Sparse principal component analysis for high-dimensional stationary time series

Kou Fujimori | Yuichi Goto | Yan Liu | Masanobu Taniguchi

1Department of Economics, Faculty of Economics and Law, Shinshu University, Matsumoto, Japan
2Department of Mathematical Sciences, Faculty of Mathematics, Kyushu University, Fukuoka, Japan
3Faculty of Science and Engineering, Waseda University, Tokyo, Japan
4Institute for Mathematical Science, Faculty of Science and Engineering, Waseda University, Tokyo, Japan

Correspondence
Yan Liu, Faculty of Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan.
Email: liu@waseda.jp

Funding information
Japan Society for the Promotion of Science, Grant/Award Numbers: 18H05290, 20K11719, 21K13271, 21K20338, 23K11018, 23K16851

Abstract
We consider the sparse principal component analysis for high-dimensional stationary processes. The standard principal component analysis performs poorly when the dimension of the process is large. We establish oracle inequalities for penalized principal component estimators for the large class of processes including heavy-tailed time series. The rate of convergence of the estimators is established. We also elucidate the theoretical rate for choosing the tuning parameter in penalized estimators. The performance of the sparse principal component analysis is demonstrated by numerical simulations. The utility of the sparse principal component analysis for time series data is exemplified by the application to average temperature data.

KEYWORDS
high-dimensional statistics, multivariate stationary process, principal component analysis, sample covariance matrix, sparse estimation

1 | INTRODUCTION

The principal component analysis (PCA) has been a standard tool for multivariate data analysis. It facilitates the understanding of the covariance matrix and becomes a central method for dimension reduction and variable selection. When the sample size is large, Anderson (1963) developed the asymptotic theory for PCA. A thorough investigation into the standard PCA is summarized in Jolliffe (2002).

The dimension $p$ of the contemporary data is often large, compared with the sample size. Johnstone (2001) investigated the distribution of the largest eigenvalue when $p$ is large, and
introduced the concept of the spiked covariance matrix model. The sparse principal PCA, combined with variable selection techniques such as Lasso (Tibshirani, 1996) or elastic net (Zou & Hastie, 2005), was introduced in Zou et al. (2006). Shen and Huang (2008) considered the sparse PCA via regularized low-rank matrix approximation. Johnstone and Lu (2009) provided a simple algorithm for selecting a subset of coordinates with the largest sample variances with consistency even when the dimension $p$ is large. Amini and Wainwright (2009) proposed two computational methods for recovering the support set of the leading eigenvector in the spiked covariance model. Cai and Zhou (2012) derived the optimal rates of convergence for sparse covariance matrix estimation. Paul and Johnstone (2012) proposed an augmented sparse PCA method and showed that the procedure attains near optimal rate of convergence. Birnbaum et al. (2013) studied the problem of estimating the leading eigenvector under the $l_2$-loss for independent high-dimensional Gaussian observations. Cai et al. (2013) considered both min-max and adaptive estimation of the principal subspace in the high-dimensional setting. Vu and Lei (2013) also considered the sparse principal subspace estimation problems and established the optimal bounds for row subspace and nearly optimal for column subspace. Berthet and Rigollet (2013) derived a minimax optimality in a finite sample analysis for sparse principal components of a high-dimensional covariance matrix. The computational aspects of the sparse PCA have also been considered so far. To name a few, Ma (2013) proposed the iterative thresholding method for estimation of the leading eigenvector, and Wang et al. (2016) studied the computationally efficient algorithm to estimate the principal subspace. Recently, Kokoszka et al. (2019) dealt with asymptotic properties of PCA of highly variable functional data. The author also applied PCA to infinite variance functional data (Kokoszka & Kulik, 2023). van de Geer (2016) formalized the theoretical development for sparse PCA on some local set, which is induced to ensure the compatibility condition (see also, e.g. Bühlmann & van de Geer, 2011; van de Geer, 2014). Notably, much of the above theory was developed under the setting of i.i.d. observations.

The PCA for dependent data has also been studied for a long time. Zhao et al. (1986) proposed a new procedure for detection of signals based on eigenvalues of covariance matrix. Taniguchi and Krishnaiah (1987) derived the asymptotic distributions of eigenvalues of the sample covariance matrix from Gaussian stationary processes. However, all these developments are restricted to the case when the dimension $p$ is finite, that is, the PCA for multivariate stationary processes. More details of analyses for multivariate stationary processes can be found in Taniguchi and Kakizawa (2000). The limiting distribution of sample covariance matrix for large-dimensional linear models was derived in Jin et al. (2009). The Marčenko–Pastur theorem for time series was obtained in an explicit way in Yao (2012). The high-dimensional covariance estimation for dependent data with some regularization techniques was introduced in Pourahmadi (2013). The theoretical development for regularized estimation in sparse high-dimensional time series models was considered in Basu and Michailidis (2015). They showed that a restricted eigenvalue condition holds with high probability. Wong et al. (2020) derived the oracle inequality of the Lasso-type estimator only for the parameters in regression models of time series. In our paper, we consider the sparse PCA for high-dimensional stationary time series and establish the oracle inequalities for the Lasso-type PCA estimator for the first time under mixing conditions. Our results serve both of $\alpha$-mixing Gaussian processes and $\beta$-mixing sub-Weibull processes. In addition, we also derived the oracle inequality for $l_0$-penalized estimators for comparison. The finite sample performance is illustrated by some numerical simulations. Empirical studies describe the utility of our method.
The rest of the paper is organized as follows. In Section 2, we provide fundamental settings for
the sparse PCA for stationary processes. Theoretical results of the Lasso-type PCA for Gaussian
processes and heavy-tailed processes are discussed in Sections 3 and 4, respectively. In addition,
the $l_0$-penalized estimation is discussed in Section 5. Section 6 gives several simulation results to
demonstrate the finite sample performance of sparse PCA. The rigorous proofs and preliminary
lemmas are relegated to Section 7.

2 | PRELIMINARIES

Let us introduce some mathematical notations. For a vector $v \in \mathbb{R}^p$, the $l_r$-norm $\|v\|_r$ is defined as
$\|v\|_r = \left( \sum_{i=1}^{p} |v_i|^r \right)^{1/r}$ for $r \in (0, \infty)$. Also, let $\|v\|_0$ and $\|v\|_\infty$ be $\|v\|_0 = \sum_{i=1}^{p} \mathbb{1}_{\{v_i > 0\}}$ and $\|v\|_\infty = \max_{1 \leq i \leq p} |v_i|$, respectively. For a $p \times p$ matrix $A$, the operator norm $\|A\|_r$ is defined as
$\|A\|_r := \sup_{\|v\|_r = 1} \|Av\|_r$, $r \in (0, \infty]$.
Moreover, the “max” norm of the matrix $A$ is defined as $\|A\|_{\max} := \max_{i,j} |A_{ij}|$. For some
sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \asymp b_n$ and $a_n \lesssim b_n$ when $a_n/b_n$ converges to a positive
constant as $n \to \infty$ and $a_n \leq Cb_n$ with a positive constant $C$ independent of $n$.

2.1 | Model setups

Let $\{X_t\}_{t \in \mathbb{Z}}$ be an $\mathbb{R}^p$-valued, strictly stationary, and centered time series on a probability space
$(\Omega, \mathcal{F}, P)$. Suppose the observation stretch $(X_1, \ldots, X_n)$, $n \in \mathbb{N}$, is available. Consider the following $p \times p$ matrices
$\Sigma_0 = E[X_t X_t^\top]$, \quad $\hat{\Sigma}_n = \frac{1}{n} \sum_{t=1}^{n} X_t X_t^\top$.

Let $q^0$ be the first principal component corresponding to the largest eigenvalue $\phi_{\max}^2 := \Lambda_{\max}(\Sigma_0)$
of $\Sigma_0$, so that $q^0$ is normalized as $\|q^0\|_2 = 1$. The parameter of interest is
$\beta^0 = \phi_{\max} q^0$,
which is a solution to the following optimization problem
$\beta^0 = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{4} \|\Sigma_0 - \beta \beta^\top\|^2_F$,
where $\| \cdot \|_F$ is the Frobenius norm. In other words, it holds that
$\Sigma_0 \beta^0 = \phi_{\max}^2 \beta^0 = \|\beta^0\|^2_2 \beta^0$.
Our primary interest is the sparse principal component estimation. Let $S$ be $S = \{j : \beta_j^0 \neq 0\}$.
Specifically, $\beta^0$ is supposed to be $s_0$-sparse, that is, $|S| = s_0$. A typical motivating example is given as follows.
**Example 1.** Consider the stationary process \( \{X_t\} \) taking the form of VAR model, that is,

\[
X_t = AX_{t-1} + \epsilon_t, \quad \epsilon_t \sim \text{i.i.d.}(0, I_p),
\]

(1)

where \( A \) is a \( p \times p \) deterministic matrix with the decomposition

\[
A = \sum_{j=1}^{p} \nu_j p_j p_j^\top,
\]

such that \( 1 > \nu_1 \geq \cdots \geq \nu_p \geq 0 \) are the eigenvalues of \( A \), with \( p_j \) the associated eigenvectors. Suppose the eigenvector \( p_1 \) is \( s_0 \)-sparse. By the holomorphic functional calculus (see, e.g., chapter 2 of Lütkepohl, 2005), we have

\[
\Sigma_0 = \sum_{j=1}^{p} \frac{1}{1 - \nu_j^2} p_j p_j^\top,
\]

which shows that the first principal component of \( \Sigma_0 \) is also \( s_0 \)-sparse.

In this paper, we consider the following penalized PCA estimators. In such a high-dimensional setting, the sparse estimation is essential for variable selection, which facilitates the interpretation of features in datasets.

**Definition 1.** The following estimator for \( \beta_0 \) is defined as

\[
\hat{\beta}_n := \arg \min_{\beta \in B} \left\{ \frac{1}{4} \| \Sigma_n - \beta \beta^\top \|_F^2 + \lambda \text{pen}(\beta) \right\}, \quad B := \{ \beta : \| \beta - \beta_0 \|_2 \leq \eta \},
\]

(2)

where \( \lambda \geq 0 \) is a tuning parameter, \( \text{pen}(\cdot) \) is some penalty function, and \( \eta > 0 \) is a suitable constant. The following estimators are focused on in this paper. The \( l_1 \)-penalized estimator \( \hat{\beta}_1^1 \) is defined as

\[
\hat{\beta}_1^1 := \arg \min_{\beta \in B} \left\{ \frac{1}{4} \| \Sigma_n - \beta \beta^\top \|_F^2 + \lambda_1 \| \beta \|_1 \right\},
\]

(3)

where \( \lambda = \lambda_1 \) and \( \text{pen}(\beta) = \| \beta \|_1 \). The estimator \( \hat{\beta}_1^1 \) is referred to as the Lasso-type estimator. The \( l_0 \)-penalized estimator \( \hat{\beta}_0 \) is defined as

\[
\hat{\beta}_0 := \arg \min_{\beta \in B} \left\{ \frac{1}{4} \| \Sigma_n - \beta \beta^\top \|_F^2 + \lambda_0 \| \beta \|_0 \right\},
\]

(4)

where \( \lambda = \lambda_0 \) and \( \text{pen}(\beta) = \| \beta \|_0 \).

