity by assuming that the variance of x_i changes over time. Formally, we would like to test

$$\mathbf{H}_0^1 : \{x_i\} \text{ is a nonstationary white noise process.}$$

Recall (2.12). Under $\mathbf{H}_0^1$, we shall have that $\phi_j(\frac{i}{n})$ are all zeros. Therefore, our estimation $\widehat{\phi}_j(\frac{i}{n})$ should be small for all pairs $i, j, i \neq j$. We hence use

$$T_1^* = \sum_{j=1}^{b} \int_0^1 \widehat{\phi}_j^2(t)\,dt. \tag{4.1}$$

The second hypothesis of interest is whether the precision matrices are banded. In our setup the Cholesky decomposition provides a convenient way to test the bandedness. Formally, for any $k_0 \equiv k_0(n) < b$, we are interested in testing the following hypothesis:

$$\mathbf{H}_0^2 : \text{The precision matrix of}\{x_i\}\text{is } k_0\text{-banded.}$$

Due to (1.2), as Ω is strictly positive definite, the Cholesky decomposition is unique. Therefore, we conclude that Φ is also k_0-banded using the discussion in [23], Section 2. Furthermore, under $\mathbf{H}_0^2$ we have that $\phi_j(\frac{i}{n}) = 0$, for $j > k_0$. Therefore, it is natural for us to use the following statistic:

$$T_2^* = \sum_{j=k_0+1}^{b} \int_0^1 \widehat{\phi}_j^2(t)\,dt.$$

It is notable that both T_1^* and T_2^* can be written into summations of quadratic forms under the null hypothesis. For instance, for T_1^* under $\mathbf{H}_0^1$, we have

$$\widehat{\phi}_j^2(t) = \big(\widehat{\phi}_j(t) - \phi_j(t)\big)^2.$$

For any fixed $j \leq b$, we have

$$\int_0^1 \big(\phi_j(t) - \widehat{\phi}_j(t)\big)^2\,dt = \sum_{k=1}^{c} (\widehat{a}_{jk} - a_{jk})^2 + O\big(n^{-d\alpha_1}\big).$$

It can be seen from the above equation that the order of smoothness and number of basis functions are important to our analysis. Under Assumption 3.5 we can see that the error $O(n^{-d\alpha_1})$ is negligible. Recall (3.8). It is easy to see that for $\Sigma := \Sigma^b$ defined in (3.3), we have that

$$\sum_{j=1}^{b} \int_0^1 \big(\phi_j(t) - \widehat{\phi}_j(t)\big)^2\,dt = \frac{\boldsymbol{\epsilon}^* Y}{n} \Sigma^{-1} \sum_{j=1}^{b} A_j^* A_j \Sigma^{-1} \frac{Y^* \boldsymbol{\epsilon}}{n} + o_{\mathbb{P}}(1), \tag{4.2}$$

where $A_j \in \mathbb{R}^{bc}$ is a diagonal block matrix whose jth diagonal block being the identity matrix and zeros otherwise. Therefore, the investigation of T_1^* boils down to the analysis of quadratic forms of a bc dimensional locally stationary time series $\{Y^* \boldsymbol{\epsilon}\}$.

REMARK 4.1. In the current paper we focus our discussion on the white noise and bandedness tests. However, many other tests could be performed using our framework. Recall (1.1). Another possible test is to check whether the time series $\{x_i\}$ is correlation stationary, that is, $\operatorname{corr}(x_i, x_j) = r(|i - j|)$ for some function r. In this situation the null hypothesis can be formulated as

$$\mathbf{H}_0^3 : \phi_{ij} \equiv \phi_j, \quad \text{for all } j.$$

Test of $\mathbf{H}_0^3$ is possible through our framework provided that an appropriate and theoretically tractable test statistic is constructed. We shall investigate this in some future work.

