Available online at www.sciencedirect.com
ScienceDirect
Stochastic Processes and their Applications xxx (xxxx) xxx
stochastic processes and their applications
www.elsevier.com/locate/spa

Extreme eigenvalues of nonlinear correlation matrices with applications to additive models

Zijian Guo[1], Cun-Hui Zhang[*,2]

Rutgers University, United States of America

Received 24 July 2020; accepted 9 April 2021
Available online xxxx

Dedicated to the memory of Larry Shepp

Abstract

The maximum correlation of functions of a pair of random variables is an important measure of stochastic dependence. It is known that this maximum nonlinear correlation is identical to the absolute value of the Pearson correlation for a pair of Gaussian random variables or a pair of finite sums of iid random variables. This paper extends these results to pairwise Gaussian vectors and processes, nested sums of iid random variables, and permutation symmetric functions of sub-groups of iid random variables. It also discusses applications to additive regression models.

Keywords: Nonlinear correlation; Extreme eigenvalue; Gaussian copula; Restricted eigenvalue; Compatibility condition; Additive model

1. Introduction

The maximum correlation of functions of a pair of random variables is an important measure of stochastic dependence. Formally, given random variables X_1 and X_2, the maximum correlation is defined as

$$R(X_1, X_2) = \sup\Big\{\mathrm{Cov}\Big(f_1(X_1), f_2(X_2)\Big) : \mathrm{Var}\big(f_1(X_1)\big) = \mathrm{Var}\big(f_2(X_2)\big) = 1\Big\}, \tag{1}$$

* Corresponding author.
E-mail address: czhang@stat.rutgers.edu (C.-H. Zhang).
1 Research was partially supported by the NSF DMS-1811857, DMS-2015373 and NIH R01GM140463-01.
2 Research was partially supported by the NSF Grants DMS-1513378, DMS-1721495, IIS-1741390 and CCF-1934924.

https://doi.org/10.1016/j.spa.2021.04.006

Please cite this article as: Z. Guo and C.-H. Zhang, Extreme eigenvalues of nonlinear correlation matrices with applications to additive models, Stochastic Processes and their Applications (2021), https://doi.org/10.1016/j.spa.2021.04.006.

where f_1 and f_2 are real functions. If X_1 and X_2 are bivariate normal, it was established in [14] that

$$R(X_1, X_2) = |\rho(X_1, X_2)| \tag{2}$$

where $\rho(X_1, X_2)$ denotes the Pearson correlation between X_1 and X_2. Dembo, Kagan and Shepp (2001) showed that the equality (2) holds with $R(X_1, X_2) = \sqrt{m/n}$, $1 \le m \le n$, if X_1 and X_2 are respectively nested sums of m and n independent and identically distributed (iid) random variables with finite second moment. Following their work, Bryc et al. [4] removed the second moment condition for the nested sums, and Yu [21] extended the result to two sums of arbitrary finite subsets of iid random variables.

The current paper extends the above results to more than two random variables and Gaussian processes. Let $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest and largest eigenvalues of matrices or linear operators, and $\mathrm{Corr}_{\ne}(X_1, \dots, X_p)$ the $p \times p$ off-diagonal correlation matrix of p random variables with elements $\rho(X_j, X_k) I_{\{j \ne k\}}$. Since the maximum correlation of a pair of random variables can be expressed as $R(X_1, X_2) = \sup_{f_1, f_2} \lambda_{\max}\big(\mathrm{Corr}_{\ne}(f_1(X_1), f_2(X_2))\big)$, a natural extension of the maximum nonlinear correlation to the multivariate setting is the extreme eigenvalue of the off-diagonal correlation matrix of marginal function transformations of $X_1, \dots, X_p$,

$$\rho_{\max}^{NL}(X_1, \dots, X_p) = \sup_{f_1, \dots, f_p} \lambda_{\max}\left(\mathrm{Corr}_{\ne}\left(f_1(X_1), \dots, f_p(X_p)\right)\right), \tag{3}$$

where the supreme is taken over all deterministic f_j with $0 < \mathrm{Var}\big(f_j^2(X_j)\big) < \infty$, and similarly

$$\rho_{\min}^{NL}(X_1, \dots, X_p) = \inf_{f_1, \dots, f_p} \lambda_{\min}\left(\mathrm{Corr}_{\ne}\left(f_1(X_1), \dots, f_p(X_p)\right)\right). \tag{4}$$

For $p = 2$, $\rho_{\max}^{NL} = -\rho_{\min}^{NL} \in [0, 1]$. However, for $p \ge 3$, $\rho_{\min}^{NL} \in [-1, 0]$ is no longer determined by $\rho_{\max}^{NL} \in [0, p-1]$, so that both quantities are needed to capture the extreme eigenvalues of the off-diagonal nonlinear correlation matrix. Moreover, (3) and (4) lead to the following further extension to stochastic processes: For any process $X_{\mathcal{T}} = \{X_t, t \in \mathcal{T}\}$ on an index set $\mathcal{T}$ equipped with a measure ν and $W_{s,t} \ge 0$ as a weight function on $\mathcal{T} \times \mathcal{T}$,

$$\begin{aligned} \rho_{\max}^{NL} &= \rho_{\max}^{NL}(X_{\mathcal{T}}, \nu, W) \\ &= \sup_{f_{\mathcal{T}} \in \mathcal{F}_{\mathcal{T}}} \sup_{\|h\|_{L_2(\nu)}=1} \int_{t \in \mathcal{T}} \int_{s \in \mathcal{T}} \rho\left(f_s(X_s), f_t(X_t)\right) W_{s,t} h(s) h(t) \nu(ds) \nu(dt), \end{aligned} \tag{5}$$

where $\|h\|_{L_2(\nu)} = \left\{\int_{\mathcal{T}} h^2(t) \nu(dt)\right\}^{1/2}$ and $\mathcal{F}_{\mathcal{T}}$ is the class of all deterministic $f_{\mathcal{T}} = \{f_t, t \in \mathcal{T}\}$ satisfying proper measurability and integrability conditions. Correspondingly,

$$\begin{aligned} \rho_{\min}^{NL} &= \rho_{\min}^{NL}(X_{\mathcal{T}}, \nu, W) \\ &= \inf_{f_{\mathcal{T}} \in \mathcal{F}_{\mathcal{T}}} \inf_{\|h\|_{L_2(\nu)}=1} \int_{t \in \mathcal{T}} \int_{s \in \mathcal{T}} \rho\left(f_s(X_s), f_t(X_t)\right) W_{s,t} h(s) h(t) \nu(ds) \nu(dt). \end{aligned} \tag{6}$$

Clearly, (3) and (4) are respectively special cases of (5) and (6) with $\mathcal{T} = \{1, \dots, p\}$, $W_{s,t} = I_{\{s \ne t\}}$ and the counting measure $\nu(A) = |A|$. We refer to (5) and (6) as the maximum, minimum or extreme nonlinear correlations of the process $X_{\mathcal{T}}$. Let $K_{W, f_{\mathcal{T}}}(s, t) = \rho\big(f_s(X_s), f_t(X_t)\big) W_{s,t}$ as a kernel and $K_{W, f_{\mathcal{T}}} : h \to \int K_{W, f_{\mathcal{T}}}(\cdot, s) h(s) \nu(ds)$ as a linear operator in $L_2(\nu)$. The extreme nonlinear correlations in (5) and (6) are expressed as the extreme eigenvalues of the operator $K_{W, f_{\mathcal{T}}} : L_2(\nu) \to L_2(\nu)$ via

$$\rho_{\max}^{NL} = \sup_{f_{\mathcal{T}} \in \mathcal{F}_{\mathcal{T}}} \lambda_{\max}(K_{W, f_{\mathcal{T}}}), \quad \rho_{\min}^{NL} = \inf_{f_{\mathcal{T}} \in \mathcal{F}_{\mathcal{T}}} \lambda_{\min}(K_{W, f_{\mathcal{T}}}). \tag{7}$$

Because the weight function W is almost completely general, it can be used to absorb the Radon–Nikodym derivative between two choices of the measure ν as follows. The pair $\{\nu', W')$ would yield the same extreme nonlinear correlations as $\{\nu, W)$ when the measures ν and ν' are absolutely continuous with respect to each other and $W'_{s,t} = W_{s,t}\sqrt{\nu(ds)/\nu'(ds)}\sqrt{\nu(dt)/\nu'(dt)}$. For $\mathcal{T} = \{1, \ldots, p\}$, we may take ν as the counting measure without loss of generality, so that the quantities in (7) are given by the extreme eigenvalues of the matrix $(K_{W,f_{\mathcal{T}}}(j,k))_{p\times p}$.