**Remark 1.** Let \( \phi_j^2 \) and \( u_j \) be the \( j \)th eigenvalue of \( \Sigma_0 \) such that \( \phi_1^2 \geq \cdots \geq \phi_p^2 \) and the corresponding eigenvector, respectively. To estimate \( \beta_0^j := \phi_j u_j \) for \( j = 1, \ldots, r \) with \( r < p \), we can consider the following direct extension of the Definition 1;

\[
(\hat{\beta}_1^1, \ldots, \hat{\beta}_r) = \arg \min_{\beta_1, \ldots, \beta_r} \frac{1}{4} \left\| \Sigma_n - \sum_{j=1}^{r} \beta_j \beta_j^\top \right\|_F^2 + \sum_{j=1}^{r} \lambda_j \text{pen}(\beta_j),
\]

(4)
where $\lambda_j, j = 1, \ldots, r$ are tuning parameters. Estimation problems similar to (4) are considered by some researchers for i.i.d. cases. For instance, Shen and Huang (2008) proposed the sequential algorithm to estimate the eigenvectors, that is, to estimate the $j + 1$st principle component, they considered the rank one approximation of the residual matrix which is constructed by using the estimator of the $j$th principle component. On the other hand, Ma (2013) dealt with the sparse principal subspace estimation, which enables us to estimate principal components simultaneously. Moreover, Mackey (2008) proposed the iterative deflation method for sparse estimation of multiple principal components. These approaches may be applicable to our settings, however, it is difficult to derive the theoretical error bounds or rates of convergence of such estimators since the concentration inequalities used to derive the error bound for time series settings are more complicated. Therefore, we focus on the first principle component in this paper, but the theoretical verification of the estimation of other principal components for high-dimensional stationary time series is not a trivial problem, which should be studied precisely.

We establish the error bound of the estimators $\hat{\beta}_1^n$ and $\hat{\beta}_0^n$ for high-dimensional time series. In the following, we list the definition of mixing coefficients for stochastic processes.

**Definition 2.** For a stationary process $\{X_t\}_{t \in \mathbb{Z}}$, we define the following quantities:

(i) The $\alpha$-mixing coefficients for $\{X_t\}_{t \in \mathbb{Z}}$ is defined as

$$\alpha(l) := \sup \{|P(A \cap B) - P(A)P(B)| : A \in \sigma(X_s, s \leq t), B \in \sigma(X_s, s \geq t + l) \text{ for all } t \in \mathbb{Z}, \ l \in \mathbb{Z} \}.$$ 

The process $\{X_t\}_{t \in \mathbb{Z}}$ is called $\alpha$-mixing if $\alpha(l) \to 0$ as $l \to \infty$.

(ii) The $\rho$-mixing coefficients for $\{X_t\}_{t \in \mathbb{Z}}$ is defined as

$$\rho(l) := \sup \{|\text{Cov}(f(X_t), g(X_{t+l}))| : E[f] = E[g] = 0, \ E[f^2] = E[g^2] = 1\}, \ l \in \mathbb{Z}.$$ 

The process $\{X_t\}_{t \in \mathbb{Z}}$ is called $\rho$-mixing if $\rho(l) \to 0$ as $l \to \infty$.

(iii) The $\beta$-mixing coefficients for $\{X_t\}_{t \in \mathbb{Z}}$ is defined by

$$\beta(l) := \sup \sum_{i=1}^{I} \sum_{j=1}^{J} |P(A \cap B) - P(A)P(B)|, \ l \in \mathbb{Z},$$

where the supremum is over all pair of partitions $\{A_i\}_{1 \leq i \leq I} \subset \sigma(X_s, s \leq t)$ and $\{B_j\}_{1 \leq j \leq J} \subset \sigma(X_s, s \geq t + l)$. The process $\{X_t\}_{t \in \mathbb{Z}}$ is called $\beta$-mixing if $\beta(l) \to 0$ as $l \to \infty$.

All these conditions are known as the weak dependence conditions (e.g., Tikhomirov, 1981), under which the convergence rate in the central limit theorem for weakly dependent random variables is evaluated.

Now consider the spectral decomposition of $\Sigma_0$ as

$$\Sigma_0 = U\Phi_0U^T.$$
where $\Phi_0^2 := \text{diag}(\phi_1^2, \ldots, \phi_p^2)$ with $\phi_1^2 \geq \cdots \geq \phi_p^2 \geq 0$ is a diagonal matrix constructed by the eigenvalues of $\Sigma_0$, and $U = (u_1, \ldots, u_p)$ satisfies that $UU^T = U^T U = I_p$. Actually, we have $\phi_{\text{max}} = \phi_1$ and $q^0 = u_1$. Hereafter, we assume the following conditions.

**Assumption 1.** There exists a constant $\sigma > 3\eta > 0$ such that 

$$\phi_{\text{max}} \geq \phi_j + \sigma, \quad j \geq 2,$$

where $\eta > 0$ is a constant appeared in the local set $B$ in Definition 1.

In the sequel, the constant $\sigma$ is assumed to be independent of the dimension $p$. However, we can consider the case where $\sigma \to \infty$ as $p \to \infty$, which means that the covariance matrix satisfies the strongly spiked structure. See remark 3.5 (i) for such a case.

## 2.2 Risk functions

We define the theoretical risk $R(\cdot)$ and empirical risk $R_n(\cdot)$ and their derivatives with respect to $\beta$ as follows:

$$R(\beta) := -\frac{1}{2} \beta^T \Sigma_0 \beta + \frac{1}{4} \Vert \beta \Vert^4_2, \quad R_n(\beta) := -\frac{1}{2} \beta^T \hat{\Sigma}_n \beta + \frac{1}{4} \Vert \beta \Vert^4_2,$$

$$\hat{R}(\beta) := -\Sigma_0 \beta + \Vert \beta \Vert^2_2 \beta, \quad \hat{R}_n(\beta) := -\hat{\Sigma}_n \beta + \Vert \beta \Vert^2_2 \beta,$$

and

$$\ddot{R}(\beta) := -\Sigma_0 + \Vert \beta \Vert^2_2 I_p + 2 \beta \beta^T, \quad \ddot{R}_n(\beta) := -\hat{\Sigma}_n + \Vert \beta \Vert^2_2 I_p + 2 \beta \beta^T.$$

Then, it holds that

$$\beta^0 = \arg\min_{\beta \in \mathbb{R}^p} R(\beta), \quad \hat{\beta}_n = \arg\min_{\beta \in B} \{ R_n(\beta) + \lambda \text{ pen}(\beta) \}.$$

For brevity, denote $\hat{\Sigma}_n - \Sigma_0$ by $W_n$. Note that

$$\ddot{R}_n(\beta) - \ddot{R}(\beta) = -W_n \beta,$$

and that

$$\ddot{R}_n(\beta) - \ddot{R}(\beta) = -W_n.$$

The following result holds for the theoretical risk function.

**Proposition 1 ((van de Geer, 2016, lemma 12.7)).** Under Assumption 1, $\ddot{R}(\beta)$ satisfies that

$$\Lambda_{\text{min}}(\ddot{R}(\beta)) \geq 2(\sigma - 3\eta), \quad \beta \in B,$$

where $\Lambda_{\text{min}}(\ddot{R}(\beta))$ is the smallest eigenvalue of $\ddot{R}(\beta)$, $\sigma$ and $\eta$ are given in Assumption 1.
By Proposition 1, the theoretical risk function $R(\cdot)$ is shown to be strictly convex on $B$. This motivates us to consider the penalized PCA estimators as an optimization problem. We always assume that the bound $\eta$ of $B$ satisfies $\eta < \sigma/3$ later on in this paper.

3 | LASSO-TYPE ESTIMATOR FOR $\alpha$-MIXING GAUSSIAN PROCESS

In this section, we deal with the Lasso-type estimator $\hat{\beta}_n^1$ for $\alpha$-mixing Gaussian stationary processes. The discussion is based on preliminary lemmas in Section 7. The proofs of theorems in this section are relegated to Section 7.

**Assumption 2.** The process $\{X_t\}_{t \in \mathbb{Z}}$ is a centered $\alpha$-mixing Gaussian stationary process.

An $\alpha$-mixing Gaussian process is also $\rho$-mixing. For every $0 < \delta < \sigma - 3\eta$, Lemma 2 suggests that

$$P \left( \sup_{v \in \mathbb{S}^{p-1}} |v^T W_n v| > 2\delta \right) \leq 2\exp(p \log 6 - cn \min\{\xi_n, \xi_n^2\}),$$

where

$$\xi_n = \frac{\delta}{\sum_{l=0}^{n} \rho(l) \phi_{\max}},$$

and $c > 0$ is a global constant appeared in Hanson–Wright’s inequality (Lemma 1). This fact implies that the empirical risk function $R_n(\beta)$ is convex with probability $1 - 2\exp(p \log 6 - cn \min\{\xi_n, \xi_n^2\})$. Let $\zeta_n$ be a sequence such that

$$\zeta_n := \sqrt{\frac{2(b + 1) \log p}{cn}},$$

where $b > 0$ is a free parameter (we can take it arbitrary). Then, it follows from Lemma 3 that

$$P \left( \|W_n\|_{\max} \leq \zeta_n \sum_{l=0}^{n} \rho(l) \phi_{\max} \right) \geq 1 - 2\exp(-2b \log p).$$

Combining these facts, we have the following inequality.

**Theorem 1.** Suppose Assumptions 1 and 2 hold. Moreover, we assume that $p/n \to q \in [0, \infty)$ as $n \to \infty$. Then, for every $n$ satisfying that

$$\frac{\log p}{n} \leq \frac{c}{2(b + 1)},$$

and $\lambda_1$ satisfying that

$$\lambda_1 > \left( \frac{C + 1}{C - 1} \|\beta_0\|_1 \right) \gamma_n,$$
where

\[ \gamma_n := \phi_{\text{max}} \left( \xi_n \sum_{l=0}^{n} \rho(l) \right), \]  

(6)

with some universal constant \( C > 1 \), it holds that

\[
\begin{align*}
P \left( \| \hat{\beta}_n^1 - \beta^0 \|_1 \leq \frac{2(C + 1)^2 s_0 \lambda_1}{\sigma - 3\eta - \delta} \right) \\
\geq 1 - 2 \exp(-2b \log p) - 2 \exp(p \log 6 - cn \min(\xi_n, \xi_0^2)) ,
\end{align*}
\]

(7)

where

\[ \xi_n = \frac{\delta}{\sum_{l=0}^{n} \rho(l) \phi_{\text{max}}} . \]

Note that \( b > 0 \) is a free parameter and we can think as \( b = 1 \) when \( n \) is sufficiently large (or \( \log p/n \) is sufficiently small). The larger \( b \) implies the larger probability in the right-hand side of the oracle inequality. On the other hand, when \( b \) is large, the sample size needs to be larger to ensure that the error bound of the estimator is small. The constant \( c > 0 \) appeared in Hanson–Wright’s inequality is a global constant independent of \( n, p, \Sigma_0 \) and other parameters. See, for example, Moshksar (2021) for the detail of the constant in Hanson–Wright’s inequality. Moreover, it follows from (5) and the left-hand side of (7) that the main order of the error bound is the order of \( s_0 \lambda_1 \), where \( \lambda_1 \) is the tuning parameter in \( l_1 \)-penalized estimator. Note that Theorem 1 holds for \( \lambda_1 \) larger than \( \gamma_n \| \beta^0 \|_1 \) up to some constant multiplication, where \( \gamma_n \) is the bound for \( \| W_n \|_{\text{max}} = \| \hat{\Sigma}_n - \Sigma_0 \|_{\text{max}} \). Since \( \gamma_n \) depends on the mixing coefficients, the tuning parameter should be chosen by considering the dependence of the process. The probability for the oracle inequality mainly depends on \( p \) and \( \xi_n \), since \( b \) can be taken arbitrarily. The sequence \( \xi_n \) also depends on the mixing coefficients. Thus, the oracle inequality of the estimator is obtained in terms of dependence residing in time series.