4.2. *Diverging dimensional Gaussian approximation.* As we have seen from the previous subsection, both test statistics are involved with high-dimensional quadratic forms. Observe that the distribution of quadratic forms of Gaussian vectors can be derived using Lindeberg's central limit theorem. Hence, our case can be tackled if we could establish a Gaussian approximation to the quadratic form (4.2) of general nonstationary time series. In this subsection we will prove a Gaussian approximation result for the quadratic form $\mathbf{Z}^*E\mathbf{Z}$, where $\mathbf{Z} := \frac{\boldsymbol{\epsilon}^* Y}{\sqrt{n}} \in \mathbb{R}^{bc}$ and E is a bounded positive semidefinite matrix. Denote $p = bc$ and $\mathbf{z}_i = (z_{i1}, \dots, z_{ip})^*$, where

$$z_{is} = x_{i-\bar{s}-1}\epsilon_i \alpha_{s'}\left(\frac{i}{n}\right), \quad \bar{s} = \left\lfloor \frac{s}{c} \right\rfloor, s' = s - \bar{s}c, i \geq b+1. \tag{4.3}$$

As a consequence we can write $\mathbf{Z} := (\mathbf{Z}_1, \dots, \mathbf{Z}_p) = \frac{1}{\sqrt{n}} \sum_{i=b+1}^n \mathbf{z}_i$. Denote $\mathbf{U} = \frac{1}{\sqrt{n}} \times \sum_{i=b+1}^n \mathbf{u}_i$, where $\{\mathbf{u}_i\}_{i=b+1}^n$ are centered Gaussian random vectors independent of $\{\mathbf{z}_i\}_{i=b+1}^n$ and preserve their covariance structure. Our task is to control the following Kolmogorov distance:

$$\rho := \sup_{x \in \mathbb{R}} \left| P\left(R^z \leq x\right) - P\left(R^u \leq x\right)\right|, \tag{4.4}$$

where $R^z = \mathbf{Z}^*E\mathbf{Z}$, $R^u = \mathbf{U}^*E\mathbf{U}$.

Denote $\xi_c := \sup_{i,t} |\alpha_i(t)|$. It is notable that ξ_c can be well controlled for the commonly used basis functions. For instance, for the trigonometric polynomials and the weighted Chebyshev polynomials of the first kind, $\xi_c = O(1)$, and, for orthogonal wavelet, $\xi_c = O(\sqrt{c})$. The following theorem establishes the Gaussian approximation for high-dimensional quadratic forms under physical dependence.

THEOREM 4.2. *Suppose Assumptions 2.2, 2.3, 3.3, 3.5 and 3.6 hold true. Then, for some constant $C > 0$, we have*

$$\rho \leq Cl(n),$$

where $l(n)$ is defined as

$$l(n) = \psi^{-1/2} + \xi_c p \psi^{\frac{q}{q+1}} M^{\frac{q(-\tau+1)}{q+1}} + \xi_c M_x^{-1} \psi^2 p^4 + \frac{M^2}{\sqrt{n}} \psi^3 p^6$$
$$+ p\psi\left(\frac{\xi_c^{1/2}}{M_x^{5/6}} + \frac{\sqrt{M}}{M_x^3}\right)\sqrt{\log \frac{p}{\gamma}} + \gamma,$$

where M_x, ψ, $M \to \infty$ and $\gamma \to 0$ when $n \to \infty$.

4.3. *Asymptotic normality of test statistics.* With the above preparation we now derive the distributions for the test statistics T_1^* and T_2^* defined in Section 4.1. First of all, under $\mathbf{H}_0^1$, we have

$$nT_1^* = \sum_{j=1}^b \sum_{k=1}^c \widehat{a}_{jk}^2 = \widehat{\boldsymbol{\beta}}^* \widehat{\boldsymbol{\beta}} = \frac{\boldsymbol{\epsilon}^* Y}{\sqrt{n}} \Sigma^{-2} \frac{Y^* \boldsymbol{\epsilon}}{\sqrt{n}} + o_{\mathbb{P}}(1), \tag{4.5}$$

where we recall (3.8). We can analyze T_2^* in the same way using

$$\begin{aligned} nT_2^* &= \sum_{j=k_0+1}^b \sum_{k=1}^c \widehat{a}_{jk}^2 \\ &= (A\widehat{\boldsymbol{\beta}})^*(A\widehat{\boldsymbol{\beta}}) \\ &= \frac{\boldsymbol{\epsilon}^* Y}{\sqrt{n}} \Sigma^{-1} A^* A \Sigma^{-1} \frac{Y^* \boldsymbol{\epsilon}}{\sqrt{n}} + o_{\mathbb{P}}(1), \end{aligned}$$

where $A \in \mathbb{R}^{bc \times bc}$ is a block diagonal matrix with the nonzero block being the lower $(b - k_0)c \times (b - k_0)c$ major part.