The main assertion of this paper is that in a number of settings, the above weighted extreme nonlinear correlations are identical to their linear counterpart:

$$\rho^{NL}_{\max} = \rho^{L}_{\max} \quad \text{and} \quad \rho^{NL}_{\min} = \rho^{L}_{\min}, \tag{8}$$

where $\rho^L_{\max}$ and $\rho^L_{\min}$ are defined by restricting the functions f_t in (5) and (6) to be the identity $f(x) = x$; e.g. in the more general stochastic process setting,

$$\rho^{L}_{\max} = \rho^{L}_{\max}(X_{\mathcal{T}}, \nu, W) = \sup_{\|h\|_{L_2(\nu)}=1} \int_{t\in\mathcal{T}}\int_{s\in\mathcal{T}} \rho\left(X_s, X_t\right) W_{s,t} h(s)h(t)\nu(ds)\nu(dt), \tag{9}$$

and

$$\rho^{L}_{\min} = \rho^{L}_{\min}(X_{\mathcal{T}}, \nu, W) = \inf_{\|h\|_{L_2(\nu)}=1} \int_{t\in\mathcal{T}}\int_{s\in\mathcal{T}} \rho\left(X_s, X_t\right) W_{s,t} h(s)h(t)\nu(ds)\nu(dt). \tag{10}$$

We note that for $W_{s,t} = 1$, $\rho^{NL}_{\max} \le \nu(\mathcal{T})$ and $\rho^{NL}_{\min} \ge 0$, so that (8) is trivial when $\rho^L_{\max} = \nu(\mathcal{T})$ and $\rho^L_{\min} = 0$. In fact, the first identify of (8) is nontrivial when $\rho^L_{\max} < \nu(\mathcal{T})$ and the second identify of (8) is nontrivial when $\rho^L_{\min} > 0$. However, for general $W_{s,t}$, there is no explicit formula for such attainable extreme solutions when the maximum and minimum are also taken over all correlation operators $\rho(X_s, X_t)$. Similar to (7), we define

$$\rho^{L}_{\max} = \lambda_{\max}(K_W), \quad \rho^{L}_{\min} = \lambda_{\min}(K_W). \tag{11}$$

where $K_W : h \to \int K_W(\cdot, s)h(s)\nu(ds)$ is the linear operator in $L_2(\nu)$ with the kernel $K_W(s,t) = \mathbb{E}\big[X_s X_t\big] W_{s,t}$. As discussed below (7), for $\mathcal{T} = \{1, \ldots, p\}$ we may take ν as the counting measure without loss of generality, so that (11) is given by the extreme eigenvalues of the matrix $(K_W(j,k))_{p\times p}$.

We will begin by proving (8) for Gaussian processes $X_{\mathcal{T}}$ on an arbitrary index set $\mathcal{T}$. Our analysis bears some resemblance to that of Lancaster [14] through the use of the Hermite polynomial expansion, but the general functional nature of our problem requires additional elements involving the spectrum boundary of the Schur product of linear operators. In fact, we prove that only a pairwise bivariate Gaussian condition is required for (8) under proper measurability and integrability conditions.

1.1. Hidden pairwise Gaussian and additive models

We generalize the results in (8) from pairwise Gaussian vectors to more general random vectors and then present two implications to the analysis of additive models. We shall say that a random vector $X_{1:p} = (X_1, \ldots, X_p)$ is *hidden Gaussian* if $X_j = T_j(Z_j)$ for a Gaussian vector $Z_{1:p} = (Z_1, \ldots, Z_p)$ and some deterministic transformations T_j, $1 \le j \le p$; $X_{1:p}$ is *hidden pairwise Gaussian* if the Gaussian requirement on $Z_{1:p}$ is reduced to pairwise Gaussian. The identities in (8) for the pairwise Gaussian process are equivalent to

$$\rho^{L}_{\min}(Z_{1:p}, \nu, W) \le \rho^{NL}_{\min}(X_{1:p}, \nu, W), \quad \rho^{NL}_{\max}(X_{1:p}, \nu, W) \le \rho^{L}_{\max}(Z_{1:p}, \nu, W),$$

for all measures ν and weights $W_{s,t}$. That is to say, if the correlation structure of $X_{1:p}$ is generated from a pairwise Gaussian distribution through marginal transformations, then their extreme nonlinear correlations are controlled within the extreme linear correlations of the underlying pairwise Gaussian distribution. When $Z_{1:p}$ is jointly Gaussian and the transformations T_j are monotone, this is the Gaussian copula model widely used in financial risk assessment and other areas of applications.

Our interest in the extreme nonlinear correlations arises from our study of the additive regression model where the response variable Y can be written as

$$Y = \sum_{j=1}^{p} f_j(X_j) + \epsilon.$$

As an important nonlinear relaxation of the linear regression, this model effectively mitigates the curse of dimensionality in the more complex multiple nonparametric regression [3,5,10,20]. Let $\|f\|_{L_2^{(0)}(\mathbb{P})}$ denote the semi-norm given by $\|f\|^2_{L_2^{(0)}(\mathbb{P})} = \mathrm{Var}(f(X_{1:p}))$. Our result on the minimum eigenvalue of the nonlinear correlation matrix has two interesting implications in the analysis of additive models as follows. Firstly, the characterization of $\rho^{NL}_{\min}$ in the current paper can be used to verify the theoretical restricted eigenvalue and compatibility conditions required for the analysis of additive models. In particular, the theoretical restricted eigenvalue and compatibility conditions on the design are critical for establishing upper bounds on the prediction error $\|\sum_{j=1}^p \widehat{f}_j - \sum_{j=1}^p f_j\|^2_{L_2^{(0)}(\mathbb{P})}$ of regularized estimators $\widehat{f}$ in the additive model [13,16–19]. Secondly, when the minimum nonlinear correlation of $X_{1:p}$ is bounded away from zero, the squared loss for the estimation of individual f_j can be derived from the prediction error bound via

$$\sum_{j=1}^{p} \|\widehat{f}_j - f_j\|^2_{L_2^{(0)}(\mathbb{P})} \le \frac{1}{\rho^{NL}_{\min}} \left\| \sum_{j=1}^{p} \widehat{f}_j - \sum_{j=1}^{p} f_j \right\|^2_{L_2^{(0)}(\mathbb{P})}$$

where $\rho^{NL}_{\min}$ is defined in (6) with the counting measure $\nu(A) = |A|$ and uniform weight $W_{s,t} = 1$. See Section 3 for more detailed discussions.

1.2. Symmetric functions

In addition to the extension of Lancaster [14] to pairwise Gaussian processes and vectors, the current paper directly extends the results of Dembo, Kagan and Shepp (2001), Bryc et al. [4] and Yu [21] by establishing (8) for nested sums $\big(X_1, X_2, \ldots, X_p\big)$ of iid random variables Y_i, with $X_j = \sum_{i=1}^{m_j} Y_i$ for some positive integers $m_1 < \cdots < m_p$. Moreover, as a natural generalization of the nested sums, we consider groups of the iid variables as random vectors $\boldsymbol{X}_j = (Y_i, i \in G_j)$ where G_j are sets of positive integers. We extend the first part of (8) by proving that for the counting measure ν and any weights $W_{j,k} \ge 0$

$$\max_{\text{symmetric } f_1,\ldots,f_p} \rho^L_{\max}\big(\big(f_1(\boldsymbol{X}_1), \ldots, f_p(\boldsymbol{X}_p)\big), \nu, W\big) = \rho^L_{\max}\big(\big(S_{G_1}, \ldots, S_{G_p}\big), \nu, W\big) \tag{12}$$

where $S_{G_j} = \sum_{i \in G_j} h_0(Y_i)$ for any deterministic function h_0 satisfying $0 < \mathrm{Var}(h_0(Y_i)) < \infty$ and the maximum is taken over all deterministic functions f_i symmetric in the permutation of its arguments. In the sequel, such f_i are simply called symmetric functions. We also establish the corresponding identity for the minimum correlation,

$$\min_{\text{symmetric } f_1,\ldots,f_p} \rho^L_{\min}\big(\big(f_1(\boldsymbol{X}_1), \ldots, f_p(\boldsymbol{X}_p)\big), \nu, W\big) = \rho^L_{\min}\big(\big(S_{G_1}, \ldots, S_{G_p}\big), \nu, W\big), \tag{13}$$

under a mild condition which holds when $\cap_{j=1}^p G_j \ne \emptyset$.

1.3. Paper organization

The rest of the paper is organized as follows. In Section 2, we study the extreme eigenvalues of nonlinear correlation matrix for pairwise Gaussian random vectors and processes; In Section 3, we discuss the implications of our results in Section 2 on additive models; In Section 4, we study the extreme eigenvalues of nonlinear correlation matrix of nested sums and also the more general symmetric functions of iid random variables.

2. Pairwise Gaussian processes

To start with, we shall explicitly specify the measurability and integrability conditions for the definition of the extreme linear and nonlinear correlations in (9), (10), (5) and (6).

Assumption A. (i) There exists $B_n \subset B_{n+1} \subset \mathcal{T}$ such that $\cup_{n=1}^{\infty} B_n = \mathcal{T}$, $\nu(B_n) < \infty$ and $\int_{B_n}\int_{B_n} W_{s,t}^2 \nu(ds)\nu(dt) < \infty$ for any positive integer $n \geq 1$.
(ii) The process $X_{\mathcal{T}}$ is standardized to $\mathbb{E}[X_t] = 0$ and $\mathbb{E}[X_t^2] = 1$, the correlation operator $\mathbb{E}\big[X_s X_t\big]$ is measurable as a function of (s,t) in the product space $\mathcal{T} \times \mathcal{T}$, and the weight function $W_{s,t}$ is element-wise nonnegative and symmetric, $W_{s,t} = W_{t,s} \geq 0$.
(iii) The operator K_W in (11) is bounded.

We note that there is no loss of generality to assume that $X_{\mathcal{T}}$ is standardized as (9) and (10) involve only the correlation between X_s and X_t. Under Assumption A (iii), the operator K_W yields finite extreme linear correlations in (9) and (10).

Assumption B. In (5) and (6), $\mathcal{F}_{\mathcal{T}}$ is the class of all function families $f_{\mathcal{T}} = \{f_t, t \in \mathcal{T}\}$ with $\mathbb{E}[f_t(X_t)] = 0$, $\mathbb{E}[f_t^2(X_t)] > 0$ and $\int_{\mathcal{T}} \mathbb{E}[f_t^2(X_t)]\nu(dt) < \infty$ such that $\mathbb{E}\big[X_t^m f_t(X_t)\big]$ are measurable functions of t on $\mathcal{T}$ for all integers $m \geq 1$, and in (7) the kernel $K_{W,f_{\mathcal{T}}}(s,t) = \mathrm{Corr}(f_s(X_s), f_t(X_t))W_{s,t}$ is a measurable function of (s,t) on $\mathcal{T} \times \mathcal{T}$.