**Remark 2.** For a simple interpretation, we can summarize Theorem 1 as follows. Let \( \delta_n > 0 \) be a decreasing sequence, \( A_n \) be a sequence and \( B \) be a constant defined by

\[
A_n := \left( \frac{C + 1}{C - 1} \phi_{\text{max}} \| \beta^0 \|_1 \right) \left( \xi_n \sum_{l=0}^{n} \rho(l) \right) \quad \text{and} \quad B := \frac{2(C + 1)^2}{\sigma - 3\eta - \delta},
\]

respectively. If \( \lambda_1 > A_n \), then it holds that \( \| \hat{\beta}_n^1 - \beta^0 \|_1 \leq B s_0 \lambda_1 \) with probability at least \( 1 - \delta_n \).

Note that \( \| \beta^0 \|_1 \) appeared in the lower bound of \( \lambda_1 \) in Theorem 1 is evaluated as follows:

\[ \| \beta^0 \|_1 \leq \sqrt{s_0} \| \beta^0 \|_2 = \sqrt{s_0} \phi_{\text{max}} . \]

To obtain the asymptotic behavior of the estimator, we induce some additional assumptions.
Assumption 3.

(i) The sample size $n$ and the dimension $p$ satisfy that $p = q_n n$, where $q_n$ is a sequence such that

$$0 < q_n < \frac{c}{\log 6} \left( \frac{\sigma - 3\eta}{C \rho \phi_{\text{max}}} \right)^2,$$

for sufficiently large $n$, where $C_p := \sum_{l=0}^{\infty} \rho(l)$ and $c > 0$ is a constant appeared in Hanson–Wright’s inequality (Lemma 1).

(ii) The largest eigenvalue $\phi_{\text{max}}^2$ of population covariance matrix $\Sigma_0$ and the $\rho$-mixing coefficients satisfy that

$$\phi_{\text{max}}^2 = O(q_n^{-1/2}), \quad \text{as } n \to \infty \quad \text{and} \quad \sum_{l=0}^{\infty} \rho(l) < \infty.$$

(iii) There exists an $N \in \mathbb{N}$ such that for every $n \geq N$,

$$\phi_{\text{max}} \sum_{l=0}^{n} \rho(l) \geq \sigma - 3\eta.$$

Condition (ii) for the largest eigenvalue $\phi_{\text{max}}^2$ is seemingly strict. For example, if $q_n = O(1)$, Condition (ii) is equivalent to $\phi_{\text{max}}^2 = O(1)$. On the other hand, if $q_n = o(1)$, $\phi_{\text{max}}^2$ is allowed to diverge. The motivating example provided by (1) satisfies this condition, and thus, this condition is general enough for our purpose. Conditions (i) and (iii) imply that the empirical risk function is convex with high probability, which is a consequence of Lemma 2. Note that van de Geer (2016) assume that $p \log p$ is sufficiently smaller than $n$, which is more restrictive than our assumption. We can relax Assumption 3 since as far as

$$p \log 6 - cn \min\{\xi_n, \xi_n^2\} \to -\infty, \quad n \to \infty,$$

where

$$\xi_n = \frac{\delta}{\sum_{l=0}^{n} \rho(l) \phi_{\text{max}}},$$

for some $\delta < \sigma - 3\eta$, under which the asymptotic behavior can be derived since the convexity is verified with high probability. The rate of convergence of the estimator is obtained as follows.

**Corollary 1.** Suppose that Assumptions 1, 2, and 3 are satisfied, and that

$$\lambda_1 \leq \phi_{\text{max}} \sqrt{s_0} \log p \frac{1}{n}.$$

Then, the following hold true.

$$\|\hat{\rho}_n - \rho_0 \|_1 = O_p \left( s_0^{3/2} \phi_{\text{max}} \sqrt{\frac{\log p}{n}} \right).$$
and
\[ \| \hat{\beta}_n^1 - \beta^0 \|_2^2 = O_p \left( \frac{s_0^3 \phi_{\max}^2 \log p}{n} \right), \]
as \( n \to \infty \).

We can consider the estimators for the largest eigenvalue \( \phi_{\max}^2 \) and the corresponding normalized eigenvector \( q^0 \) as follows;
\[ \hat{\phi}_{\max}^2 := \| \hat{\beta}_n^1 \|_2^2, \quad \hat{q}_n := \frac{\hat{\beta}_n^1}{\| \hat{\beta}_n^1 \|_2}. \]

Noting that
\[ |\hat{\phi}_{\max} - \phi_{\max}| = |\|\hat{\beta}_n^1\|_2 - \|\beta^0\|_2| \leq |\|\hat{\beta}_n^1 - \beta^0\|_2|, \]
and
\[ \| \hat{q}_n - q^0 \|_2 = \left\| \frac{\hat{\beta}_n^1}{\phi_{\max}} - \frac{\beta^0}{\phi_{\max}} \right\|_2 \]
\[ \leq \left| \frac{\| \hat{\beta}_n^1 - \beta^0 \|_2^2}{\phi_{\max}} + \| \hat{\beta}_n^1 \|_2 \left| \frac{1}{\phi_{\max}} - \frac{1}{\phi_{\max}} \right| \right| \]
\[ = \left| \frac{\| \hat{\beta}_n^1 - \beta^0 \|_2}{\phi_{\max}} + \phi_{\max} \frac{|\hat{\phi}_{\max} - \phi_{\max}|}{\phi_{\max} \phi_{\max}} \right| \]
\[ \leq \frac{2}{\phi_{\max}^2} \| \hat{\beta}_n^1 - \beta^0 \|_2, \]
we have the following rates of convergence of these estimators.

**Corollary 2.** Suppose that Assumptions 1, 2, and 3 are satisfied, and that
\[ \lambda_1 \asymp \phi_{\max} \sqrt{\frac{\log p}{n}}. \]

Then, the following hold true.
\[ |\hat{\phi}_{\max} - \phi_{\max}| = O_p \left( s_0^{3/2} \phi_{\max} \sqrt{\frac{\log p}{n}} \right), \]
and
\[ \| \hat{q}_n - q^0 \|_2 = O_p \left( s_0^{3/2} \sqrt{\frac{\log p}{n}} \right). \]

**Remark 3.**

(i) A reviewer of this paper suggests that if we consider the case where \( \sigma \to \infty \) as \( p \to \infty \), (e.g., the case where the first two eigenvalues \( \phi_1^2 \) and \( \phi_2^2 \) satisfies that \( \phi_1^2 \to \infty \) and
we may relax the condition for the largest eigenvalue $\phi_n^2$ and $q_n = p/n$ for estimation of $q^0$. Actually, inequality (7.1) in subsection 7.1 implies that, as far as

$$\sqrt{\frac{\log 6}{c}}C_p < \frac{\sigma - 3\eta}{q_n^{1/2} \phi_{\max}},$$

for sufficiently large $n$, $q_n$ and $\phi_{\max}$ may be allowed to diverge and the consistency of $\hat{q}_n$ may be verified by Corollary 2. However, since this is beyond the scope of this paper, we assume that the spikiness $\sigma > 0$ is a constant independent of the dimension $p$.

(ii) When $\rho(l) = 0$, $l \geq 1$, $\{X_t\}$ is an i.i.d. normal sequence. Our result suggests that even for i.i.d. case,

$$||\hat{\beta}_n^1 - \beta^0||_1 = O_p \left( s_0^{3/2} \phi_{\max} \sqrt{\frac{\log p}{n}} \right).$$

On the other hand, the rate of convergence derived in chapter 12 of van de Geer (2016) is

$$||\hat{\beta}_n^1 - \beta^0||_1 = O_p \left( s_0 \sqrt{\frac{\log p}{n}} \right),$$

when $\phi_{\max}^2 = O(1)$. The difference is caused by the concentration inequality used to derive the error bound of the $W_n = \hat{\Sigma}_n - \Sigma_0$. More precisely, we apply the Hanson–Wright’s inequality to the random vectors, while van de Geer (2016) apply the matrix Bernstein inequality, see, for example, Tropp (2012) for the detail of such inequalities. For i.i.d. (or matrix-martingale) cases, we can evaluate the eigenvalues of the matrix-valued random variables by Tropp’s approach, however, for our setting, it may be difficult to evaluate the errors by the similar way.

(iii) We can consider the non-centered case, that is, $E[X_t] = \mu \neq 0$, when the mean vector $\mu$ is known, since we can construct the centered time series by $X_t - \mu$. When $\mu$ is unknown, the centering step by $X_t - \bar{X}_n$, where $\bar{X}_n$ is a sample mean, causes a bias in high-dimensional settings. Therefore, it may be more difficult to establish the error bound of sparse PCA estimators than that for centered cases. Note that, however, in our empirical study in Section 6, we apply the centering without theoretical verification.

4 LASSO-TYPE ESTIMATOR FOR $\beta$-MIXING SUB-WEIBULL PROCESS

Next, we consider the Lasso-type estimator $\hat{\beta}_n^1$ for stationary processes with heavier tails.

Definition 3 (Sub-Weibull random variables). Let $\gamma > 0$. The sub-Weibull($\gamma$) is defined as follows.
(i) An \( \mathbb{R} \)-valued random variable \( X \) is called the sub-Weibull(\( \gamma \)) if it satisfies that there exists a constant \( K > 0 \) such that
\[
(E[|X|^q])^{\frac{1}{q}} \leq Kq^{\gamma}, \quad \forall q \geq 1 \land \gamma.
\]
The sub-Weibull(\( \gamma \))-norm \( || \cdot ||_{\psi\gamma} \) is defined for sub-Weibull(\( \gamma \)) random variable \( X \) as follows:
\[
||X||_{\psi\gamma} := \sup_{q \geq 1} (E[|X|^q])^{\frac{1}{q}} q^{-\frac{1}{\gamma}}.
\]
(ii) The \( \mathbb{R}^p \)-valued random variable \( X \) is called the sub-Weibull(\( \gamma \)) if it satisfies that each component of \( X \) is sub-Weibull(\( \gamma \)). Then, the sub-Weibull(\( \gamma \))-norm \( || \cdot ||_{\psi\gamma} \) is defined for sub-Weibull(\( \gamma \)) random variable \( X \) as follows:
\[
||X||_{\psi\gamma} := \sup_{v \in \mathbb{S}^{p-1}} ||v^\top X||_{\psi\gamma},
\]
where \( \mathbb{S}^{p-1} \) is the unit sphere on \( \mathbb{R}^p \).

The sub-Gaussian random variables is sub-Weibull(2); the sub-exponential random variable is sub-Weibull(1). Note that, for \( \gamma < 1 \), sub-Weibull(\( \gamma \)) random variables have heavier tail than sub-exponential and sub-Gaussian random variables.

We consider the stationary process \( \{X_t\}_{t \in \mathbb{Z}} \) satisfying the following conditions in this section.