Note that $\frac{1}{\sqrt{n}} Y^* \boldsymbol{\epsilon} \in \mathbb{R}^p$ is a block vector with size c, where the jth entry of the i-block is $\frac{1}{\sqrt{n}} \sum_{k=b+1}^n x_{k-i} \epsilon_k \alpha_j(\frac{k}{n})$. We can therefore rewrite it as

$$\frac{1}{\sqrt{n}} Y^* \boldsymbol{\epsilon} = \frac{1}{\sqrt{n}} \sum_{i=b+1}^{n} \mathbf{h}_i \otimes \mathbf{b}\left(\frac{i}{n}\right),$$

where $\mathbf{h}_i = \mathbf{x}_i \epsilon_i$. For $i > b$, $\mathbf{h}_i$ can be regarded as a locally stationary time series, that is, $\mathbf{h}_i = \mathbf{U}(\frac{i}{n}, \mathcal{F}_i)$. Denote the long-run covariance matrix of $\{\mathbf{h}_i\}$ as

$$\bar{\Delta}(t) = \sum_{j=-\infty}^{\infty} \operatorname{Cov}(\mathbf{U}(t, \mathcal{F}_j), \mathbf{U}(t, \mathcal{F}_0)),$$

and we further define

$$\Delta = \int_0^1 \bar{\Delta}(t) \otimes (\mathbf{b}(t)\mathbf{b}^*(t))\, dt. \tag{4.6}$$

For $k \in \mathbb{N}$, denote

$$f_k = (\operatorname{Tr}[(\Delta^{1/2} \Sigma^{-2} \Delta^{1/2})^k])^{1/k}, \qquad g_k = (\operatorname{Tr}[(\Delta^{1/2} \Sigma^{-1} A^* A \Sigma^{-1} \Delta^{1/2})^k])^{1/k}.$$

We next summarize the limiting distributions of T_1^* and T_2^*.

THEOREM 4.3. *Suppose Assumptions* 2.2, 2.3, 3.3, 3.5 *and* 3.6 *hold true. Then, if* $l(n) \to 0$, *we have*

(1) *Under* $\mathbf{H}_0^1$, *we have*

$$\frac{nT_1^* - f_1}{f_2} \Rightarrow \mathcal{N}(0, 2).$$

Furthermore, there exist some positive constants $c_i, C_i, i = 1, 2$, *such that*

$$c_1 \le \frac{f_1}{bc} \le C_1, \qquad c_2 \le \frac{f_2}{\sqrt{bc}} \le C_2.$$

(2) *Under* $\mathbf{H}_0^2$, *we have*

$$\frac{nT_2^* - g_1}{g_2} \Rightarrow \mathcal{N}(0, 2).$$

Furthermore, there exist some positive constants $w_i, W_i, i = 1, 2$, *such that*

$$w_1 \le \frac{g_1}{(b - k_0)c} \le W_1, \qquad w_2 \le \frac{g_2}{\sqrt{(b - k_0)c}} \le W_2.$$

Finally, we discuss the local power of our tests. We will only focus on the white noise test and similar discussion can be applied to the bandedness test. Consider the local alternative

$$\mathbf{H}_a : \frac{n \sum_{j=1}^{\infty} \int_0^1 \gamma^2(t, j)\, dt}{\sqrt{bc}} \to \infty.$$

The following proposition states that, under $\mathbf{H}_a$, the power of our test will asymptotically be 1.

PROPOSITION 4.4. *Under Assumptions 2.2, 2.3, 3.3, 3.5 and 3.6, when the alternative hypothesis $\mathbf{H}_a$ holds true, for any given significant level α, we have*

$$\mathbb{P}\left(\left|\frac{nT_1^* - f_1}{f_2}\right| \geq \sqrt{2}\mathcal{Z}_{1-\alpha}\right) \to 1, \quad n \to \infty,$$

where $\mathcal{Z}_{1-\alpha}$ is the $(1-\alpha)\%$ quantile of the standard normal distribution.

Proposition 4.4 states that the white noise test has asymptotic power 1 whenever $\sum_{j=1}^{\infty}\int_0^1 \gamma^2(t,j)\,dt \gg \sqrt{bc}/n$. In an interesting special case when $\int_0^1 \gamma^2(t,j_i)\,dt \gg \sqrt{bc}/(nk)$, $i=1,2,\ldots,k$, T_1^* achieves asymptotic power 1. Note that if k here is large, then we conclude that alternatives consist of many very small deviations from the null can be picked up by the $\mathcal{L}^2$ test T_1^*. On the contrary, maximum deviation or $\mathcal{L}^\infty$ norm based tests will not be sensitive to such alternatives.