In the discrete case where $\mathcal{T} = \{1, \ldots, p\}$, Assumption A always holds when $\mathbb{E}[X_t] = 0$ and $\mathbb{E}[X_t^2] = 1$ and Assumption B always holds when $\mathcal{F}_{\mathcal{T}}$ is the set of all $f_{\mathcal{T}} = \{f_1, \ldots, f_p\}$ satisfying $\mathbb{E}[f_j(X_j)] = 0$ and $0 < \mathbb{E}[f_j^2(X_j)] < \infty$, $j = 1, \ldots, p$.

We first establish some equivalent expressions to (5) and (6) in the following lemma.

Lemma 1. *Let $\rho_{\max}^{NL}$ and $\rho_{\min}^{NL}$ be as in (5) and (6) with the function class $\mathcal{F}_{\mathcal{T}}$ specified in Assumption B. Then,*

$$\rho_{\max}^{NL} = \sup_{f_{\mathcal{T}} \in \mathcal{F}_{\mathcal{T}}} \frac{\int_{t\in\mathcal{T}}\int_{s\in\mathcal{T}} \mathbb{E}\big[f_s(X_s), f_t(X_t)\big] W_{s,t}\nu(ds)\nu(dt)}{\int_{t\in\mathcal{T}} \mathbb{E}\big[f_t^2(X_t)\big]\nu(dt)}, \tag{14}$$

and

$$\rho_{\min}^{NL} = \inf_{f_{\mathcal{T}} \in \mathcal{F}_{\mathcal{T}}} \frac{\int_{t\in\mathcal{T}}\int_{s\in\mathcal{T}} \mathbb{E}\big[f_s(X_s), f_t(X_t)\big] W_{s,t}\nu(ds)\nu(dt)}{\int_{t\in\mathcal{T}} \mathbb{E}\big[f_t^2(X_t)\big]\nu(dt)}. \tag{15}$$

A proof of Lemma 1 can be found in the Appendix. The more explicit expressions established in the lemma will facilitate the Hermite polynomial expansion of the covariance in our analysis. Another ingredient of our analysis, stated in the following lemma, concerns the extreme eigenvalues of the Schur product.

Lemma 2. *Let $\rho^L_{\max}$ and $\rho^L_{\min}$ be as in (9) and (10) respectively and $K_W(s,t) = \mathbb{E}[X_s X_t]W_{s,t}$. Under Assumption A,*

$$\rho^L_{\min} \le \int_{t\in\mathcal{T}}\int_{s\in\mathcal{T}} \big(\mathbb{E}[X_s X_t]\big)^{m-1} K_W(s,t)h(s)h(t)\nu(ds)\nu(dt) \le \rho^L_{\max}. \tag{16}$$

for any integer $m \ge 1$ and function $h(t)$ with $\int h^2(t)\nu(dt) = 1$.

The above lemma establishes that the spectrum of the operator given by the Schur power kernel $\big(\mathbb{E}[X_s X_t]\big)^{m-1} K_W(s,t) = \big(\mathbb{E}[X_s X_t]\big)^{m} W_{s,t}$ is controlled inside that of $K_W(s,t)$, so that the Schur multiplication of a correlation matrix is a contraction. The proof of the lemma, given in the Appendix, utilizes an interesting construction of the Schur power kernel with iid copies of $X_{\mathcal{T}}$. Such a proof technique is simple but quite useful.

We are now ready to state the equivalence between the extreme nonlinear correlation and the extreme linear correlation for pairwise Gaussian processes.

Theorem 1. *Let $X_{\mathcal{T}} = \{X_t\}_{t\in\mathcal{T}}$ be a pairwise Gaussian process in the sense that (X_s, X_t) are bivariate Gaussian vectors for all pairs $(s,t) \in \mathcal{T}\times\mathcal{T}$. Under Assumptions A and B,*

$$\rho^{NL}_{\max} = \rho^L_{\max} \quad \text{and} \quad \rho^{NL}_{\min} = \rho^L_{\min},$$

where $\rho^{NL}_{\max}$ and $\rho^{NL}_{\min}$ are defined in (5) and (6) respectively, and $\rho^L_{\max}$ and $\rho^L_{\min}$ are defined in (9) and (10) respectively.

Proof. As the normalized Hermite polynomials

$$H_m(x) = (m!)^{-1/2}(-1)^m e^{x^2/2}(d/dx)^m e^{-x^2/2}$$

form a orthonormal system with $\mathbb{E}[H_m(Z)] = 0$ and $\mathbb{E}[H_m^2(Z)] = 1$ for $Z \sim N(0,1)$, by Assumptions A and B we may write $f_t(X_t) = \sum_{m=1}^\infty a_m(t)H_m(X_t)$ in the sense of L_2 convergence. Let $K_W^m(s,t) = \big(\mathbb{E}\big[X_s, X_t\big]\big)^{m-1}K_W(s,t) = \big(\mathbb{E}\big[X_s, X_t\big]\big)^{m}W_{s,t}$. As (X_s, X_t) is bivariate normal with $\mathrm{Var}(X_s) = \mathrm{Var}(X_t) = 1$, $\mathbb{E}[H_m(X_s)H_n(X_t)]W_{s,t} = K_W^m(s,t)$ as in [14]. It follows that $\mathbb{E}\big[f_s(X_s)f_t(X_t)\big]W_{s,t} = \sum_{m=1}^\infty K_W^m(s,t)a_m(s)a_m(t)$. As $|K_W^m(s,t)| \le K_W^2(s,t)$, Lemma 2 provides

$$\begin{aligned}
&\int_{s\in\mathcal{T}}\int_{t\in\mathcal{T}} \mathbb{E}\big[f_s(X_s),\, f_t(X_t)\big]W_{s,t}\nu(ds)\nu(dt)\\
&= \int_{s\in\mathcal{T}}\int_{t\in\mathcal{T}} \Big\{\sum_{m=1}^\infty K_W^m(s,t)a_m(s)a_m(t)\Big\}\nu(ds)\nu(dt)\\
&\le \int_{s\in\mathcal{T}}\int_{t\in\mathcal{T}} K_W(s,t)a_1(s)a_1(t)\nu(ds)\nu(dt)\\
&\quad + \sum_{m=2}^\infty \int_{s\in\mathcal{T}}\int_{t\in\mathcal{T}} K_W^2(s,t)\big|a_m(s)a_m(t)\big|\nu(ds)\nu(dt)\\
&\le \rho^L_{\max}\sum_{m=1}^\infty \int a_m^2(t)\nu(dt)\\
&= \rho^L_{\max}\int_{t\in\mathcal{T}} \mathbb{E}\big[f_t^2(X_t)\big]\nu(dt).
\end{aligned}$$

Moreover, as the exchange of summation and integration is allowed as the above,

$$\begin{aligned}
&\int_{s\in\mathcal{T}}\int_{t\in\mathcal{T}} \mathbb{E}\big[f_s(X_s),\, f_t(X_t)\big] W_{s,t}\,\nu(ds)\nu(dt)\\
&= \sum_{m=1}^{\infty}\int_{s\in\mathcal{T}}\int_{t\in\mathcal{T}} \big\{K_W^m(s,t)a_m(s)a_m(t)\big\}\nu(ds)\nu(dt)\\
&\geq \rho_{\min}^L \sum_{m=1}^{\infty}\int a_m^2(t)\nu(dt)\\
&= \rho_{\min}^L \int_{t\in\mathcal{T}} \mathbb{E}\big[f_t^2(X_t)\big]\nu(dt).
\end{aligned}$$

The proof is complete as inequalities in the other direction are trivial. □

Theorem 1 establishes the equality of the extreme eigenvalues of the nonlinear and linear correlation operators. However, as we have mentioned in the introduction, such results could be trivial when $\rho^L_{\max}$ and $\rho^L_{\min}$ attain the extreme eigenvalues among all correlation operators. In the following three subsections, we discuss the discrete case $\mathcal{T} = \{1, \dots, p\}$, the continuous case $\mathcal{T} = [0, 1]$, and stationary processes as three nontrivial examples and state the implications of Theorem 1 as corollaries.

2.1. Hidden pairwise Gaussian vectors

The following part demonstrates the application of Theorem 1 to a finite number of pairwise Gaussian random variables, that is, $\mathcal{T} = \{1, 2, \dots, p\}$. As discussed below (7) and (11), we take ν as the counting measure without loss of generality throughout the subsection.