**Assumption 4.**

(i) The process \( \{X_t\}_{t \in \mathbb{Z}} \) is geometrically \( \beta \)-mixing, that is, there exist constants \( c, \gamma_1 > 0 \) such that
\[
\beta(n) \leq 2 \exp(-cn^{\gamma_1}), \quad \forall n \in \mathbb{N}.
\]
(ii) For a constant \( \gamma_2 > 0 \), the process \( \{X_t\}_{t \in \mathbb{Z}} \) is sub-Weibull(\( \gamma_2 \)), that is, there exists a constant \( K_1 > 0 \) such that
\[
||X_t||_{\psi_{\gamma_2}} \leq K_1, \quad \forall t \in \mathbb{Z}.
\]
(iii) It holds that
\[
\left( \frac{1}{\gamma_1} + \frac{2}{\gamma_2} \right)^{-1} < 1,
\]
where \( \gamma_1 \) and \( \gamma_2 \) are defined in (i) and (ii), respectively.
(iv) It holds that \( p = q_n n^\gamma \), where \( q_n \) is a sequence such that
\[
0 < q_n < \frac{(\sigma - 3\eta)^\gamma}{K_2^2 C_1 \log 6},
\]
for sufficiently large \( n \) with a constant \( C_1 \) depending only on \( c, \gamma_1, \) and \( \gamma_2 \).

\[
\gamma := \left( \frac{1}{\gamma_1} + \frac{2}{\gamma_2} \right)^{-1},
\]
and
\[ K_2 := 2^{2/\gamma} + K_1^2. \]

In the sequel, we assume that \( K_1 \) and \( K_2 \) are constants independent of the dimension \( p \) for brevity, however, we can consider the case where they may depend on \( p \). Noting that the largest eigenvalue \( \phi_{\max}^2 \) satisfies that \( \phi_{\max} \lesssim K_1 \), we can see that \( \phi_{\max}^2 = O(1) \) if \( K_1 \) is a constant independent of \( p \). On the other hand, when \( q_n = p/n^\gamma \to 0 \) as \( n \to \infty \), \( K_1 \) and \( K_2 \) can diverge as long as \( K_2 q_n^{1/\gamma} = O(1) \). Introducing \( K_2 \) can reduce terms in the oracle inequality in the following. As a remark, for any \( X_i \), we have
\[
\|X^2_i\|_{\psi_{2/2}} \leq K_2, \quad \|(X_{i,t} + X_{j,t})^2\|_{\psi_{2/2}} \leq K_2, \quad i, j = 1, \ldots, p.
\]

Let \( \tilde{b} \) and \( \delta \) be constants and \( \tau_n \) be a sequence such that
\[
q_n^{1/\gamma} K_2 C_1^{1/\gamma} (\log 6)^{1/\gamma} < \delta < \sigma - 3\eta,
\]
and
\[
\tau_n := \max \left\{ \frac{K_2 C_1^{1/\gamma} (\log np^2 + 2\tilde{b} \log p)^{1/\gamma}}{n}, \frac{\sqrt{2C_2 (\tilde{b} + 1) \log p}}{\sqrt{n}} \right\},
\]
respectively. The constant \( \delta \) ensures the convexity of the empirical risk with high probability, and the sequence \( \{ \tau_n \} \) controls the error bound for \( \|W_n\|_{\psi_{2/2}} \), respectively. The oracle inequality is obtained as follows.

**Theorem 2.** Suppose Assumptions 1 and 4 hold. Let \( \zeta_n > \tau_n \). For \( n > 4 \), a constant \( \delta \) satisfying (8) and \( \lambda_1 \) satisfying that
\[
\lambda_1 > \frac{C + 1}{C - 1} \zeta_n \|\beta^0\|_1,
\]
where \( C \geq 1 \) is a universal constant, it holds that
\[
P \left( \|\hat{\beta}_n - \beta^0\|_1 \leq \frac{2(C + 1)^2 S_0 \lambda_1}{\sigma - 3\eta - \delta} \right)
\geq 1 - 2 \exp(-\tilde{b} \log p^2) - 2n \exp \left( p \log 6 - \frac{(\delta n)^2}{K_2 C_1} \right) - 2 \exp \left( p \log 6 - \frac{\delta^2 n}{K_2^2 C_2} \right),
\]
where \( \tilde{b} > 0 \) is a free parameter.

As well as Gaussian case, the parameter \( \tilde{b} > 0 \) is a free parameter. The condition (8) ensures that \( \sigma - 3\eta - \delta > 0 \). Since \( C > 1 \) is a universal constant, the main part of the error bound is still the order of \( S_0 \lambda_1 \), where \( \lambda_1 \) is the tuning parameter in \( l_1 \)-penalized estimator. Theorem 2 holds for \( \lambda_1 \) larger than \( \zeta_n \|\beta^0\|_1 \) up to some constant multiplication, where \( \zeta_n \) is the bound for \( \|W_n\|_{\psi_{2/2}} = \|\hat{\Sigma}_n - \Sigma_0\|_{\psi_{2/2}} \). Interestingly, the error bound and the tail probability do not depend on the mixing coefficients, since we assume the geometric \( \beta \)-mixing condition here for the process. However, it is clear that the decay of the probability is slower than that for Gaussian case.
Remark 4. We can summarize Theorem 2 as follows. Let $\delta_n > 0$ be a decreasing sequence, $A_n$ be a sequence and $B$ be a constant defined by

$$A_n := \frac{C + 1}{C - 1} c_n ||\beta^0||_1 \quad \text{and} \quad B := \frac{2(C + 1)^2}{\sigma - 3\eta - \delta},$$

respectively. If $\lambda_1 > A_n$, then it holds that $||\hat{\beta}^1_n - \beta^0||_1 \leq Bs_0 \lambda_1$ with probability at least $1 - \delta_n$.

The rate of convergence of the estimator $\hat{\beta}^1_n$ is established under some additional conditions.

Corollary 3. Suppose that the same assumptions as Theorem 2 hold. Assume moreover that

$$\zeta_n \asymp \sqrt{\frac{\log p}{n}},$$

and

$$\lambda_1 \asymp \sqrt{s_0 \frac{\log p}{n}}.$$

Then, the following hold true.

$$||\hat{\beta}^1_n - \beta^0||_1 = O_p \left( s_0^{3/2} \sqrt{\frac{\log p}{n}} \right),$$

and

$$||\hat{\beta}^1_n - \beta^0||_2^2 = O_p \left( s_0^3 \frac{\log p}{n} \right),$$

as $n \to \infty$.

Remark 5. As far as $\phi_{\text{max}}$ obeys the constant order, the rates of convergence of the estimator in both cases are $||\hat{\beta}^1_n - \beta^0||_1 = O_p(s_0^{3/2} \sqrt{\log p/n})$. However, for sub-Weibull case, we can see that the decay of the tail probability derived in Theorem 2 is slower than the corresponding result for Gaussian processes, which is caused by the heavy-tail property of sub-Weibull distribution when $\gamma_2 < 1$. In addition, note that the mixing condition assumed in Section 4 is stronger than that for the Gaussian case.

The rate of convergence of the estimators of largest eigenvalue $\phi_{\text{max}}^2$ and the corresponding normalized eigenvector $q^0$ can be derived as well as the $\alpha$-mixing Gaussian case.

5 L0-PENALIZED ESTIMATOR FOR GAUSSIAN PROCESS

In this section, we discuss the $l_0$-penalized estimator $\hat{\beta}^0_n$ for $\alpha$-mixing Gaussian stationary processes for comparison. The estimator $\hat{\beta}^0_n$ is defined as

$$\hat{\beta}^0_n := \arg\min_{\beta \in B} \{ R_n(\beta) + \lambda_0 ||\beta||_0 \}. \quad (9)$$
where $\lambda_0 \geq 0$ is a tuning parameter for $l_0$-penalized estimator. Note that for the penalized estimator (9), there exists a nonnegative constant $s \geq 0$ such that

$$\hat{\beta}_n^0 := \arg\min_{\beta \in \mathcal{B}_0(s)} R_n(\beta),$$

where

$$\mathcal{B}_0(s) := \{ \beta \in \mathbb{R}^p : \|\beta\|_0 \leq s \}.$$

Let $s_0 = \max\{s_0, s\}$, and $\zeta_n$ be a sequence such that

$$\tilde{\zeta}_n := \sqrt{\left(\tilde{b} + 4s\right) \log p \over cn},$$

where $\tilde{b}$ and $c$ are some positive constants. Let $\tilde{\gamma}_n$ be defined as

$$\tilde{\gamma}_n = \phi_{\max}\left(\tilde{\zeta}_n \sum_{l=0}^{n} \rho(l)\right).$$

This is parallel to the definition of $\gamma_n$ in (6) for Lasso-type estimator. The oracle inequality is then derived as follows.

**Theorem 3.** Suppose that Assumptions 1 and 2 hold. If

$$\tilde{s} \log p \over n \leq c \over 4\tilde{b},$$

it holds that

$$P\left(\|\hat{\beta}_n^0 - \beta^0\|_2 \leq \delta_n\right) \geq 1 - 2 \exp\left(-\tilde{b} \log p\right) - 2 \exp(-b \log p),$$

where

$$\delta_n := \sqrt{2\tilde{s}\gamma_n \|\beta^0\|_1 + 2\tilde{s}\gamma_n^2 \|\beta^0\|_1^2 + 4(\sigma - 3\eta - \tilde{\gamma}_n)s_0 \lambda_0 \over 2(\sigma - 3\eta - \tilde{\gamma}_n)}.$$

**Remark 6.** By the inequality of arithmetic and geometric means, it is easy to see that $\delta_n$ takes its lower bound of the order when it holds that

$$\sqrt{2\tilde{s}\gamma_n \|\beta^0\|_1} = \sqrt{\tilde{s}\gamma_n^2 \|\beta^0\|_1^2 + 4(\sigma - 3\eta - \tilde{\gamma}_n)s_0 \lambda_0}.$$

In other words, $\lambda_0 = O_p(\sqrt{s_0 \gamma_n})$. This implies that $\delta_n = O_p\left(\sqrt{s} \|\beta^0\|_1 \gamma_n\right)$. Now we can establish the rate of convergence of the estimator $\hat{\beta}_n^0$ as follows.

**Corollary 4.** Suppose that $\sum_{l=0}^{\infty} \rho(l) < \infty$, $\log p \to \infty$ as $n \to \infty$. If $\lambda_0$ and $\tilde{s}$ satisfy that

$$\tilde{s} \log p \over n \to 0, \quad n \to \infty,$$
and
\[ \lambda_0 \asymp \phi_{\text{max}}^4 \frac{\bar{s} \log p}{n}. \]

Then, under the same assumptions in Theorem 3, it holds that
\[ \| \hat{\beta}_n^0 - \beta^0 \|_2^2 = O_p \left( \bar{s} s_0 \phi_{\text{max}}^4 \frac{\log p}{n} \right), \quad n \to \infty. \]

Remark 7. We do not assume some conditions such that \( p = q_n n \) for some sequence \( q_n > 0 \) and \( \phi_{\text{max}}^2 = O(q_n^{-1/2}) \) to obtain the rate of convergence of the \( l_0 \) penalized estimator. It can be seen from Corollaries 1 and 4 that \( l_0 \)-penalized estimators and Lasso-type estimators have similar rate of convergence \( O_p(\log p/n) \) in terms of squared errors if \( \phi_{\text{max}}^2, s_0 \) and \( \bar{s} \) are constant order.

6 | SIMULATION STUDIES

In this section, we investigate the finite sample performance of the sparse PCA for stationary processes. We also provide a real data example of average temperatures in Kyoto analyzed by the sparse PCA.