4.4. *Practical implementation.* It can been seen from Theorem 4.3 that the key to implementing the tests is to estimate the covariance matrix of the high dimensional vector $\{\mathbf{x}_i\epsilon_i\}$. A disadvantage of using (4.5) is that the basis functions are mixed with $\{x_i\}$. In the present subsection we provide practical implementation by representing nT_1^* and nT_2^* into different forms in order to separate the data and the basis functions. We focus our discussion on nT_1^*.

For $i > b$, $j \leq b$, denote the vector $\mathbb{B}_j(\frac{i}{n}) \in \mathbb{R}^{bc}$ with b-blocks, where the jth block is the basis $\mathbf{b}(\frac{i}{n})$ and zeros otherwise. Therefore, for all $j \leq b$, $b < i \leq n$, we have

$$\left(\phi_j\left(\frac{i}{n}\right) - \widehat{\phi}_j\left(\frac{i}{n}\right)\right)^2 = \mathbb{B}_j^*\left(\frac{i}{n}\right)\Sigma^{-1}\frac{Y^*\boldsymbol{\epsilon}}{n}\frac{\boldsymbol{\epsilon}^*Y}{n}\Sigma^{-1}\mathbb{B}_j\left(\frac{i}{n}\right) + o_{\mathbb{P}}(1). \tag{4.7}$$

Denote $\mathbf{q}_{ij}^* = \mathbb{B}_j^*(\frac{i}{n})\Sigma^{-1} \in \mathbb{R}^{bc}$ and $\mathbf{q}_{ijk}$ as the kth block of $\mathbf{q}_{ij}$ of size c. As a consequence we can write

$$\mathbf{q}_{ij}^*\frac{Y^*\boldsymbol{\epsilon}}{n} = \frac{1}{n}\sum_{k=b+1}^{n}\mathbf{h}_k^*\mathbf{q}_{ij}^{\widetilde{k}}, \tag{4.8}$$

where we recall $\mathbf{h}_k = \epsilon_k\mathbf{x}_k$, $\mathbf{q}_{ij}^{\widetilde{k}} \in \mathbb{R}^b$ is denoted by $(\mathbf{q}_{ij}^{\widetilde{k}})_s = \mathbf{q}_{ijs}^*\mathbf{b}(\frac{k}{n})$. Denote $\mathbf{Q}_{ij} \in \mathbb{R}^{(n-b)b\times(n-b)b}$ as a block matrix with size $b \times b$ whose (k_1,k_2)th block is $\mathbf{q}_{ij}^{\widetilde{k_1}}(\mathbf{q}_{ij}^{\widetilde{k_2}})^*$. Furthermore, we denote

$$Q\left(\frac{i}{n}\right) = \sum_{j=1}^{b}\mathbf{Q}_{ij}, \qquad Q_{k_0}\left(\frac{i}{n}\right) = \sum_{j=k_0}^{b}\mathbf{Q}_{ij}.$$

By (4.8) and Theorem 4.3 it suffices to study the following quantity:

$$n^2T_1^{**} = (\Sigma_L^{1/2}\mathbf{z}_L)^*\left(\int_0^1 Q(t)\,dt\right)(\Sigma_L^{1/2}\mathbf{z}_L),$$

where Σ_L is the covariance matrix of $\mathbf{h} = (\mathbf{h}_{b+1},\ldots,\mathbf{h}_n)^*$ and $\mathbf{z}_L \sim \mathcal{N}(\mathbf{0},\mathbf{I})$, $\mathbf{I} \in \mathbb{R}^{(n-b)b}$. Similarly, we use the following statistic to study $\mathbf{H}_0^2$:

$$n^2T_2^{**} = (\Sigma_L^{1/2}\mathbf{z}_L)^*\left(\int_0^1 Q_{k_0}(t)\,dt\right)(\Sigma_L^{1/2}\mathbf{z}_L).$$

The above expressions are useful for our practical implementation as they provide us a way to separate the deterministic basis functions and the random part. Hence, we only need to estimate the covariance matrix Σ_L for $\mathbf{h}$. Next, we will provide a nonparametric estimator for Σ_L. Similar ideas have been employed to estimate the long-run covariance matrix in [40] for fixed dimensional random vectors.