Corollary 1. *Let $X_1, X_2, \dots, X_p$ be pairwise Gaussian random variables with $X_j \sim N(0, 1)$ and a correlation matrix $\Sigma = (\Sigma_{j,k})_{p\times p}$. Let $W = (W_{j,k})_{p\times p}$ be a matrix with elements $W_{j,k} = W_{k,j} \geq 0$ and $\Sigma \circ W = (\Sigma_{j,k}W_{j,k})_{p\times p}$ be the Schur product. Then, for all functions f_j satisfying $\mathbb{E} f_j(X_j) = 0$ and $0 < \mathbb{E} f^2(X_j) < \infty$,*

$$\lambda_{\min}(\Sigma\circ W) \leq \frac{\mathbb{E}\left[\sum_{j=1}^p\sum_{k=1}^p W_{j,k} f_j(X_j) f_j(X_k)\right]}{\sum_{j=1}^p \mathbb{E} f_j^2(X_j)} \leq \lambda_{\max}(\Sigma\circ W). \tag{17}$$

In particular, for $\Sigma\circ W = \Sigma$ with $W_{j,k} = 1$,

$$\lambda_{\min}(\Sigma)\cdot\sum_{j=1}^p \mathbb{E} f_j^2(X_j) \leq \mathbb{E}\left(\sum_{j=1}^p f_j(X_j)\right)^2 \leq \lambda_{\max}(\Sigma)\cdot\sum_{j=1}^p \mathbb{E} f_j^2(X_j). \tag{18}$$

Equivalently, for $W_{j,k} = I_{\{j\neq k\}}$, (3) and (4) are given by their linear version,

$$\rho^{NL}_{\max}(X_1,\dots,X_p) = \lambda_{\max}(\Sigma) - 1 \quad \text{and} \quad \rho^{NL}_{\min}(X_1,\dots,X_p) = \lambda_{\min}(\Sigma) - 1. \tag{19}$$

In the setting of the above corollary, the operator K_W in (11) is given by the Schur product matrix $K_W = \Sigma\circ W$, and for general weights W (18) and (19) are nontrivial with $\lambda_{\min}(\Sigma) > 0$ and $\lambda_{\max}(\Sigma) < p$ when Σ is of full rank.

Finally, we state in the following corollary the implication of Theorem 1 on Gaussian copula and other hidden pairwise Gaussian variables.

Corollary 2. *Suppose $X_{1:p} = \big(X_1, X_2, \dots, X_p\big)$ follows a hidden Gaussian distribution in the sense of $X_j = T_j(Z_j)$ for a Gaussian vector $Z_{1:p} = (Z_1, \dots, Z_p)$ and some deterministic functions T_j with $0 < Var(T_j(Z_j)) < \infty$. Let Σ^z be the covariance matrix of the hidden vector $Z_{1:p}$. Then, for the counting measure ν and any symmetric W with $W_{j,k} \geq 0$,*

$$\lambda_{\min}(\Sigma^z \circ W) \leq \rho^{NL}_{\min}(X_{1:p}, \nu, W), \quad \rho^{NL}_{\max}(X_{1:p}, \nu, W) \leq \lambda_{\max}(\Sigma^z \circ W),$$

and the above inequalities become equality when T_j are almost surely invertible. In particular, (18) holds with Σ replaced by Σ^z. Moreover, the Gaussian assumption on $Z_{1:p}$ can be weakened to pairwise Gaussian.

Similarly to Corollary 1, the upper and lower bounds in the above corollary are nontrivial when the covariance matrix of $Z_{1:p}$ is of full rank. The above corollary has interesting implications as it states that the extreme eigenvalues of nonlinear correlation matrix fall into the spectrum range of the covariance matrix of the underlying generating Gaussian distribution. This is meaningful in statistical applications, that is, the well conditioning of the covariance matrix of general nonlinear transformations follows from that of the underlying generating Gaussian covariance matrix.

2.2. Processes on finite intervals

Our result for a general pairwise Gaussian process with general index set also directly leads to the same for Gaussian process on finite intervals. As discussed below (7), we take $\mathcal{T} = [0, 1]$ and the Lebesgue measure $\nu(dt) = dt$ without much loss of generality.

Corollary 3. *Let $\{X_t, 0 \leq t \leq 1\}$ be a Gaussian process with correlation $\rho(X_s, X_t)$ and $W_{s,t}$ be a nonnegative symmetric square integrable function of (s, t) in $[0, 1]^2$. Let $K_W(s, t) = \rho(X_s, X_t)W_{s,t}$. Let K_W be the linear operator $h(\cdot) \to \int_0^1 K_W(\cdot, s)h(s)ds$. Then,*

$$\rho^{NL}_{\max} = \lambda_{\max}\left(K_W\right), \quad \rho^{NL}_{\min} = \lambda_{\min}\left(K_W\right),$$

for the extreme nonlinear correlations in (5) and (6), where $\lambda_{\max}(K_W)$ and $\lambda_{\min}(K_W)$ are the extreme eigenvalues of the operator K_W in $L_2([0, 1])$. In particular, for K_1 with $W = 1$,

$$Var\left(\int_0^1 f(X_t, t)dt\right) \leq \lambda_{\max}\left(K_1\right) \int_0^1 Var\big(f(X_t, t)\big)dt$$

for all bounded continuous bivariate functions $f(x, t)$.

In the above corollary, $\lambda^2_{\max}(K_1) \leq \int_0^1\int_0^1 \rho^2(X_s, X_t)dsdt \leq 1$, so that the maximum eigenvalue of the correlation operator K_1 is nontrivial unless $\rho^2(X_s, X_t) \equiv 1$. As the operator K_1 is Hilbert–Schmidt, the minimum eigenvalue $\lambda_{\min}(K_1) = 0$ is always trivial. However, $\lambda_{\min}(K_W)$ may take nontrivial negative values for general W.

The setting here is related to the nested sum problem considered in Section 4 as follows. Let $S_j = \sum_{i=1}^j Y_i$, $1 \leq j \leq p$, where Y_i are iid random variables with $\mathbb{E}[Y_i] = 0$ and $\mathbb{E}[Y_i^2] = 1$. As the correlation kernel of $\{X_t = S_{\lfloor pt \rfloor}/\sqrt{p}, 0 \leq t \leq 1\}$ uniformly converges to the correlation kernel $K_1(s, t) = (s \wedge t)/\sqrt{st}$ of the standard Brownian motion as $p \to \infty$,

$$\lambda^L_{\max}(S_{1:p}, \nu^{(p)}, W^{(p)}) \to \lambda_{\max}(K_W), \quad \lambda^L_{\min}(S_{1:p}, \nu^{(p)}, W^{(p)}) \to \lambda_{\min}(K_W),$$

where $\nu^{(p)}(A) = |A|/p$ is the normalized counting measure and $W^{(p)}_{j,k} = W_{j/p,k/p}$ for $A \cup \{j, k\} \subseteq \{1, \dots, p\}$, and K_W is treated as the operator in $L_2([0, 1])$ as in Corollary 3.

2.3. Stationary processes

In this subsection we consider stationary processes on the entire set of integers and the entire real line. In both cases we consider stationary pairwise Gaussian X_t with $\mathbb{E}[X_t] = 0$, $\mathbb{E}[X_t^2] = 1$ and autocorrelation function

$$\rho(s, t) = \rho(t - s) = \mathbb{E}[X_s X_t].$$

We consider weight functions $W_{s,t} = W_{t,s} = W(t - s) \geq 0$ and write the kernel as

$$K_W(s, t) = K_W(s - t), \quad K_W(t) = \rho(t)W(t).$$

We shall first consider the discrete case.

Corollary 4. *Let $\{X_t, t \in \mathcal{T}\}$, $\mathcal{T} = \{0, \pm 1, \pm 2, \ldots\}$, be a pairwise Gaussian stationary sequence with autocorrelation $\rho(t) = \mathbb{E}[X_t X_0]$. Let ν be the counting measure on $\mathcal{T}$, $W_{s,t} = W_{t,s} = W(t - s) \geq 0$, $K_W(t) = \rho(t)W(t)$, and $K_W : h \to \sum_{s\in\mathcal{T}} K_W(\cdot - s)h(s)$ as an operator in ℓ_2. Suppose $\sum_{t\in\mathcal{T}} |K_W(t)| < \infty$. Then, (8) and (11) hold with*

$$\rho_{\max}^{NL} = \lambda_{\max}(K_W) = \sup_{|\omega|\leq\pi} \left|K_W^*(\omega)\right|, \quad \rho_{\min}^{NL} = \lambda_{\min}(K_W) = \inf_{|\omega|\leq\pi} \left|K_W^*(\omega)\right|,$$

where $K_W^*(\omega) = \sum_{s\in\mathcal{T}} K_W(s)\cos(\omega s)$. In particular, for the autoregression sequence with $\rho(t) = \beta^{|t|}$, $|\beta| < 1$ necessarily, we have $K_1^*(\omega) = (1 - \beta^2)/(1 + \beta^2 - 2\beta\cos(\omega))$, $\lambda_{\max}(K_1) = (1 + |\beta|)/(1 - |\beta|)$ and $\lambda_{\min}(K_1) = (1 - |\beta|)/(1 + |\beta|)$ for $W(t) = 1$.

We note that Corollary 4 gives the autoregression and implicitly many other examples in which $\nu(\mathcal{T}) = \infty$ and $\lambda_{\max}(K_W) < \infty$ and $\lambda_{\min}(K_W) > 0$ are both nontrivial. For a pairwise Gaussian stationary process $\{X_t, t \in \mathcal{T}\}$, the process $\{f_t(X_t), t \in \mathcal{T}\}$ is in general non-Gaussian and non-stationary, and its spectrum is typically not tractable. Still, Corollary 4 shows that the spectrum of the nonlinear $\{f_t(X_t), t \in \mathcal{T}\}$ is contained within the spectrum of the underlying process $\{X_t, t \in \mathcal{T}\}$ under mild conditions.