6.1 | Finite sample performance

We generate the observation stretch from the model (1). The first principal component vector \( p_1 \) of the coefficient matrix \( A \) is supposed to be
\[ p_1 = \frac{(f(1/p), \ldots, f(p/p))^T}{\| (f(1/p), \ldots, f(p/p)) \|}, \]
where the function \( f \) is specified by each one of the functions in Figure 1. The function in the left figure is known as the three-peak function and the function in the right figure is known as the step function.

FIGURE 1 Functions to generate the first principal component vector \( p_1 \).
The eigenvalues \( v_1, \ldots, v_p \) of \( A \) are determined by \( v_j = v_j^j, \ j = 1, \ldots, p \), where \( v \in \{0.85, 0.6, 0.35, 0.1\} \). Under this setting, the ratio of the first eigenvalue of the covariance matrix \( \Sigma_0 \) to the second eigenvalue is \( 1 + v^2 \). We compare the average squared errors \( \| \hat{\beta}_n - \beta^0 \| / p \) in \( l_0 \)-Penalized estimator, \( l_1 \)-Penalized estimator and standard PCA. We implemented the simulations by the iterative thresholding sparse PCA which is provided by Ma (2013). Note that this algorithm is known to lead to the same result as the penalized Frobenius minimization problem (see Shen & Huang, 2008). The dimension \( p \) of the stationary process is specified as \( p = 512 \), while the number of the observation is specified as \( n = 256 \). The penalty parameter \( \lambda \) is taken as \( \lambda_0 = 3 \log(p)/n \) and \( \lambda_1 = (\log(p)/n)^{1/2}/3 \). The numerical results reported in Table 1 are obtained over 1000 runs for the model (1) with i.i.d. innovations as centered Gaussian distribution and centered two-sided Weibull distribution with the shape parameter 0.5, respectively. Here, the covariance matrix of innovations is the identity matrix.

From Table 1, we see that the penalized PCA performs better than the standard PCA in terms of the loss for all cases. The penalized PCA show similar performance, while the \( l_1 \)-Penalized estimator performs better in almost all cases in these simulations. The \( l_0 \)-Penalized estimators are prone to choose a small number of features while the \( l_1 \)-Penalized estimators balance the average squared error and the number of features.

### Table 1 Comparison of principal component analyses (PCA).

|          | Gaussian | \( l_0 \)-Penalized estimator | \( l_1 \)-Penalized estimator | Standard PCA |
|----------|----------|-------------------------------|-------------------------------|--------------|
|          | \( v_1 \) | Loss (Size)                  | Loss (Size)                  | Loss         |
| Peak     | 0.85     | 0.00304 (29.01)              | 0.00314 (47.92)              | 0.00422      |
|          | 0.60     | 0.00350 (35.98)              | 0.00297 (49.43)              | 0.00513      |
|          | 0.35     | 0.00331 (35.73)              | 0.00273 (49.42)              | 0.00505      |
|          | 0.10     | 0.00324 (35.55)              | 0.00265 (49.42)              | 0.00501      |
| Step     | 0.85     | 0.00410 (42.35)              | 0.00411 (47.17)              | 0.00424      |
|          | 0.60     | 0.00362 (47.03)              | 0.00298 (51.58)              | 0.00513      |
|          | 0.35     | 0.00342 (46.89)              | 0.00272 (51.51)              | 0.00505      |
|          | 0.10     | 0.00336 (46.98)              | 0.00263 (51.41)              | 0.00501      |
| Weibull  | \( \nu_1 \) | Loss (Size)                  | Loss (Size)                  | Loss         |
| Peak     | 0.85     | 0.00566 ( 5.33)              | 0.00523 ( 9.07)              | 0.00607      |
|          | 0.60     | 0.00523 (13.61)              | 0.00470 (19.69)              | 0.00595      |
|          | 0.35     | 0.00508 (13.68)              | 0.00453 (19.37)              | 0.00584      |
|          | 0.10     | 0.00500 (13.92)              | 0.00444 (19.86)              | 0.00578      |
| Step     | 0.85     | 0.00587 (14.16)              | 0.00540 (20.08)              | 0.00603      |
|          | 0.60     | 0.00522 (14.01)              | 0.00467 (20.04)              | 0.00595      |
|          | 0.35     | 0.00506 (14.09)              | 0.00449 (19.98)              | 0.00582      |
|          | 0.10     | 0.00505 (13.91)              | 0.00448 (19.73)              | 0.00583      |

Notes: Peak refers to the three-peak function and Step refers to the step function. Loss refers to the average squared error and Size refers to the number of the estimated elements in the first principal component.
Real data example

We apply the sparse PCA to the dataset of daily average temperatures in Kyoto, Japan. The data are from January 1, 1901 to December 31, 2020, which are over the last 120 years. We removed the temperature of February 29, if exists, to make each year have 365 days. To summarize, the dimension of the data is 365 and the sample size is 120.

In general, the temperature data appear to increase over the years. Assuming a linear trend in the temperature, we removed the trend from the original data and obtained detrended data, of which the partial data are shown in the left figure in Figure 2. The detrended temperature data appear to be stationary. We also plot the eigenvalues of the covariance matrix obtained from the detrended data. The spikiness condition in Assumption 1 seems to be satisfied. The sparsity feature of these data can be confirmed from the right figure in Figure 2.

The sparse PCA and the standard PCA are applied to the data. The penalty parameter $\lambda$ is taken as $\lambda_0 = 3 \log(p)/n$ and $\lambda_1 = (\log(p)/n)^{1/2}/3$ for $l_0$-penalty and $l_1$-penalty, respectively. The numerical results are obtained as Figure 3. The nonzero coefficients are shown in colors with rainbow plots. The estimates of large absolute value are in red while those of small absolute value are in blue. We can also find that the $l_0$-penalized estimate shows the smallest number of features in the first principal component, compared with other two methods, while the $l_1$-penalized estimate harmonize the standard PCA with the $l_0$-penalized one.

The sparse PCAs explain the feature of the temperatures in Kyoto well. It is well known that the temperature in Kyoto is radically going up and down during February and March over years. Thus there is much more temperature variation during these months. This result matches the data provided by Japan Meteorological Agency. In summary, this feature is extracted by the sparse PCA.

7 | PROOFS

In this section, we complement the rigorous proofs for the main results in Sections 3–5. First, we summarize some crucial preliminary lemmas in Sections 7.1 and 7.2 for $\alpha$-mixing Gaussian processes and $\beta$-mixing sub-Weibull processes, respectively. The proofs for preliminary lemmas in Section 7.1 can be found in Section 7.3, while the proofs for preliminary lemmas in Section 7.2 can be found in Section 7.4. Section 7.5 offers the proofs for $l_0$-penalized estimators.
7.1 Lemmas for Gaussian process

We use the following concentration inequality, which is a modified form of Hanson–Wright’s inequality.

**Lemma 1.** Let $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$ be an $n$-dimensional normal random vector. Then, there exists a universal constant $c > 0$ such that for any $\eta > 0$,

$$
P \left( \frac{1}{n} \left| \left| \mathbf{Y} \right| \right|_2^2 - E[\left| \left| \mathbf{Y} \right| \right|_2^2] > \eta \left| \left| \mathbf{Q} \right| \right|_2 \right) \leq 2 \exp \left( -cn \min \{\eta, \eta^2\} \right).$$

See Rudelson and Vershynin (2013), Basu and Michailidis (2015), and Wong et al. (2020) for the detail of this inequality.

The next lemma guarantees that the empirical risk $R_n(\cdot)$ is also strictly convex on $\mathcal{B}$ with high probability.

**Lemma 2.** Suppose that the same assumptions as Proposition 1 and Assumption 2 hold. Then, for every $\delta > 0$, it holds that

$$
P \left( \sup_{\mathbf{v} \in \mathbb{S}^{p-1}} |\mathbf{v}^T \mathbf{W}_n \mathbf{v}| > 2\delta \right) \leq 2 \exp \left( p \log 6 - cn \min \{\tilde{\xi}_n, \tilde{\xi}_n^2\} \right),$$

where

$$
\tilde{\xi}_n = \frac{\delta}{\sum_{l=0}^{n-1} \rho(l) \phi_{\max}},
$$

and $c > 0$ is a universal constant appeared in Lemma 1.

Especially, if Assumption 3 holds, and $\delta$ satisfies

$$
\sqrt{\frac{q_n \log 6}{c}} C_{\rho} \phi_{\max} < \delta < \sigma - 3\eta,
$$

where

\begin{align*}
q_n &= \sum_{l=0}^{n-1} \rho(l) \phi_{\max}, \\
C_{\rho} &= \max_{0 \leq \rho \leq 1} \frac{\int_{\rho}^{\rho+1} \frac{d\phi(t)}{\phi(t)}}{\phi_{\max}},
\end{align*}

**Figure 3** Plots for the estimated first principal components. From the top to bottom, plots are for the $l_0$-penalized estimate, the $l_1$-penalized estimate, and the standard one.
For $\alpha$-mixing Gaussian processes, by Lemma 2, we find that for any convex penalty function $\text{pen}(\cdot)$, $R_n(\cdot) + \lambda \text{pen}(\cdot)$ is still asymptotically strictly convex on $B$, which follows from the fact that the conical combination of convex functions is also convex. This also implies that $\hat{\beta}_n^1$ is a unique solution to the optimization problem (2), that is, the Lasso-type PCA estimator is well-defined with high probability.

We establish the oracle inequality for the estimator $\hat{\beta}_n^1$. To this end, we evaluate the difference between the empirical risk and theoretical risk, which is achieved by the following lemma.

**Lemma 3.** Let $b > 0$ be a free parameter and $c > 0$ be a constant appeared in Lemma 1. Define that

$$\zeta_n := \sqrt{\frac{2(b + 1) \log p^2}{cn}}.$$

Under Assumptions 1 and 2 and the same assumption as Proposition 1, for every $n$ satisfying $\log p/n \leq 1$, it holds that

$$P\left(\|W_n\|_{\max} \leq \zeta_n \sum_{l=0}^{n} \rho(l)\phi_{\max}\right) \geq 1 - 2 \exp(-2b \log p),$$

where $W_n = \hat{\Sigma}_n - \Sigma_0$.

### 7.2 Lemmas for sub-Weibull process

To derive the oracle inequality for the $\beta$-mixing sub-Weibull process, we use the following concentration inequality, which is established by Merlevède et al. (2011).

**Lemma 4.** Let $\{X_t\}_{t \in \mathbb{Z}}$ be an $\mathbb{R}$-valued $\beta$-mixing zero mean strictly stationary process, which satisfies that

$$\beta(n) \leq 2 \exp(-cn^{\gamma_1}),$$

and

$$\|X_t\|_{\psi_{2}} \leq K, \quad \forall t \in \mathbb{Z},$$

for constants $c, \gamma_1, \gamma_2$, and $K > 0$. Let $\gamma$ be

$$\gamma := \left(\frac{1}{\gamma_1} + \frac{1}{\gamma_2}\right)^{-1} < 1.$$

Then, for every $n > 4$ and $\epsilon > 1/n$, it holds that

$$P\left(\left\|\frac{1}{n} \sum_{t=1}^{n} X_t \right\| > \epsilon\right) \leq n \exp\left(-\frac{(en)^{\gamma}}{K^2 C_1}\right) + \exp\left(-\frac{e^2 n}{K^2 C_2}\right).$$
where $C_1$ and $C_2$ are constants depending only on $\gamma_1, \gamma_2$, and $c$.