We observe that the covariance matrix of $\mathbf{h}$ is a $(n-b) \times (n-b)$ block matrix with block size b. We first consider the diagonal part, where each block Λ_k is the covariance matrix of $\mathbf{h}_k, k = b+1, \dots, n$. Recall that we can write $\{\mathbf{h}_k\}$ into a sequence of locally stationary time series $\{\mathbf{U}(\frac{k}{n}, \mathcal{F}_k)\}_{k=b+1}^n$. Denote

$$\Lambda(t, j) = \operatorname{Cov}\big(\mathbf{U}(t, \mathcal{F}_0), \mathbf{U}(t, \mathcal{F}_j)\big).$$

The following lemma shows that Λ_{kk}, which is the kth diagonal block of Σ_L, can be well estimated by $\Lambda(\frac{k}{n}, 0)$ for any $k > b$.

LEMMA 4.5. *Under Assumptions 2.2 and 2.3 and the assumption (2.7), we have*

$$\sup_{k>b}\left\|\Lambda\left(\frac{k}{n}, 0\right) - \Lambda_{kk}\right\| = O\big(n^{-1+4/\tau}\big).$$

Next, we consider the upper-off-diagonal blocks. For any $b < k \leq n-b+1$, we find that for $j > b+k$, for some constant $C > 0$, we have

$$\|\Lambda_{kj}\| \leq C(j-b)^{-\tau+1}, \tag{4.9}$$

where we use a discussion similar to Lemma 2.6 and Gershgorin circle theorem. As a consequence we only need to estimate the blocks Λ_{kj} for $k < j \leq k+b$. Similar to Lemma 4.5, we have

$$\left\|\Lambda\left(\frac{k}{n}, j\right) - \Lambda_{kj}\right\| = O\big(n^{-1+4/\tau}\big).$$

Hence, we propose to estimate $\Lambda(t, j), 0 \leq j \leq b$ using the kernel estimators. For a smooth symmetric density function K_h defined on $\mathbb{R}$ supported on $[-1, 1]$, where $h \equiv h_n$ is the bandwidth such that $h \to 0$, $nh \to \infty$. We write

$$\widehat{\Lambda}(t, j) = \frac{1}{nh}\sum_{k=b+1}^{n-j} K\left(\frac{k/n - t}{h}\right)\mathbf{h}_k\mathbf{h}_{k+j}^*, \quad 0 \leq j \leq b.$$

Finally we define $\widehat{\Sigma}_L$ as the estimator by setting its blocks

$$(\widehat{\Sigma}_L)_{kk} = \widehat{\Lambda}\left(\frac{b+k}{n}, 0\right), \qquad (\widehat{\Sigma}_L)_{kj} = \widehat{\Lambda}\left(\frac{k+b}{n}, j\right), \tag{4.10}$$

and zeros otherwise, where $k = 1, 2, \dots, n-b$, $k < j \leq k+b$. We can prove that our estimators are consistent under mild assumptions.

THEOREM 4.6. *Under Assumptions 2.2 and 2.3 and the assumption (2.7), let $h \to 0$ and $nh \to \infty$, for $j = 0, 1, 2, \dots, b$, we have*

$$\sup_t \|\Lambda(t, j) - \widehat{\Lambda}(t, j)\| = O_{\mathbb{P}}\left(b\left(\frac{1}{\sqrt{nh}} + h^2\right)\right). \tag{4.11}$$

As a consequence we have

$$\|\Sigma_L - \widehat{\Sigma}_L\| = O_{\mathbb{P}}\left(b^2\left(\frac{1}{\sqrt{nh}} + h^2\right)\right). \tag{4.12}$$

In practice, the true ϵ_i is unknown, and we have to use $\widehat{\epsilon}_i$ defined in (3.13). We then define

$$\widetilde{\Lambda}(t, j) = \frac{1}{nh} \sum_{k=b+1}^{n-j} K\left(\frac{k/n - t}{h}\right) \widehat{\mathbf{h}}_k \widehat{\mathbf{h}}_{k+j}^*, \quad 0 \leq j \leq b,$$

where $\widehat{\mathbf{h}}_k := \mathbf{x}_k \widehat{\epsilon}_k$. Similarly, we can define the estimation $\widetilde{\Sigma}_L$. The analog of Theorem 4.6 is the following result.