Proof. Let $F : h \to Fh$ be the mapping from complex $h \in L_2([-\pi, \pi])$ to its Fourier series $(Fh)(t) = (2\pi)^{-1/2} \int_{-\pi}^{\pi} e^{i\omega t} h(\omega)d\omega$, $t \in \mathcal{T}$. For h with finitely many nonzero coefficients,

$$(K_W Fh)(t) = \sum_{s\in\mathcal{T}} K_W(|t - s|) \int_{-\pi}^{\pi} \frac{e^{i\omega s}}{(2\pi)^{1/2}} h(\omega)d\omega = \int_{-\pi}^{\pi} \frac{e^{i\omega t}}{(2\pi)^{1/2}} K_W^*(\omega)h(\omega)d\omega.$$

Let K_W^* be the mapping $h(\omega) \to K_W^*(\omega)h(\omega)$ in $L_2([-\pi, \pi])$. As h with finitely many nonzero Fourier coefficients are dense in $L_2([-\pi, \pi])$ and F is isometric from $L_2([-\pi, \pi])$ onto ℓ_2, the above calculation implies $K_W F = F K_W^*$. Moreover, the spectrum decomposition $K_W = \int \lambda dP_\lambda$ is given by projections $P_\lambda = F P_\lambda^* F^{-1}$ where $P_\lambda^* h(\omega) = h(\omega)I\{K_W^*(\omega) \leq \lambda\}$ gives the spectrum decomposition $K_W^* = \int \lambda dP_\lambda^*$. This gives the main conclusion. For $K_1(t) = \beta^{|t|}$, $K_1^*(\omega) = \sum_{s\in\mathcal{T}} \beta^{|s|} e^{i\omega s} = (1 - \beta^2)/(1 + \beta^2 - 2\beta\cos(\omega))$ gives the spectrum. $\square$

Next, we consider the continuous case. The spectrum of the Ornstein–Uhlenbeck process $K_1(t) = e^{-|t|}$ was studied in [15] by directly solving the eigenvalue problem for the restriction of the kernel $K_1(s, t) = e^{-|t-s|}$ on $L_2([a, b])$ in the proof of Theorem 5 there.

Corollary 5. *Let $\{X_t, t \in \mathbb{R}\}$ be a stationary pairwise Gaussian process on the entire real line with autocorrelation $\rho(t) = \mathbb{E}[X_t X_0]$. Let ν be the Lebesgue measure on $\mathbb{R}$, $W_{s,t} = W_{t,s} =$*

$W(t-s) \geq 0$, $K_W(t) = \rho(t)W(t)$, and $K_W : h \to \int_{-\infty}^{\infty} K_W(\cdot - s)h(s)ds$ as an operator in $L_2(\mathbb{R})$. Suppose $\int_{-\infty}^{\infty} |K_W(t)|dt < \infty$. Then, (8) and (11) hold with

$$\rho_{\max}^{NL} = \lambda_{\max}(K_W) = \sup_{\omega} \left|K_W^*(\omega)\right|, \quad \rho_{\min}^{NL} = \lambda_{\min}(K_W) = \inf_{\omega} \left|K_W^*(\omega)\right|,$$

where $K_W^*(\omega) = \int_{-\infty}^{\infty} K_W(s)\cos(\omega s)ds$. In particular, for the Ornstein–Uhlenbeck process with $\rho(t) = e^{-|t|}$, $K_1^*(\omega) = 2/(1+\omega^2)$, $\lambda_{\max}(K_1) = 2$ and $\lambda_{\min}(K_1) = 0$ for $W(t) = 1$.

Proof. Here F is the Fourier transformation $(Fh)(t) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} e^{i\omega t} h(\omega)d\omega$, and

$$(K_W F h)(t) = \int_{-\infty}^{\infty} K_W(|t-s|) \int_{-\infty}^{\infty} \frac{e^{i\omega s}}{(2\pi)^{1/2}} h(\omega)d\omega ds = \int_{-\infty}^{\infty} \frac{e^{i\omega t}}{(2\pi)^{1/2}} K_W^*(\omega)h(\omega)d\omega.$$

This certainly holds for $h \in L_2(\mathbb{R}) \cap L_1(\mathbb{R})$ which is dense in $L_2(\mathbb{R})$. Again, as F is isometric in $L_2(\mathbb{R})$, $K_W F = F K_W^*$ with the operator $K_W^* : h(\omega) \to K_W^*(\omega)h(\omega)$. This gives the spectrum decomposition and spectrum limits of K_W as in the proof of Corollary 4. For $K_1(t) = e^{-|t|}$, $K_1^*(\omega) = \int e^{-|s|}e^{i\omega s}ds = 1/(1-i\omega) + 1/(1+i\omega)$ gives the spectrum. □

Our problem is also related to mixing conditions on stochastic processes. For example, when $\{X_t, -\infty < t < \infty\}$ is a Gauss–Markov process, its ρ-mixing coefficient, given by

$$\rho^*(n) = \sup_t \sup_{f \in \mathscr{F}_{(-\infty,t]}, g \in \mathscr{F}_{[t+n,\infty)}} \mathrm{Corr}(f, g)$$

with $\mathscr{F}_A$ being the set of all nonzero square integrable functions of $\{X_t, t \in A\}$, can be characterized by $\rho^*(n) = \sup_t \mathrm{Corr}(X_t, X_{t+n})$ by (2). However, as this paper is mainly motivated by the application in the additive model as discussed in Section 3 and the multivariate extension of Dembo, Kagan and Shepp (2001) as discussed in Section 4, our results do not yield a direct extension of the above explicit calculation of the ρ-mixing condition to more general processes. We refer to Bradley et al. [2] for a survey of the relationship among different mixing conditions.

3. Applications to additive models

In this section, we discuss applications of our results to additive models, including justification of theoretical restricted eigenvalue and compatibility conditions and derivation of convergence rates for the estimation of individual component functions from prediction error bounds. In the additive regression model, the relationship between the response variable Y and design vector $X_{1:p} = (X_1, \ldots, X_p)$ is given by $Y = \sum_{j=1}^{p} f_j(X_j) + \varepsilon$, or $\mathbb{E}[Y|X_{1:p}] = \sum_{j=1}^{p} f_j(X_j)$ in terms of the conditional expectation, where f_j are assumed to be smooth functions and ε is the noise variable independent of $X_{1:p}$ with $\mathbb{E}[\varepsilon] = 0$.

Additive models have been important tools for practical data analysis [5,8], mainly due to the fact that it relaxes the stringent model assumption in linear regression and at the same time mitigates the curse of dimensionality in multiple nonparametric regression with $Y = f(X_{1:p}) + \varepsilon$. Another advantage of the additive model is the natural interpretation of its components. For example, the rate of change of the jth function f_j represents the effect of the covariate X_j as in linear regression.

Due to its importance, additive models have been extensively investigated in both the classical low-dimension setting and the more contemporary high-dimensional setting where only a much smaller number than p of the components f_j is actually nonzero. In both cases, one of the main assumptions is the invertibility condition of the additive model. In the very special linear regression setting $Y = \sum_{j=1}^{p} X_j\beta_j + \varepsilon$, this invertibility condition is that the minimum eigenvalue of the sample covariance matrix of the design vector $X_{1:p}$ is bounded away from zero. In the more general additive model, the component functions f_j are not necessarily linear and the invertibility condition is typically imposed on the sample covariance matrix of certain basis functions of f_j. In the following, we connect the invertibility condition to $\rho_{\min}^{NL}$ in (10) and verify them over a large class of distributions, in both the low- and high-dimensional settings.

3.1. Implications to low-dimensional additive models

We start with the low-dimensional setting where the number of covariates p is fixed or much smaller than the diverging sample size n. A useful way of understanding and implementing the additive models is to consider the projection of the response to the linear span of suitable bases of the component functions f_j. Denote a set of basis functions for f_j by $B_j = B_j(x_j) = (B_{j,1}(x_j)\cdots B_{j,M_j}(x_j))^{\intercal} \in \mathbb{R}^{M_j}$ with some positive integer M_j, where the basis can be taken as Fourier, spline, wavelet or other constructions and M_j is allowed to grow as the sample size increases. Under proper smoothness conditions, $f_j(x_j)$ can be adequately approximated by a linear combination $a_j^{\intercal} B_j(x_j)$ of its basis functions, resulting in a d^*-dimensional regression $E[Y|X_{1:p}] \approx \sum_{j=1}^{p} a_j^{\intercal} B_j(X_j)$ with very large $d^* = \sum_{j=1}^{p} M_j$.

When iid copies $\{(X_{i,1},\ldots,X_{i,p},Y_i)\}_{1\le i\le n}$ of $(X_1,\ldots,X_p,Y)$ are observed, the invertibility condition in this d^*-dimensional regression can be written as

$$\mathbb{P}\left\{\min_{\sum_{j=1}^{p}\|\alpha_j\|_2^2=1}\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}a_j^{\intercal}B_j(X_{i,j})-\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}a_j^{\intercal}B_j(X_{i,j})\right)\right)^2\ge\kappa_0\right\}\to 1 \quad (20)$$

with some fixed positive constant κ_0. While probabilistic methods such as empirical process and noncommutative Bernstein inequality can be used to verify the above condition, such analysis invariably requires the following population invertibility condition:

$$\min_{\sum_{j=1}^{p}\|\alpha_j\|_2^2=1}\left\|\sum_{j=1}^{p}\alpha_j^{\intercal}B_j(X_j)\right\|_{L_2^{(0)}(\mathbb{P})}^2\ge\kappa_0,$$

as $\|f(X_{1:p})\|_{L_2^{(0)}(\mathbb{P})}^2 = \mathrm{Var}(f(X_{1:p}))$ for all functions $f:\mathbb{R}^p\to\mathbb{R}$. This population invertibility condition can be decomposed into a component-wise invertibility condition

$$\left\|\alpha_j^{\intercal}B_j(X_j)\right\|_{L_2^{(0)}(\mathbb{P})}^2\ge\kappa_1\|\alpha_j\|_2^2,\ \text{ for } j=1,\ldots,p \text{ and } \kappa_1>0,$$

and a population predictive invertibility condition

$$\left\|\sum_{j=1}^{p}\alpha_j^{\intercal}B_j(X_j)\right\|_{L_2^{(0)}(\mathbb{P})}^2\ge\kappa_2\sum_{j=1}^{p}\left\|\alpha_j^{\intercal}B_j(X_j)\right\|_{L_2^{(0)}(\mathbb{P})}^2 \quad \text{for } \kappa_2>0. \quad (21)$$

Assume that after proper centering and scaling the support of $X_{1:p}$ is $[0,1]^p$. The component-wise invertibility condition is fulfilled when B_j is orthonormal in $L_2([0,1])$ and the marginal density of X_j is uniformly greater than κ_1 in $[0,1]$. The orthonormal condition on B_j can be further weakened to $\left\|\alpha_j^{\intercal}B_j(X_j)\right\|_{L_2^{(0)}([0,1])}^2 \gtrsim \|\alpha_j\|_2^2$ as in the case of B-spline. It is also well known that the population predictive invertibility condition (21) holds when the joint density of $X_{1:p}$ is uniformly greater than κ_0 in $[0,1]^p$. However, while the lower bound assumption on the individual marginal densities approximately holds after the quantile transformation of individual samples $(X_{1,j},\ldots,X_{n,j})$, the lower bound assumption on the joint density is much harder to ascertain. Consequently, it is unclear from the existing literature the extent of the validity of the largely theoretical assumption (21) beyond the restrictive condition on the lower bound of the joint density.