Using Lemma 4, we obtain the following result which is corresponding to the Lemma 2 for Gaussian case.

**Lemma 5.** Suppose that the same assumptions as Proposition 1 and Assumption 4 hold. Then, for every $\delta > 0$, it holds that

$$ P \left( \sup_{\nu \in \mathbb{S}^p} |\nu^T W_n \nu| > 2\delta \right) \leq n \exp \left( p \log 6 - \frac{(\delta n)^{\gamma_1}}{K_2 C_1} \right) + \exp \left( p \log 6 - \frac{\delta^2 n}{K_2 C_2} \right), $$

where $C_1$ and $C_2$ are constants depending only on $c, \gamma_1$ and $\gamma_2$ described in Assumption 4 and $K_2 := 2^{2/\gamma_2 + 1} K_1^2$.

Especially, if $p = q_n n^\gamma$, where $q_n > 0$ is a sequence such that

$$ q_n < \frac{(\sigma - 3\eta)^{\gamma_1}}{K_2^1 C_1 \log 6}, \tag{11} $$

$$ K_2^1 C_1^{1/\gamma} q_n^{1/\gamma} (\log 6)^{1/\gamma} < \delta < \sigma - 3\eta, \tag{12} $$

for some $\delta > 0$ and $K_2 q_n^{1/\gamma} = O(1)$, it holds that

$$ \lim_{n \to \infty} P \left( \inf_{\nu \in \mathbb{S}^p} \nu^T \hat{R}_n(\beta) \nu > 0 \right) = 1, \quad \beta \in B. $$

Therefore, we conclude that for $\beta$-mixing sub-Weibull processes, $R_n(\cdot) + \lambda \quad \text{pen}(\cdot)$ is also asymptotically strictly convex on $B$ with high probability for any convex penalty function $\text{pen}(\cdot)$. As for the bound for $\|W_n\|_{\max}$, we have the following lemma.

**Lemma 6.** Suppose that Assumptions 1 and 4 hold. For every $\zeta_n$ satisfying

$$ \zeta_n > \max \left\{ \frac{K_2 C_1^{1/\gamma} (\log np^2 + 2\tilde{b} \log p)^{1/\gamma}}{n}, K_2 \sqrt{\frac{2C_2 (\tilde{b} + 1) \log p}{n}} \right\}, $$

where $\tilde{b} > 0$ is a free parameter, it holds that

$$ P \left( \|W_n\|_{\max} \leq \zeta_n \right) \geq 1 - 2 \exp(-\tilde{b} \log p^2). $$

### 7.3 Proofs for Sections 7.1 and 3

In this subsection, we provide proofs for main results and lemmas for Lasso-type estimator for $\alpha$-mixing Gaussian process.

**Proof of Lemma 2.** Put $X(n) = (X_1, \ldots, X_n)$. It follows from the stationarity of $\{X_t\}_{t \in \mathbb{Z}}$ that

$$ \hat{\Sigma}_n = \frac{1}{n} X(n) X(n)^T, \quad \Sigma_0 = \frac{1}{n} E[X(n) X(n)^T], $$

and the proof is similar to the Gaussian case.
which implies that
\[
\mathbf{v}^\top W_n \mathbf{v} = \frac{1}{n} \mathbf{v}^\top (X_{(n)} X_{(n)}^\top - E[X_{(n)} X_{(n)}^\top]) \mathbf{v} \\
= \frac{1}{n} (\|X_{(n)}^\top \mathbf{v}\|_2^2 - E[\|X_{(n)}^\top \mathbf{v}\|_2^2]).
\]

Let \( Q_n \) be the covariance matrix of the random variable \( X_{(n)}^\top \mathbf{v} \). By the simple calculation, we can find that
\[
Q_n = \begin{pmatrix}
\mathbf{v}^\top E[X_{1} X_{1}^\top] & \cdots & \mathbf{v}^\top E[X_{1} X_{n}^\top] \\
\vdots & \ddots & \vdots \\
\mathbf{v}^\top E[X_{n} X_{1}^\top] & \cdots & \mathbf{v}^\top E[X_{n} X_{n}^\top]
\end{pmatrix}.
\]

Noting that \( \{X_t\}_{t \in \mathbb{Z}} \) is a centered Gaussian time series, we have that \( X_{(n)}^\top \mathbf{v} \sim \mathcal{N}(0, Q_n) \). We can apply Lemma 1 to \( X_{(n)}^\top \mathbf{v} \) to deduce that for every \( \xi > 0 \),
\[
P \left( \frac{1}{n} \left| \|X_{(n)}^\top \mathbf{v}\|_2^2 - E[\|X_{(n)}^\top \mathbf{v}\|_2^2] \right| > \xi \|Q_n\|_2 \right) \leq 2 \exp(-cn \min\{\xi, \xi^2\}),
\]
where \( c > 0 \) is a universal constant. Noting that the \( \alpha \)-mixing Gaussian time series \( \{X_t\}_{t \in \mathbb{Z}} \) is also \( \rho \)-mixing, we have
\[
\|Q_n\|_2 \leq \sum_{l=0}^{n} \rho(l) \|\Sigma_0\|_2 = \sum_{l=0}^{n} \rho(l) \varphi_{\max}.
\]

We therefore obtain that
\[
P \left( \frac{1}{n} \left| \|X_{(n)}^\top \mathbf{v}\|_2^2 - E[\|X_{(n)}^\top \mathbf{v}\|_2^2] \right| > \xi \sum_{l=0}^{n} \rho(l) \varphi_{\max} \right) \leq 2 \exp(-cn \min\{\xi, \xi^2\}).
\]

Especially, we can take \( \xi \) satisfying that
\[
\xi = \frac{\delta}{\sum_{l=0}^{n} \rho(l) \varphi_{\max}},
\]
which concludes that
\[
P (|\mathbf{v}^\top W_n \mathbf{v}| > \delta) \leq 2 \exp(-cn \min\{\xi, \xi^2\}),
\]
for every \( \mathbf{v} \in \mathbb{S}^{p-1} \). Then, we apply the \( \epsilon \)-net argument described in chapter 4 of Vershynin (2018). Let \( \mathcal{N}_{1/2} \) be the \( 1/2 \)-net of \( \mathbb{S}^{p-1} \), that is, for every \( \mathbf{v} \in \mathbb{S}^{p-1} \), there exists \( \mathbf{v}_0 \in \mathcal{N}_{1/2} \) such that
\[
|\mathbf{v} - \mathbf{v}_0|_2 \leq \frac{1}{2}.
\]

Then, by lemma 4.4.1 of Vershynin (2018), we have that
\[
\sup_{\mathbf{v} \in \mathbb{S}^{p-1}} |\mathbf{v}^\top W_n \mathbf{v}| \leq 2 \sup_{\mathbf{v} \in \mathcal{N}_{1/2}} |\mathbf{v}^\top W_n \mathbf{v}|.
\]
Moreover, it is known that $|\mathcal{N}_{1/2}| \leq 6^p$. We thus have that

$$P \left( \sup_{\mathbf{v} \in \mathbb{S}^{p-1}} |\mathbf{v}^T \mathbf{W}_n \mathbf{v}| > 2\delta \right) \leq P \left( 2 \sup_{\mathbf{v} \in \mathcal{N}_{1/2}} |\mathbf{v}^T \mathbf{W}_n \mathbf{v}| > 2\delta \right)$$

$$= \sum_{\mathbf{v} \in \mathcal{N}_{1/2}} P \left( |\mathbf{v}^T \mathbf{W}_n \mathbf{v}| > \delta \right)$$

$$= 6^p \cdot 2 \exp(-cn \min\{\xi, \xi^2\}).$$

which ends the proof.

**Proof of Lemma 3.** For every $i, j \in \{1, \ldots, p\}$, it holds that

$$W_{nij} = e_i^T \mathbf{W}_n e_j,$$

where $e_k, k = 1, \ldots, p$ is the $k$-th canonical basis of $\mathbb{R}^p$. Since

$$\mathbf{W}_n = \hat{\Sigma}_n - \Sigma_0 = \frac{1}{n} (\mathbf{X}(n)\mathbf{X}(n)^T - E[\mathbf{X}(n)\mathbf{X}(n)^T]),$$

it holds that

$$W_{nij} = \frac{1}{n} \left( e_i^T \mathbf{X}(n) \mathbf{X}(n)^T e_j - E[e_i^T \mathbf{X}(n) \mathbf{X}(n)^T e_j] \right).$$

Putting $\mathbf{Y}_k = \mathbf{X}(n)^T e_k, k = 1, \ldots, p$, we can rewrite that

$$W_{nij} = \frac{1}{n} \left( \mathbf{Y}_i^T \mathbf{Y}_j - E[\mathbf{Y}_i^T \mathbf{Y}_j] \right).$$

Note that

$$\mathbf{Y}_i^T \mathbf{Y}_j - E[\mathbf{Y}_i^T \mathbf{Y}_j] = \frac{1}{2} \left\{ (\|\mathbf{Y}_i + \mathbf{Y}_j\|_2^2 - E[\|\mathbf{Y}_i + \mathbf{Y}_j\|_2^2]) + (\|\mathbf{Y}_i\|_2^2 - E[\|\mathbf{Y}_i\|_2^2]) = (\|\mathbf{Y}_j\|_2^2 - E[\|\mathbf{Y}_j\|_2^2]) \right\}. $$

We then find that

$$|W_{nij}| = \frac{1}{n} \left| \mathbf{Y}_i^T \mathbf{Y}_j - E[\mathbf{Y}_i^T \mathbf{Y}_j] \right|$$

$$\leq \frac{1}{2n} \left| (\|\mathbf{Y}_i + \mathbf{Y}_j\|_2^2 - E[\|\mathbf{Y}_i + \mathbf{Y}_j\|_2^2]) + (\|\mathbf{Y}_i\|_2^2 - E[\|\mathbf{Y}_i\|_2^2]) \right| + \frac{1}{2n} \left| (\|\mathbf{Y}_j\|_2^2 - E[\|\mathbf{Y}_j\|_2^2]) \right|.$$
and
\[ P \left( \frac{1}{n} \left\| Y_i + Y_j \right\|^2_2 - E[\left\| Y_i + Y_j \right\|^2_2] > \zeta \left\| \Sigma_{Y_i,Y_j} \right\|^2_2 \right) \leq 2 \exp(-cn \min\{\zeta, \zeta^2\}). \]

After some tedious computation, we have
\[ \left\| \Sigma_{Y_i,Y_j} \right\|^2_2 \leq \frac{\left\| \Sigma_{Y_i} \right\|^2_2 + \left\| \Sigma_{Y_j} \right\|^2_2}{2}. \]

Therefore, we have that
\[ P \left( \frac{1}{n} \left\| Y_i + Y_j \right\|^2_2 - E[\left\| Y_i + Y_j \right\|^2_2] > \zeta \left\| \Sigma_{Y_i} \right\|^2_2 + \left\| \Sigma_{Y_j} \right\|^2_2 \right) \leq 2 \exp(-cn \min\{\zeta, \zeta^2\}). \]

(15)
The inequalities (13)–(15) imply that for every \( \zeta > 0 \),
\[ P(|W_nij| > \zeta(\left\| \Sigma_{Y_i} \right\|^2_2 + \left\| \Sigma_{Y_j} \right\|^2_2)) \leq 6 \exp(-cn \min\{\zeta, \zeta^2\}). \]