THEOREM 4.7. *Under the assumptions of Theorem* 4.6 *and Assumptions* 3.3, 3.5 *and* 3.6, *we have*

$$\sup_t \| \Lambda(t, j) - \widetilde{\Lambda}(t, j) \| = O_{\mathbb{P}}\left(b\left(\frac{1}{\sqrt{nh}} + h^2 + \theta_n\right)\right),$$

where θ_n is defined as

$$\theta_n = \sqrt{\frac{b}{nh}}\left(\zeta_c \sqrt{\frac{\log n}{n}} + n^{-d\alpha_1}\right).$$

As a consequence we have

$$\| \Sigma_L - \widetilde{\Sigma}_L \| = O_{\mathbb{P}}\left(b^2\left(\frac{1}{\sqrt{nh}} + h^2 + \theta_n\right)\right).$$

By Theorems 4.3, 4.6 and 4.7, we now propose the following practical procedure to test $\mathbf{H}_0^1$ (the implementation for $\mathbf{H}_0^2$ is similar):

1. For $j = 1, 2, \ldots, b, i = b + 1, \ldots, n$, estimate Σ^{-1} using $n(Y^*Y)^{-1}$ and calculate $\mathbf{Q}_{ij}$ by the definitions.
2. Choose the tuning parameters b and c according to Section 4.5.
3. Estimate Σ_L using (4.10) from the samples $\{\widehat{\mathbf{h}}_k\}_{k=b+1}^n$.
4. Generate B (say 2000) i.i.d. copies of Gaussian random vectors $\mathbf{z}_i, i = 1, 2, \ldots, B$. Here, $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. For each $k = 1, 2, \ldots, B$, calculate the following Riemann summation:

$$T_k^1 = \frac{1}{n^2} \sum_{j=1}^{b} \sum_{i=b+1}^{n} (\widehat{\Sigma}_L \mathbf{z}_k)^* \mathbf{Q}_{ij} (\widehat{\Sigma}_L \mathbf{z}_k).$$

5. Let $T_{(1)}^1 \leq T_{(2)}^1 \leq \cdots \leq T_{(B)}^1$ be the order statistics of $T_k^1, k = 1, 2, \ldots, B$. Reject $\mathbf{H}_0^1$ at the level α if $T_1^* > T^1_{(\lfloor B(1-\alpha) \rfloor)}$, where $\lfloor x \rfloor$ stands for the largest integer smaller or equal to x. Let $B^* = \max\{k : T_{(k)}^1 \leq T_1^*\}$, the p-value can be denoted as $1 - B^*/B$.

4.5. *Choices of tuning parameters.* In this subsection we briefly discuss the practical choices of the key parameters, that is, the lag b of the auto-regression in Cholesky decomposition, the number of basis functions in sieve estimation and the bandwith selection in the nonparametric estimation of covariance matrix.

Similar to the discussion in Section 4.1, by Proposition 2.7, Lemma 2.8 and Theorem 3.7, for any given sufficiently large $b_0 \equiv b_0(n)$, the following statistic should be small enough:

$$\mathcal{T}_b = \sum_{j=b_1}^{b_0} \int_0^1 \widehat{\phi}_j^2(t)\, dt.$$

If the b_1th to b_0th off-diagonal elements of Φ are zero, by Theorem 4.3, $\mathcal{T}_b$ is normally distributed. Hence, we can follow the procedure described in the end of Section 4.4. For each fixed $b_1 < b_0$, we can formulate the null hypothesis as $\mathbf{H}_0^{b_1}$: kth off-diagonal element of Φ is zero for all $k \geq b_1$. Given level α, denote

$$b^* = \max_{b_1}\{b_1 < b_0 : \mathbf{H}_0^b \text{ is rejected}\}.$$

Then, we can choose $b = b^*$. Note that $b^* + 1$ is the first off diagonal where all its entries are effectively zeros in terms of statistical significance.

The number of basis functions can be chosen using model selection methods for nonparametric sieve estimation. However, due to nonstationarity, the classic Akaike information criterion (AIC) may fail under heteroskedasticity. In the present paper we use the cross-validation method described in [16], Section 8, where the cross-validation criterion is defined as

$$\mathrm{CV}(c) = \frac{1}{n}\sum_{i=2}^{n} \frac{\widehat{\epsilon}_{ic}^2}{(1-\upsilon_{ic})^2},$$

where $\{\widehat{\epsilon}_{ic}\}$ are the estimation residuals using sieve method with order of c and υ_{ic} is the leverage defined as $\upsilon_{ic} = \mathbf{y}_i^*(Y^*Y)\mathbf{y}_i$, where we recall (3.7). Hence, we can choose

$$\widehat{c} = \operatorname*{argmin}_{1\leq c\leq c_0} \mathrm{CV}(c),$$

where c_0 is a pre-chosen large value.