Our results provide the validity of (21) in a broad collection of new scenarios as follows. When $X_{1:p} = \big(T_1(Z_1),\ldots,T_p(Z_p)\big)$ follows a hidden pairwise Gaussian distribution, Corollary 2 provides (21) with $\kappa_2 = \lambda_{\min}(\Sigma^z)$, Thus, the population predictive invertibility condition is satisfied as long as the covariance matrix Σ^z is well conditioned.

3.2. Implications to high-dimensional additive models

The prediction performance of additive models has also been carefully investigated in the high-dimensional setting through regularized estimation. In the high-dimensional setting where $d^* > n$, e.g. $p > n$, the sample invertibility condition (20) would not hold for any $\kappa_0 > 0$ as the rank of the matrix $((B_j^\top(X_{i,j}), j \le p)^\top, i \le n)$ cannot be greater than n. A popular remedy to this impasse is to impose the sparsity condition that only a small unknown subset of components $f_1,\ldots,f_p$ is actually non-zero. This is referred to as the sparse additive model and has the natural interpretation that the response Y depends on the design variables only through a small number of them. We use s, the number of non-zero f_j, and the smoothness index of the nonzero f_j to measure the complexity of the sparse additive model. A core assumption in the theory of penalized estimation in the sparse additive model is the restricted eigenvalue and compatibility conditions. Let $\mathcal{I} = \{j : f_j \neq 0\}$ be the unknown index set of real signals and κ_0 and ξ_0 be positive constants, the theoretical restricted eigenvalue and compatibility conditions can be defined as

$$\phi^* = \inf\left\{\frac{|\mathcal{I}|^{2-q}\left\|\sum_{j=1}^p f_j(X_j)\right\|^2_{L_2^{(0)}(P)}}{\left(\sum_{j\in\mathcal{J}}\left\|f_j(X_j)\right\|^q_{L_2^{(0)}(P)}\right)^{2/q}} : \frac{\sum_{j\in\mathcal{I}}\mathrm{Pen}_j(f_j)}{\sum_{j\in\mathcal{I}^c}\mathrm{Pen}_j(f_j)} > \xi_0\right\} \ge \kappa_0 \tag{22}$$

with the convention $0/0 = 0$, where $q = 2$ and $\mathcal{J} = \{1,\ldots,p\}$ for the restricted eigenvalue condition, $q = 1$ and $\mathcal{J} = \mathcal{I}$ for the compatibility coefficient, and $\mathrm{Pen}_j(f_j)$, typically a certain norm of f_j as a regularizer, is the penalty function.

In high-dimensional linear regression, the sample restricted eigenvalue and compatibility conditions were respectively proposed in [1] and [9]. Condition (22), which generalizes the population predictive invertibility condition (21) imposed in the low-dimensional setting, is comparable to the key invertibility conditions imposed in [13] and [18] for $q = 2$ and [16] and [19] for $q = 1$.

To make the dependence on the compatibility condition more explicit, Theorem 1 of Meier et al. [16] establishes that, in the case where all the unknown functions are twice differentiable, the rate of convergence in terms of the in-sample prediction accuracy is $s(\log p/n)^{4/5}/\phi_n$, where ϕ_n is a sample version of ϕ^* defined in (22). Moreover, Theorem 2 of Meier et al. [16] and its proof show that as long as the theoretical compatibility condition (22) holds, ϕ_n and ϕ^* are of the same order and the rate of the population prediction error is $s(\log p/n)^{4/5}/\phi^*$. This directly illustrates the role of ϕ^* defined in the theoretical compatibility condition (22) on the rate of convergence.

Despite the importance of (22) in theoretical justification of regularized prediction in sparse additive models, it has been typically imposed as a condition but without further verification of its validity other than in some very special cases such as the class of densities of $X_{1:p}$ on $[0,1]^p$ uniformly bounded away from 0 and ∞. The result of the current paper on the minimum eigenvalue of the nonlinear correlation matrix sheds light on the theoretical restricted eigenvalue and compatibility conditions for additive models in the sense that condition (22) is satisfied with κ_0 being the minimum eigenvalue of the correlation matrix of the latent pairwise Gaussian vector $Z_{1:p}$ as in Corollary 2.

Corollary 6. *Suppose $\left(X_1, X_2, \ldots, X_p\right)$ follows a hidden Gaussian distribution with $X_j = T_j(Z_j)$ for a pairwise Gaussian vector $(Z_1, \ldots, Z_p)$ with $Corr(Z_1, \ldots, Z_p) = \Sigma^z$ and some deterministic functions T_j with $0 < Var(T_j(Z_j)) < \infty$. Then, condition (22) holds with $\kappa_0 = \lambda_{\min}(\Sigma^z)$. In particular, the theoretical restricted eigenvalue and compatibility conditions hold when $\lambda_{\min}(\Sigma^z)$ is strictly bounded away from zero.*

The above corollary implies that condition (22) holds for the Gaussian copula model. To the best of the authors' knowledge, this is a new connection of the theoretical restricted eigenvalue and compatibility conditions to the widely-used model of multivariate dependency.

In addition to verifying the important condition (22), our results also provide the following connection between the rate of convergence in the estimation of the individual components f_j and the prediction rate [13,16–19].

Corollary 7. *Under the same assumption as Corollary 6,*

$$\lambda_{\min}(\Sigma^z) \sum_{i=1}^{p} \|\widehat{f_i} - f_i\|^2_{L_2^{(0)}(\mathbb{P})} \leq \left\| \sum_{i=1}^{p} \widehat{f_i} - \sum_{i=1}^{p} f_i \right\|^2_{L_2^{(0)}(\mathbb{P})} \leq \lambda_{\max}(\Sigma^z) \sum_{i=1}^{p} \|\widehat{f_i} - f_i\|^2_{L_2^{(0)}(\mathbb{P})}.$$

4. Symmetric functions of iid random variables

In this section, we move beyond the pairwise Gaussianity and consider the extreme nonlinear correlation for symmetric functions of iid random variables. We first consider multiple nested sums of iid random variables to directly generalize the results for a pair of nested sums established in Dembo, Kagan and Shepp (2001) and [4]. In Section 4.2, we consider the class of symmetric functions defined on groups of iid random variables and establish the extreme nonlinear correlation in the much broader setting.

4.1. Nested sums

In this section, we consider the extreme nonlinear correlation for multiple nested sums of iid random variables. Specifically, given positive integers $m_1 < m_2 < \cdots < m_p$ and iid non-degenerate random variables $Y_1, Y_2, \ldots$, we consider

$$X_j = S_{m_j} = \sum_{i=1}^{m_j} Y_i \quad \text{for} \quad j = 1, \ldots, p. \tag{23}$$

Here, the non-degeneracy means that the distribution of the random variable is not concentrated at a single point. In the case of $p = 2$, Dembo, Kagan and Shepp (2001) proved that the maximum correlation of S_{m_1} and S_{m_2} is equal to $\sqrt{m_1/m_2}$ if Y has finite second moment, and Bryc et al. [4] proved the same result even without assuming the finite second order moment by investigating the characteristic functions of sums of Y_i. The following theorem extends their results from $p = 2$ to general finite p. Further extensions to general symmetric functions of arbitrary groups of Y_i are given in the next subsection.

Theorem 2. *Let $Y, Y_1, Y_2, \ldots$ be iid non-degenerate random variables and $X_1, X_2, \ldots, X_p$ be nested sums of Y_i with sample sizes $1 \le m_1 \le \cdots \le m_p$ as defined in (23). Then,*

$$\rho^{NL}_{\max}(X_{1:p}; \nu, W) = \lambda_{\max}(R \circ W), \quad \rho^{NL}_{\min}(X_{1:p}; \nu, W) = \lambda_{\min}(R \circ W), \tag{24}$$

where $R = (R_{j,k})_{p\times p}$ is the matrix with elements $R_{jk} = (m_j \wedge m_k)/\sqrt{m_j m_k}$, ν is taken as the counting measure for the extreme nonlinear correlations defined in (5) and (6), and $\circ$ denotes the Schur product. If Y has a finite second moment, then $R \in \mathbb{R}^{p\times p}$ is the correlation matrix of the nested sums $X_j = S_{m_j}$, $1 \le j \le p$, so that (8) holds with $X_{1:p}$ for all measures ν and weights $W_{j,k} = W_{k,j} \ge 0$,

$$\rho^{NL}_{\max}(X_{1:p}; \nu, W) = \rho^{L}_{\max}(X_{1:p}; \nu, W), \quad \rho^{NL}_{\min}(X_{1:p}; \nu, W) = \rho^{L}_{\min}(X_{1:p}; \nu, W).$$

As discussed below Corollary 3, for $m_j = j$ and large p, $\lambda_{\max}(R \circ W)/p$ with weight matrix $W_{j,k}$ is approximately the maximum eigenvalue of the operator $K_{W'}$ in $L_2([0, 1])$ when $W_{j,k} = W'_{j/p,k/p}$ for a $W'_{s,t}$ continuous in $(s, t) \in [0, 1]^2$.