Let \( b > 0 \) be a free parameter. We take \( \zeta \) as
\[ \zeta := \zeta_n = \sqrt{\frac{(b + 1) \log p^2}{cn}}. \]

Then, for every \( n \) such that \( \zeta_n^2 < \zeta_n \), it holds that
\[ P \left( \|W_n\|_{\max} > 4\zeta \sum_{l=0}^n \rho(l)\phi_{\max} \right) \leq P \left( \|W_n\|_{\max} > \max_{i,j} \zeta(\left\| \Sigma_{Y_i} \right\|^2_2 + \left\| \Sigma_{Y_j} \right\|^2_2) \right) \]
\[ = P \left( \max_{i,j} |e_i^T W_n e_j| > \max_{i,j} \zeta(\left\| \Sigma_{Y_i} \right\|^2_2 + \left\| \Sigma_{Y_j} \right\|^2_2) \right) \]
\[ \leq \sum_{i,j} P \left( |e_i^T W_n e_j| > \zeta(\left\| \Sigma_{Y_i} \right\|^2_2 + \left\| \Sigma_{Y_j} \right\|^2_2) \right) \]
\[ \leq 6p^2 \exp(-cn \min\{\zeta, \zeta^2\}) \]
\[ = 6 \exp\left( \log p^2 - cn \frac{(b + 1) \log p^2}{cn} \right) \]
\[ = 6 \exp(-b \log p^2), \]

which completes the proof.

\[ \text{Proof of Theorem 1.} \] Let \( \gamma_n \) be
\[ \gamma_n = \zeta_n \sum_{l=0}^n \rho(l)\phi_{\max}. \]

For the constant \( \delta < \sigma - 3\eta \), it suffices to show the inequality
\[ \| \hat{\beta}_n^1 - \beta^0 \|_1 \leq \frac{2(C + 1)^2 \lambda_1 s_0}{\sigma - 3\eta - \delta}, \]
on the event 
\[ \{ \| W_n \|_{\max} \leq \gamma_n \} \cap \left\{ \sup_{v \in \mathbb{S}^{p-1}} |v^T W_n v| \leq \delta \right\}. \]

Following lemma 7.1 of van de Geer (2016), it holds that
\[ -R_n (\hat{\beta}_n^1)^T (\beta^0 - \hat{\beta}_n^1) \leq \lambda_1 \| \beta^0 \|_1 - \lambda_1 \| \hat{\beta}_n^1 \|_1. \tag{17} \]

Moreover, it follows from Proposition 1 and the Taylor expansion that
\[ R(\beta^0) - R(\hat{\beta}_n^1) - R(\hat{\beta}_n^1)^T (\beta^0 - \hat{\beta}_n^1) \geq (\sigma - 3\eta) \| \beta^0 - \hat{\beta}_n^1 \|_2^2 \geq 0. \tag{18} \]

Combining (17) and (18), we have that
\[ R(\hat{\beta}_n^1) - R(\beta^0) + \lambda_1 \| \hat{\beta}_n^1 \|_1 \leq \left( R_n (\hat{\beta}_n^1 - R(\hat{\beta}_n^1))^T (\beta^0 - \hat{\beta}_n^1) + \lambda_1 \| \beta^0 \|_1. \tag{19} \]

Noting that
\[ \hat{R}_n (\hat{\beta}_n^1) - R(\beta^0) = -W_n \hat{\beta}_n^1, \]

we have, on the event \( \{ \sup_{v \in \mathbb{S}^{p-1}} |v^T W_n v| \leq \delta \} \),
\[ \left( \hat{R}_n (\hat{\beta}_n^1) - R(\hat{\beta}_n^1) \right)^T (\beta^0 - \hat{\beta}_n^1) = (\hat{\beta}_n^1 - \beta^0)^T W_n (\hat{\beta}_n^1 - \beta^0) + \beta^0^T W_n (\hat{\beta}_n^1 - \beta^0) \leq \delta \| \beta_n^1 - \beta^0 \|_2^2 + \| W_n \|_{\max} \| \beta^0 \|_1 \| \hat{\beta}_n^1 - \beta^0 \|_1. \tag{20} \]

Since \( \beta^0 \) satisfies that \( \hat{R}(\beta^0) = 0 \), it follows from Proposition 1 and the Taylor expansion that
\[ R(\hat{\beta}_n^1) - R(\beta^0) \geq (\sigma - 3\eta) \| \beta_n^1 - \beta^0 \|_2^2, \]
which implies
\[ R(\hat{\beta}_n^1) - R(\beta^0) + \lambda_1 \| \hat{\beta}_n^1 \|_1 \geq (\sigma - 3\eta) \| \beta_n^1 - \beta^0 \|_2^2 + \lambda_1 \| \hat{\beta}_n^1 \|. \tag{21} \]

Therefore, by (21), (19), and (20), we find that, on the event \( \{ \| W_n \|_{\max} \leq \gamma_n \} \),
\[ (\sigma - 3\eta) \| \beta_n^1 - \beta^0 \|_2^2 + \lambda_1 \| \hat{\beta}_n^1 \|_1 \leq \eta_n \| \beta_n^1 - \beta^0 \|_2^2 + \gamma_n \| \beta^0 \|_1 \| \hat{\beta}_n^1 - \beta^0 \|_1 + \lambda_1 \| \beta^0 \|_1, \]
and thus,
\[ (\sigma - 3\eta - \delta) \| \beta_n^1 - \beta^0 \|_2^2 \leq \gamma_n \| \beta^0 \|_1 \| \beta_n^1 - \beta^0 \|_1 + \lambda_1 \| \beta^0 \|_1 - \lambda_1 \| \hat{\beta}_n^1 \|. \tag{22} \]

The right-hand side of (22) can be bounded as follows.
\[ \gamma_n \| \beta^0 \|_1 \| \beta_n^1 - \beta^0 \|_1 + \lambda_1 \| \beta^0 \|_1 - \lambda_1 \| \hat{\beta}_n^1 \|_1 \]
\[ = \gamma_n \| \beta^0 \|_1 \| \beta_{nS} - \beta^0 \|_1 + \gamma_n \| \beta^0 \|_1 \| \hat{\beta}_{nS} - \beta^0 \|_1 \]
In this subsection, we provide proofs for main results and lemmas for Lasso-type estimator for \( \beta \)-mixing sub-Weibull process.

Proof of Lemma 5. For every \( \mathbf{v} \in \mathbb{S}^{p-1} \), it holds that

\[
\mathbf{v}^T W_n \mathbf{v} = \frac{1}{n} (\|X^T_{(n)} \mathbf{v}\|^2_2 - E[\|X^T_{(n)} \mathbf{v}\|^2_2]).
\]

Define the process \( \{Z_t(\mathbf{v})\}_{t \in \mathbb{N}} \) by

\[
Z_t(\mathbf{v}) = |\mathbf{v}^T X_t|^2 - E[|\mathbf{v}^T X_t|^2].
\]

7.4 Proofs for Sections 7.2 and 4

In this subsection, we provide proofs for main results and lemmas for Lasso-type estimator for \( \beta \)-mixing sub-Weibull process.

Proof of Lemma 5. For every \( \mathbf{v} \in \mathbb{S}^{p-1} \), it holds that

\[
\mathbf{v}^T W_n \mathbf{v} = \frac{1}{n} (\|X^T_{(n)} \mathbf{v}\|^2_2 - E[\|X^T_{(n)} \mathbf{v}\|^2_2]).
\]
Then, we have that
\[ \|X_{(n)}^T\nu\|_2^2 - E[\|X_{(n)}^T\nu\|_2^2] = \sum_{i=1}^{n} Z_i(\nu). \]

Since \( \{X_t\}_{t \in \mathbb{Z}} \) is the sub-Weibull(\( \gamma_2 \)), the process \( \{Z_t(\nu)\}_{t \in \mathbb{Z}} \) is the sub-Weibull(\( \gamma_2/2 \)).

Therefore, it follows from Lemma 4 that
\[
P(|\nu^T W_n \nu| > \delta) = P\left( \left| \frac{1}{n} \sum_{t=1}^{n} Z_t(\nu) \right| > \delta \right)
\leq n \exp\left( -\left( \frac{\delta n}{K_2 C_1} \right)^{\frac{\gamma}{2}} \right) + \exp\left( -\frac{\delta^2 n}{K_2^2 C_2} \right).
\]

The union bound over \( \mathbb{S}^{p-1} \) is shown in the similar way as Gaussian case.

\[ \Box \]

**Proof of Lemma 6.** As well as Proof of Lemma 3, we can find that
\[
|W_{nij}| \leq \frac{1}{2n} \left| \left( \|Y_i + Y_j\|_2^2 - E[\|Y_i + Y_j\|_2^2] \right) \right|
+ \frac{1}{2n} \left| \left( \|Y_i\|_2^2 - E[\|Y_i\|_2^2] \right) \right|
+ \frac{1}{2n} \left| \left( \|Y_j\|_2^2 - E[\|Y_j\|_2^2] \right) \right|
\]
where \( Y_k = X_{(n)}^T e_k, \ k = 1, \ldots, p \). Note that
\[
\|Y_i\|_2^2 = \sum_{t=1}^{n} X_{i,t}^2, \quad \|Y_i + Y_j\|_2^2 = \sum_{t=1}^{n} (X_{i,t} + X_{j,t})^2, \quad i, j = 1, \ldots, p,
\]
where \( X_{i,t} \) is the \( i \)th component of \( X_t \). Since the process \( \{X_t\}_{t \in \mathbb{Z}} \) is sub-Weibull(\( \gamma_2 \)), it follows that \( \{X_{i,t}^2\}_{t \in \mathbb{Z}} \) is sub-Weibull(\( \gamma_2/2 \)). More precisely, we have that
\[
\|X_{i,t}^2\|_{\psi_{2/2}} \leq 2^{2/\gamma_2} \|X_{i,t}\|_{\psi_{2}}^2 \leq 2^{2/\gamma_2} K_2 \leq K_2,
\]
and
\[
\|(X_{i,t} + X_{j,t})^2\|_{\psi_{2/2}} \leq 2^{2/\gamma_2} \|X_{i,t} + X_{j,t}\|_{\psi_{2}}^2 \leq 2^{2/\gamma_2} 2K_2 = K_2.
\]
Therefore, we can apply Lemma 4 and take a union bound over \( \{e_1, \ldots, e_p\}^2 \) to deduce the conclusion.

\[ \Box \]

**Proof of Theorem 2.** It suffices to show the inequality on the event
\[
\{ \|W_n\|_{\max} \leq \zeta_n \} \cap \left\{ \sup_{\nu \in \mathbb{S}^{p-1}} \nu^T W_n \nu \leq \xi_n \right\}.
\]

The remaining part of the proof is similar to the proof of Theorem 1.

\[ \Box \]

**7.5 l₀-penalized estimator for α-mixing Gaussian process**

In this subsection, we prove main results for the \( l_0 \)-penalized estimator for \( \alpha \)-mixing process.
Lemma 7. Suppose that Assumptions 1 and 2 hold. Let $\zeta_n$ be

$$\tilde{\zeta}_n = \sqrt{\frac{(b + 4\bar{s}) \log p}{cn}},$$

where $b > 0$ is a free parameter and $c > 0$ is a constant. For every $p > 6$ and $n$ such that $\tilde{\zeta}_n^2 < \tilde{\zeta}_n$, it holds that

$$P\left( \sup_{v \in B_0(2\bar{s}) \cap \mathbb{S}^{p-1}} |v^T W_n v| > \tilde{\zeta}_n \sum_{l=0}^{n} \rho(l) \phi_{\max} \right) \leq \exp(-b \log p).$$