Finally, the bandwidth can be chosen using the standard leave-one-out cross-validation criterion for nonparametric estimation. Denote

$$\widehat{J}(h) := \sup_j \left\| \int_0^1 \widetilde{\Lambda}(t,j) \circ \widetilde{\Lambda}(t,j)\,dt - \frac{2}{n}\sum_{k=b+1}^{n} \widetilde{\Lambda}_{-k}(t_k,j) \right\|,$$

where $t_i = \frac{i}{n}$, $\circ$ is the Hadamard (entrywise) product for matrices and $\widetilde{\Lambda}_{-k}$ is the estimation excluding the sample $\widehat{\mathbf{h}}_k\widehat{\mathbf{h}}_{k+j}^*$. Therefore, the selected bandwidth is

$$\widehat{h} = \operatorname*{argmin}_h \widehat{J}(h).$$

Acknowledgments. The authors are grateful for the comments and suggestions of two anonymous referees, the Associate Editor and the Editor which have improved the paper significantly.

SUPPLEMENTARY MATERIAL

Supplement to "Estimation and inference for precision matrices of nonstationary time series" (DOI: 10.1214/19-AOS1894SUPP; .pdf). This supplementary material contains numerical simulations, auxiliary lemmas and technical proofs of the paper.

REFERENCES

[1] BELLONI, A., CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *J. Econometrics* **186** 345–366. MR3343791 https://doi.org/10.1016/j.jeconom.2015.02.014

[2] BICKEL, P. J. and LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969 https://doi.org/10.1214/009053607000000758

[3] BROCKWELL, P. J. and DAVIS, R. A. (1987). *Time Series*: *Theory and Methods*. *Springer Series in Statistics*. Springer, New York. MR0868859 https://doi.org/10.1007/978-1-4899-0004-3

[4] CAI, T. T., REN, Z. and ZHOU, H. H. (2013). Optimal rates of convergence for estimating Toeplitz covariance matrices. *Probab. Theory Related Fields* **156** 101–143. MR3055254 https://doi.org/10.1007/s00440-012-0422-7

[5] CAI, T. T. and ZHOU, H. H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Ann. Statist.* **40** 2389–2420. MR3097607 https://doi.org/10.1214/12-AOS998

[6] CHEN, X. (2007). Large sample sieve estimation of semi-nonparametric models. In *Handbook of Econometrics, Vol. 6B* (J. J. Heckman and E. E. Leamer, eds.). Chapter 76. https://doi.org/10.1016/S1573-4412(07)06076-X

[7] CHEN, X. and CHRISTENSEN, T. M. (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *J. Econometrics* **188** 447–465. MR3383220 https://doi.org/10.1016/j.jeconom.2015.03.010

[8] CHEN, X., XU, M. and WU, W. B. (2013). Covariance and precision matrix estimation for high-dimensional time series. *Ann. Statist.* **41** 2994–3021. MR3161455 https://doi.org/10.1214/13-AOS1182

[9] DEMKO, S., MOSS, W. F. and SMITH, P. W. (1984). Decay rates for inverses of band matrices. *Math. Comp.* **43** 491–499. MR0758197 https://doi.org/10.2307/2008290

[10] DING, X. (2017). Asymptotics of empirical eigen-structure for high dimensional sample covariance matrices of general form. Available at arXiv:1708.06296.

[11] DING, X. and YANG, F. (2018). A necessary and sufficient condition for edge universality at the largest singular values of covariance matrices. *Ann. Appl. Probab.* **28** 1679–1738. MR3809475 https://doi.org/10.1214/17-AAP1341

[12] DING, X. and ZHOU, Z. (2020). Supplement to "Estimation and inference for precision matrices of nonstationary time series." https://doi.org/10.1214/19-AOS1894SUPP.