Proof. As $f_j(X_j) = f_j(S_{m_j})$, $m_1 \le \cdots \le m_p$, are symmetric functions of nested variable groups $\{Y_i, i \in G_j\}$ with $G_j = \{1, 2, \ldots, m_j\}$ and $\cap_{j=1}^p G_j = G_1 \ne \emptyset$, it follows from Theorem 3 in the next subsection that

$$\rho^{NL}_{\max}(X_{1:p}, \nu, W) \le \lambda_{\max}(R \circ W), \quad \rho^{NL}_{\min}(X_{1:p}, \nu, W) \ge \lambda_{\min}(R \circ W).$$

We note that ν is the counting measure here. It remains to prove that $\lambda_{\max}(R\circ W)$ and $\lambda_{\min}(R\circ W)$ are attainable by functions $f_j(X_j)$. This would be simple under the second moment condition on Y as we may simply set $f_j(X_j) = X_j$ to achieve $\mathrm{Corr}(X_{1:p}) = R$. In the case of $\mathbb{E}[Y^2] = \infty$, we prove that R is in the closure of the correlation matrices generated by $(f_j(X_j), j \le p)$. This will be done below by proving

$$\lim_{t\to 0+} \rho\big(\sin(tX_j - m_j c_t), \sin(tX_j - m_k c_t)\big) = R_{j,k}, \quad 1 \le j < k \le p, \tag{25}$$

where $c_t \in (-\pi/2, \pi/2)$ is the solution of

$$\mathbb{E}[\sin(tY - c_t)] = 0, \qquad \text{or equivalently} \quad \frac{\mathbb{E}[\sin(tY)]}{\mathbb{E}[\cos(tY)]} = \tan(c_t).$$

We shall choose the sequence $t \to 0+$ such that for each t, $\mathbb{P}\{\sin(t(Y_1 - Y_2)) = 0\} < 1$ so that $\mathbb{P}\{\sin(tY) = 0\} < 1$ and $\mathbb{P}\{\sin(tY - c_t) = 0\} < 1$. This is always feasible when Y is non-degenerate.

As $\mathbb{E}[\sin(tY)] \to 0$ and $\mathbb{E}[\cos(tY)] \to 1$, it suffices to consider small $t > 0$ satisfying $|c_t| \le 1$. Let $Y' = tY - c_t$. As $\big|\sin(y)(1 - \cos(y))\big| \le \sin^2(y) + 2|\sin(y)|I_{\{|y|>2\}}$, we have

$$\begin{aligned}\Big|\mathbb{E}\big[\sin(Y')\cos(Y')\big]\Big| &= \Big|\mathbb{E}\big[\sin(Y')(1 - \cos(Y'))\big]\Big| \\ &\le \mathbb{E}\big[\sin^2(Y')\big] + \sqrt{\mathbb{E}\big[\sin^2(Y')\big]\mathbb{P}\{|Y| > 1/t\}}.\end{aligned} \tag{26}$$

Let $Y_i' = tY_i - c_t$ and $S_{a:m}' = \sum_{i=a}^m Y_i'$. We shall prove that for $a \le b \le m \le n$

$$\lim_{t\to 0+} \rho\big(\sin(S_{a:m}'), \sin(S_{b:n}')\big) = \frac{(m-b+1)}{(m-a+1)^{1/2}(n-b+1)^{1/2}}. \tag{27}$$

This implies (25) with $a = b = 1$, $m = m_j$ and $n = m_k$, but the more general a and b would provide the extension to sums of arbitrary subgroups of Y_i later in Corollary 8.

Let $f_{a,m} = \sin(S_{a:m}')$. As $\sin(y+z) = \sin(y)\cos(z) + \cos(y)\sin(z)$. We write

$$f_{a,m} = \sum_{u=a}^m f_{a,m,u} \quad \text{where} \quad f_{a,m,u} = \Bigg(\prod_{i=a}^{u-1}\cos(Y_i')\Bigg)\sin(Y_u')\cos(S_{(u+1):m}').$$

Let $a \le b \le m \le n$. As $\mathbb{E}[\sin(Y_a')] = 0$, we have $\mathbb{E}[f_{a,m}] = 0$ and $\mathbb{E}[f_{a,m,u}f_{b,n,v}] = 0$ for $a \le u < b$ or for $m < v \le n$. For $b \le u \wedge v \le u \vee v \le m$,

$$\begin{aligned} & f_{a,m,u}f_{b,n,v} \\ &= \Bigg(\prod_{i=a}^{u-1}\cos(Y_i')\Bigg)\sin(Y_u')\cos(S_{(u+1):m}')\Bigg(\prod_{i=b}^{v-1}\cos(Y_i')\Bigg)\sin(Y_v')\cos(S_{(v+1):n}') \\ &= \sin(Y_{u\wedge v}')\cos(Y_{u\wedge v}')\sin(Y_{u\vee v}')g(Y_i', a \le i \le n, i \ne u\wedge v) \end{aligned}$$

for a certain function g bounded by 1. Thus, as a consequence of (26)

$$\begin{aligned} \big|\mathbb{E}\big[f_{a,m,u}f_{b,n,v}\big]\big| &\le \Big|\mathbb{E}\big[\sin(Y')\cos(Y')\big]\Big|\mathbb{E}\big[|\sin(Y')|\big] \\ &\le \mathbb{E}\big[\sin^2(Y')\big]\Big(\sqrt{\mathbb{E}\big[\sin^2(Y')\big]} + \sqrt{\mathbb{P}\{|Y| > 1/t\}}\Big) \end{aligned}$$

for $b \le u \wedge v < u \vee v \le m$. Moreover, for $b \le u \le m$,

$$\begin{aligned} & \mathbb{E}\big[f_{a,m,u}f_{b,n,u}\big] \\ &= \mathbb{E}\big[\sin^2(Y_u')\big]\mathbb{E}\Bigg[\Bigg(\prod_{i=a}^{b-1}\cos(Y_i')\Bigg)\Bigg(\prod_{i=b}^{u-1}\cos^2(Y_i')\Bigg)\cos\big(S_{(u+1):m}'\big)\cos\big(S_{(u+1):n}'\big)\Bigg]. \end{aligned}$$

Thus, as $Y_i' = tY_i - c_t \to 0$ in probability, we find that for all $a \le b \le m \le n$

$$\lim_{t\to 0+}\frac{\mathbb{E}\big[\sin(S_{a:m}')\sin(S_{b:n}')\big]}{\mathbb{E}\big[\sin^2(Y')\big]} = \lim_{t\to 0+}\sum_{u=a}^m\sum_{v=b}^n\frac{\mathbb{E}\big[f_{a,m,u}f_{b,n,v}\big]}{\mathbb{E}\big[\sin^2(Y')\big]} = \#\big\{b \le u = v \le m\big\}.$$

This implies (27) and completes the proof. $\square$

4.2. Symmetric functions of groups of variables

In this section, we consider a broader setting than nested sums considered in Section 4.1. We use $\{Y_i\}_{i\ge 1}$ to denote an infinite sequence of iid random variables and define random vectors $\boldsymbol{X}_j = (Y_i, i \in G_j)$ for arbitrary sets of positive integers G_j of finite size $m_j = |G_j| < \infty$. Again we are interested in the extreme nonlinear correlation among $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$.

As $\boldsymbol{X}_j$ are vectors, we adjust the definition of the extreme nonlinear correlations in (5) and (6) as follows: Given a $p \times p$ symmetric matrix $W = (W_{j,k})$ with $W_{j,k} \ge 0$, define

$$\rho_{\max,\,\text{symm}}^{NL} = \rho_{\max,\,\text{symm}}^{NL}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p, W) = \sup_{f_{1:p}\in\mathcal{F}_{1:p}} \lambda_{\max}\big(K_{W,f_{1:p}}\big), \tag{28}$$

where $\mathcal{F}_{1:p} = \{(f_1, \ldots, f_p) : 0 < \mathrm{Var}(f_j(\boldsymbol{X}_j)) < \infty, f_j(y_1, \ldots, y_{m_j}) \text{ symmetric } \forall 1 \le j \le p\}$ and $K_{W,f_{1:p}} = \big(\mathrm{Corr}\big(f_j(\boldsymbol{X}_j), f_k(\boldsymbol{X}_k)\big) W_{j,k}\big)_{p\times p}$. Correspondingly, define

$$\rho^{NL}_{\min,\,\mathrm{symm}} = \rho^{NL}_{\min,\,\mathrm{symm}}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p, W) = \inf_{f_{1:p}\in\mathcal{F}_{1:p}} \lambda_{\min}\big(K_{W,f_{1:p}}\big). \tag{29}$$

We omit ν in the notation as it is taken as the counting measure in $\{1, \ldots, p\}$ without loss of generality as discussed below (7). Here, the symmetry of f_j means permutation invariance, $f_j(y_1, \ldots, y_{m_j}) = f_j(y_{i_1}, \ldots, y_{i_{m_j}})$ for all permutations $i_1, \ldots, i_{m_j}$ of $1, \ldots, m_j$. To avoid confusion, we call the above quantities extreme symmetric nonlinear correlations. We extend Theorem 2 to groups satisfying the following assumption.