Proof of Lemma 7. First, we fix $v \in B_0(2\bar{s}) \cap \mathbb{S}^{p-1}$ arbitrarily. In view of the proof of Lemma 2, we have that

$$|v^T W_n v| = \frac{1}{n} \left| \|X(n) v\|_2^2 - E[\|X(n) v\|_2^2]\right|.$$ 

It follows from Lemma 1 that there exists a constant $c > 0$ such that for every $\zeta > 0$,

$$P\left( \frac{1}{n} \left| \|X(n) v\|_2^2 - E[\|X(n) v\|_2^2]\right| > \zeta \|Q_n\|_2 \right) \leq 2 \exp(-cn \min\{\zeta, \zeta^2\}),$$

where $Q_n$ is the covariance matrix of $X(n)$ $v$. Noting that

$$\|Q_n\|_2 \leq \sum_{l=0}^{n} \rho(l) \phi_{\max},$$

we have

$$P\left( \frac{1}{n} \left| \|X(n) v\|_2^2 - E[\|X(n) v\|_2^2]\right| > \zeta \sum_{l=0}^{n} \rho(l) \phi_{\max} \right) \leq 2 \exp(-cn \min\{\zeta, \zeta^2\}).$$

Then, we apply the $\epsilon$-net argument. Let $\tilde{\mathcal{N}}_{1/2}$ be the $1/2$-net of $B_0(2\bar{s}) \cap \mathbb{S}^{p-1}$. It is easy to see that

$$|\tilde{\mathcal{N}}_{1/2}| \leq \left( \frac{p}{2\bar{s}} \right)^{2\bar{s}},$$

which deduce that

$$P\left( \sup_{v \in B_0(2\bar{s}) \cap \mathbb{S}^{p-1}} \frac{1}{n} \left| \|X(n) v\|_2^2 - E[\|X(n) v\|_2^2]\right| > 2 \zeta \sum_{l=0}^{n} \rho(l) \phi_{\max} \right) \leq \sum_{v \in \tilde{\mathcal{N}}_{1/2}} P\left( \frac{1}{n} \left| \|X(n) v\|_2^2 - E[\|X(n) v\|_2^2]\right| > \zeta \sum_{l=0}^{n} \rho(l) \phi_{\max} \right) \leq \left( \frac{p}{2\bar{s}} \right)^{2\bar{s}} \cdot 2 \exp(-cn \min\{\zeta, \zeta^2\}) \leq (6p)^{2\bar{s}} \cdot 2 \exp(-cn \min\{\zeta, \zeta^2\}).$$
\[
= 2 \exp(2\delta \log p + 2\delta \log 6 - cn \min\{\zeta, \zeta^2\}) \\
\leq 2 \exp(4\delta \log p - cn \min\{\zeta, \zeta^2\}).
\]

If we replace \(\zeta\) with \(\tilde{\zeta}_n\), then we obtain the conclusion. \(\blacksquare\)

**Proof of Theorem 3.** It suffices to show the inequality
\[
\|\hat{\beta}_n^0 - \beta^0\|_2 \leq \delta_n,
\]
on the event
\[
\{|\|W_n\|_\text{max} \leq \gamma_n\} \cap \left\{ \sup_{v \in B_0(2\delta) \cap S^{p-1}} |v^T W_n v| \leq \tilde{\gamma}_n \right\}.
\]

Note that \(\hat{\beta}_n^0 - \beta^0 \in B_0(2\delta)\) and that
\[
R_n(\hat{\beta}_n^0) + \lambda_0\|\hat{\beta}_n^0\|_0 \leq R_n(\beta^0) + \lambda_0\|\beta^0\|_0.
\]
It follows from the Taylor expansion and the fact that \(\hat{R}(\beta^0) = 0\) that
\[
\lambda_0\|\beta^0\|_0 - \lambda_0\|\hat{\beta}_n^0\|_0 \geq R_n(\hat{\beta}_n^0) - R_n(\beta^0)
\]
\[
= \hat{R}_n(\beta^0)^T(\hat{\beta}_n^0 - \beta^0) + \frac{1}{2}(\hat{\beta}_n^0 - \beta^0)^T \hat{R}_n(\hat{\beta}_n^0)(\hat{\beta}_n^0 - \beta^0)
\]
\[
\geq (\hat{R}_n(\beta^0) - \hat{R}(\beta^0))^T(\hat{\beta}_n^0 - \beta^0) + \hat{R}(\beta^0)^T(\hat{\beta}_n^0 - \beta^0)
\]
\[
+ (\sigma - 3\eta - \tilde{\gamma}_n)\|\hat{\beta}_n^0 - \beta^0\|_2^2
\]
\[
= -\beta^0^T W_n(\hat{\beta}_n^0 - \beta^0) + (\sigma - 3\eta - \tilde{\gamma}_n)\|\hat{\beta}_n^0 - \beta^0\|_2^2.
\]

We thus obtain that
\[
(\sigma - 3\eta - \tilde{\gamma}_n)\|\hat{\beta}_n^0 - \beta^0\|_2^2 \leq \beta^0^T W_n(\hat{\beta}_n^0 - \beta^0) + \lambda_0\|\beta^0\|_0 - \lambda_0\|\hat{\beta}_n^0\|_0
\]
\[
\leq |W_n|_{\text{max}}\|\beta^0\|_1\|\hat{\beta}_n^0 - \beta^0\|_1 + \lambda_0\|\beta^0\|_0
\]
\[
\leq \sqrt{2\gamma_n}\|\beta^0\|_1\|\hat{\beta}_n^0 - \beta^0\|_2 + \lambda_0s_0.
\]

We reach the conclusion after solving the quadratic inequality with respect to
\[\|\hat{\beta}_n^0 - \beta^0\|_2.\] \(\blacksquare\)

**ACKNOWLEDGMENTS**
The authors are grateful to the associate editor and reviewers for their insightful comments that have helped to improve the paper. The last two authors would like to express their thanks to the Institute for Mathematical Science (IMS) and Research Institute for Science & Engineering, Waseda University, respectively, for their support. Kou Fujimori is supported by JSPS Grant-in-Aid for Early-Career Scientists 21K13271. Yuichi Goto is supported by JSPS Grant-in-Aid for Research Activity Start-up 21K20338 and for Early-Career Scientists 23K16851. Yan Liu is supported by JSPS Grant-in-Aid for Scientific Research (C) 20K11719 and 23K11018. M. Taniguchi is supported by JSPS Grant-in-Aid for Scientific Research (S) 18H05290.
REFERENCES

Amini, A. A., & Wainwright, M. J. (2009). High-dimensional analysis of semidefinite relaxations for sparse principal components. *Annals of Statistics*, 37(5B), 2877–2921.

Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics*, 34(1), 122–148.

Basu, S., & Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *Annals of Statistics*, 43(4), 1535–1567.

Berthet, Q., & Rigollet, P. (2013). Optimal detection of sparse principal components in high dimension. *Annals of Statistics*, 41(4), 1780–1815.

Birnbaum, A., Johnstone, I. M., Nadler, B., & Paul, D. (2013). Minimax bounds for sparse pca with noisy high-dimensional data. *Annals of Statistics*, 41(3), 1055–1084.

Bühlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Springer.

Cai, T. T., Ma, Z., & Wu, Y. (2013). Sparse pca: Optimal rates and adaptive estimation. *Annals of Statistics*, 41(6), 3074–3110.

Cai, T. T., & Zhou, H. H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Annals of Statistics*, 40(5), 2389–2420.

Jin, B., Wang, C., Miao, B., & Huang, M.-N. L. (2009). Limiting spectral distribution of large-dimensional sample covariance matrices generated by varma. *Journal of Multivariate Analysis*, 100(9), 2112–2125.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, 29(2), 295–327.

Johnstone, I. M., & Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486), 682–693.

Jolliffe, I. T. (2002). *Principal component analysis*. Springer.

Kokoszka, P., & Kulik, R. (2023). Principal component analysis of infinite variance functional data. *Journal of Multivariate Analysis*, 193, 105–123.

Kokoszka, P., Stoev, S., & Xiong, Q. (2019). Principal components analysis of regularly varying functions. *Bernoulli*, 25(4B), 3864–3882.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.

Ma, Z. (2013). Sparse principal component analysis and iterative thresholding. *Annals of Statistics*, 41(2), 772–801.

Mackey, L. (2008). Deflation methods for sparse PCA. *Advances in Neural Information Processing Systems*, 21. https://proceedings.neurips.cc/paper_files/paper/2008/hash/85d8ce590ad8981ca2c8286f79f59954-Abstract.html

Merlevède, F., Peligrad, M., & Rio, E. (2011). A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, 151(3-4), 435–474.

Moshksar, K. (2021). On the absolute constant in Hanson-Wright inequality. *arXiv preprint arXiv:2111.00557*.

Paul, D., & Johnstone, I. M. (2012). Augmented sparse principal component analysis for high dimensional data. *arXiv:1202.1242v1*.

Pourahmadi, M. (2013). *High-dimensional covariance estimation*. John Wiley & Sons.

Rudelson, M., & Vershynin, R. (2013). Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 1–9.

Shen, H., & Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99(6), 1015–1034.

Taniguchi, M., & Kakizawa, Y. (2000). *Asymptotic theory of statistical inference for time series*. Springer-Verlag.

Taniguchi, M., & Krishnaiah, P. (1987). Asymptotic distributions of functions of the eigenvalues of sample covariance matrix and canonical correlation matrix in multivariate time series. *Journal of Multivariate Analysis*, 22(1), 156–176.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodology)*, 58(1), 267–288.

Tikhomirov, A. N. (1981). On the convergence rate in the central limit theorem for weakly dependent random variables. *Theory of Probability and its Applications*, 25(4), 790–809.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4), 389–434.

van de Geer, S. (2014). Weakly decomposable regularization penalties and structured sparsity. *Scandinavian Journal of Statistics*, 41(1), 72–86.

van de Geer, S. A. (2016). *Estimation and testing under sparsity*. Springer.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press.

Vu, V. Q., & Lei, J. (2013). Minimax sparse principal subspace estimation in high dimensions. *Annals of Statistics*, 41(6), 2905–2947.

Wang, T., Berthet, Q., & Samworth, R. J. (2016). Statistical and computational trade-offs in estimation of sparse principal components. *Annals of Statistics*, 44(5), 1896–1930.

Wong, K. C., Li, Z., & Tewari, A. (2020). Lasso guarantees for $\beta$-mixing heavy-tailed time series. *Annals of Statistics*, 48(2), 1124–1142.

Yao, J. (2012). A note on a Marčenko–Pastur type theorem for time series. *Statistics & Probability Letters*, 82(1), 22–28.

Zhao, L., Krishnaiah, P., & Bai, Z. (1986). On detection of the number of signals in presence of white noise. *Journal of Multivariate Analysis*, 20(1), 1–25.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B (Methodology)*, 67(2), 301–320.

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2), 265–286.

---

**How to cite this article:** Fujimori, K., Goto, Y., Liu, Y., & Taniguchi, M. (2023). Sparse principal component analysis for high-dimensional stationary time series. *Scandinavian Journal of Statistics*, 50(4), 1953–1983. [https://doi.org/10.1111/sjos.12664](https://doi.org/10.1111/sjos.12664)