[13] ERDŐS, L., YAU, H.-T. and YIN, J. (2012). Rigidity of eigenvalues of generalized Wigner matrices. *Adv. Math.* **229** 1435–1515. MR2871147 https://doi.org/10.1016/j.aim.2011.12.010

[14] FAN, J., LIAO, Y. and LIU, H. (2016). An overview of the estimation of large covariance and precision matrices. *Econom. J.* **19** C1–C32. MR3501529 https://doi.org/10.1111/ectj.12061

[15] HAMILTON, J. D. (1994). *Time Series Analysis*. Princeton Univ. Press, Princeton, NJ. MR1278033

[16] HANSEN, B. E. (2014). Nonparametric sieve regression: Least squares, averaging least squares, and cross-validation. In *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics* 215–248. Oxford Univ. Press, Oxford. MR3306927

[17] KNOWLES, A. and YIN, J. (2017). Anisotropic local laws for random matrices. *Probab. Theory Related Fields* **169** 257–352. MR3704770 https://doi.org/10.1007/s00440-016-0730-4

[18] LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. MR2572459 https://doi.org/10.1214/09-AOS720

[19] MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models. Monographs on Statistics and Applied Probability*. CRC Press, London. MR3223057 https://doi.org/10.1007/978-1-4899-3242-6

[20] MCMURRY, T. L. and POLITIS, D. N. (2015). High-dimensional autocovariance matrices and optimal linear prediction. *Electron. J. Stat.* **9** 753–788. MR3331856 https://doi.org/10.1214/15-EJS1000

[21] PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617–1642. MR2399865

[22] POURAHMADI, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **86** 677–690. MR1723786 https://doi.org/10.1093/biomet/86.3.677

[23] RAN, R.-S. and HUANG, T.-Z. (2009). An inversion algorithm for a banded matrix. *Comput. Math. Appl.* **58** 1699–1710. MR2557547 https://doi.org/10.1016/j.camwa.2009.07.069

[24] RÖLLIN, A. (2013). Stein's method in high dimensions with applications. *Ann. Inst. Henri Poincaré Probab. Stat.* **49** 529–549. MR3088380 https://doi.org/10.1214/11-aihp473

[25] STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053. MR0673642

[26] TAO, T. and VU, V. (2011). Random matrices: Universality of local eigenvalue statistics. *Acta Math.* **206** 127–204. MR2784665 https://doi.org/10.1007/s11511-011-0061-3

[27] TRETTER, C. (2008). *Spectral Theory of Block Operator Matrices and Applications*. Imperial College Press, London. MR2463978 https://doi.org/10.1142/9781848161122

[28] TROPP, J. (2015). *An Introduction to Matrix Concentration Inequalities. Foundations and Trends in Machine Learning*. Now Publishers Inc.

[29] WADE, W. (2014). *Introduction to Analysis*, 4th ed. Prentice Hall, Upper Saddle River, NJ.

[30] WU, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proc. Natl. Acad. Sci. USA* **102** 14150–14154. MR2172215 https://doi.org/10.1073/pnas.0506715102

[31] WU, W. B. and POURAHMADI, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* **90** 831–844. MR2024760 https://doi.org/10.1093/biomet/90.4.831

[32] XIAO, H. and WU, W. B. (2012). Covariance matrix estimation for stationary time series. *Ann. Statist.* **40** 466–493. MR3014314 https://doi.org/10.1214/11-AOS967

[33] XU, M., ZHANG, D. and WU, W. (2015). L^2 asymptotics for high-dimensional data. Available at arXiv:1405.7244.

[34] YUAN, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **11** 2261–2286. MR2719856

[35] YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824 https://doi.org/10.1093/biomet/asm018

[36] ZHANG, C. and ZHANG, T. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Statist. Sci.* **27** 576–593.

[37] ZHANG, X. and CHENG, G. (2018). Gaussian approximation for high dimensional vector under physical dependence. *Bernoulli* **24** 2640–2675. MR3779697 https://doi.org/10.3150/17-BEJ939

[38] ZHOU, Z. (2013). Inference for non-stationary time-series autoregression. *J. Time Series Anal.* **34** 508–516. MR3070872 https://doi.org/10.1111/jtsa.12028

[39] ZHOU, Z. and WU, W. B. (2009). Local linear quantile estimation for nonstationary time series. *Ann. Statist.* **37** 2696–2729. MR2541444 https://doi.org/10.1214/08-AOS636

[40] ZHOU, Z. and WU, W. B. (2010). Simultaneous inference of linear models with time varying coefficients. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 513–531. MR2758526 https://doi.org/10.1111/j.1467-9868.2010.00743.x