Assumption C. There exist certain sets $G_{0,j}$ of positive integers such that

$$|G_{0,j} \cap G_{0,k}| = \big(|G_j \cap G_k| - 1\big)_+ \ \forall 1 \le j < k \le p, \quad |G_{0,j}| \le |G_j| - 1 \ \forall 1 \le j \le p.$$

Assumption C holds when $\cap_{j=1}^p G_j \ne \emptyset$, as we can simply set $G_{0,j} = G_j \setminus \{i_0\}$ for a fixed $i_0 \in \cap_{j=1}^p G_j$. Hence, for the special case that G_j are nested with $\emptyset \ne G_1 \subset G_2 \subset \cdots \subset G_p$, Assumption C holds automatically. However, $G_{0,j}$ do not need to have anything to do with G_j beyond the specified conditions on their size and the size of their intersections.

Theorem 3. *Let $Y, Y_1, Y_2, \ldots$ be iid non-degenerate random variables and $\boldsymbol{X}_j = (Y_i, i \in G_j)$ for arbitrary groups of positive integers $G_1, \ldots, G_p$ of finite size $m_j = |G_j| < \infty$. Let $\rho^{NL}_{\max,\,\mathrm{symm}}$ and $\rho^{NL}_{\min,\,\mathrm{symm}}$ be the extreme symmetric nonlinear correlations among $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$ as defined in (28) and (29) with weight matrix W. Let $R^{(\ell)} \in \mathbb{R}^{p\times p}$ be the matrix with elements*

$$R^{(\ell)}_{j,k} = \binom{|G_j \cap G_k|}{\ell}\binom{|G_j|}{\ell}^{-1/2}\binom{|G_k|}{\ell}^{-1/2} \tag{30}$$

for $1 \le \ell \le \ell^*$, with the convention $0/0 = 0$, where $\ell^* = \max_{1\le j\le p} |G_j|$. Then,

$$\rho^{NL}_{\max,\mathrm{symm}} = \lambda_{\max}(R\circ W), \quad \rho^{NL}_{\min,\,\mathrm{symm}} = \min_{1\le\ell\le\ell^*} \lambda_{\min}\Big(\big(R^{(\ell)} \circ W\big)_{J^{(\ell)},J^{(\ell)}}\Big), \tag{31}$$

with $R = R^{(1)}$, $J^{(\ell)} = \{1 \le j \le p : |G_j| \ge \ell\}$ and $\circ$ being the Schur product. If in addition Assumption C holds, then

$$\rho^{NL}_{\min,\,\mathrm{symm}} = \lambda_{\min}\big(R\circ W\big). \tag{32}$$

The first part of (31) asserts that the maximum symmetric nonlinear correlation is identical to its linear version, while the second part gives a formula for the minimum symmetric nonlinear correlation. Under Assumption C, (32) asserts the equality between the minimum symmetric nonlinear correlation and its linear version. The connection between Theorems 2 and 3 can be built under the observation that $f_j(\sum_{i=1}^{m_j} Y_i)$ is a symmetric function of $\boldsymbol{X}_j = \{Y_i\}_{i\in G_j}$ when $G_j = \{1, 2, \ldots, m_j\}$, and the corresponding index sets G_j satisfy Assumption C due to the nested structure of $\{G_j\}_{1\le j\le p}$. For the case $p = 2$, Theorem 3 serves as an extension of Dembo, Kagan and Shepp (2001) from functions $f_j(\sum_{i=1}^{m_j} Y_i)$ of the two sums to any symmetric functions of iid random variables and of Yu [21] from two $f_j(\sum_{i\in G_j} Y_i)$ with arbitrary G_j.

An interesting aspect of Theorem 3 is that under Assumption C the extreme symmetric nonlinear correlation is attained by sums of the form

$$f_j(\boldsymbol{X}_j) = \sum_{i\in G_j} h_0(Y_i) \quad \text{for } 1 \le j \le p, \tag{33}$$

for any function h_0 with $0 < \mathrm{Var}(h_0(Y)) < \infty$, e.g. $h_0(Y_i) = Y_i$ when Y_i has finite variance. That is to say, among symmetric functions, the most extreme multivariate correlations are achieved by the linear summation of iid random variables. The following corollary, based on Theorem 3 and (27) in the proof of Theorem 2, asserts that the extreme symmetric nonlinear correlations for groups of Y_i are achieved by functions of the corresponding sums of Y_i without assuming the finite second moment condition.

Corollary 8. *Let $\boldsymbol{X}_j = (Y_i, i \in G_j)$ and $S_{G_j} = \sum_{i\in G_j} Y_i$ with iid non-degenerate Y_i. Then,*

$$\rho^{NL}_{\max,\,\mathrm{symm}}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p, W) = \rho^{NL}_{\max}\big((S_{G_1}, \ldots, S_{G_p}), \nu, W\big) = \lambda_{\max}\big(R\circ W\big),$$
$$\rho^{NL}_{\min,\,\mathrm{symm}}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p, W) = \rho^{NL}_{\min}\big((S_{G_1}, \ldots, S_{G_p}), \nu, W\big) = \lambda_{\min}\big(R\circ W\big),$$

under Assumption C, where ν is taken as the counting measure in the extreme nonlinear correlations in (5) and (6). Consequently, (8) holds for $X_j = S_{G_j}$ when $\mathbb{E}[Y^2] < \infty$.

The proof of Theorem 3 relies on the Hoeffding [11,12] decomposition of symmetric functions of random variables, stated as Lemma 3; See Lemma 1 in [12], the decomposition lemma in [7], and Lemma 1 in [6].

Lemma 3. *Let $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$ with iid components Y_i and $f_0(\boldsymbol{Y}) = f_0(Y_1, \ldots, Y_m)$ with a symmetric function $f_0(y_1, \ldots, y_m)$. Suppose $\mathbb{E}[f_0(\boldsymbol{Y})] = 0$ and $\mathbb{E}[f_0^2(\boldsymbol{Y})] < \infty$. Define $f_{0,1}(y_1) = \mathbb{E}[f_0(\boldsymbol{Y})|Y_1 = y_1]$ and for $k = 2, \ldots, m$ define*

$$f_{0,k}(y_{1:k}) = \mathbb{E}\left[f_0(\boldsymbol{Y}) - \sum_{j=1}^{k-1} \sum_{1\le i_1<\cdots<i_j\le m} f_{0,j}(Y_{i_1}, \ldots, Y_{i_j}) \middle| Y_{1:k} = y_{1:k} \right].$$

Then, the following expansion holds,

$$f_0(\boldsymbol{Y}) = \sum_{\ell=1}^{m} \sum_{1\le i_1<\cdots<i_\ell\le m} f_{0,\ell}(Y_{i_1}, \ldots, Y_{i_\ell}), \tag{34}$$

and that for all $s = 1, \ldots, \ell$ and $\ell = 1, \ldots, m$

$$\mathbb{E}\Big[f_{0,\ell}(Y_{i_1}, \ldots, Y_{i_\ell}) \Big| \{Y_{i_1}, \ldots, Y_{i_\ell}\}\backslash Y_{i_s} \Big] = 0. \tag{35}$$

Consequently,

$$\mathbb{E}\big[f_0^2(\boldsymbol{Y})\big] = \sum_{\ell=1}^{m} \binom{m}{\ell} \mathbb{E}\Big[f_{0,\ell}^2(Y_{1:\ell}) \Big]. \tag{36}$$

Proof of Theorem 3. Assume without generality $\mathbb{E}[f_j(\boldsymbol{X}_j)] = 0, \mathbb{E}[f_j^2(\boldsymbol{X}_j)] = 1$ for all j as $\rho^{NL}_{\max,\mathrm{symm}}$ and $\rho^{NL}_{\min,\,\mathrm{symm}}$ are defined through the correlations between $f_j(\boldsymbol{X}_j)$ and $f_k(\boldsymbol{X}_k)$. Let $G^{(\ell)} = \{(i_1, \ldots, i_\ell) : i_1 < \cdots < i_\ell, i_s \in G \text{ for } 1 \le s \le \ell\}$ for all subsets G of positive integers. Since $f_j(\boldsymbol{X}_j)$ are symmetric functions of $\{Y_i\}_{i\in G_j}$, (34) gives

$$f_j(\boldsymbol{X}_j) = \sum_{\ell=1}^{m_j} \sum_{(i_1,\ldots,i_\ell)\in G_j^{(\ell)}} f_{j,\ell}(Y_{i_1}, \ldots, Y_{i_\ell}). \tag{37}$$

We first apply (35) and obtain the following expression for the cross-product,

$$\mathbb{E}\Big[f_{j,\ell}(Y_{i_1}, \ldots, Y_{i_\ell}) f_{k,\ell'}(Y_{i'_1}, \ldots, Y_{i'_{\ell'}}) \Big] = 0$$

when $\{i_1, \ldots, i_\ell\} \neq \{i'_1, \ldots, i'_{\ell'}\}$. It follows that

$$\mathbb{E} f_j(\boldsymbol{X}_j) f_k(\boldsymbol{X}_k) = \mathbb{E} \sum_{\ell=1}^{|G_j \cap G_k|} \sum_{(i_1,\ldots,i_\ell)\in(G_j\cap G_k)^{(\ell)}} f_{j,\ell}(Y_{i_1}, \ldots, Y_{i_\ell}) f_{k,\ell}(Y_{i_1}, \ldots, Y_{i_\ell})$$

$$= \sum_{\ell=1}^{\ell^*} \binom{|G_j \cap G_k|}{\ell} \mathbb{E}\Big[f_{j,\ell}(Y_{1:\ell}) f_{k,\ell}(Y_{1:\ell}) \Big] \tag{38}$$

with the convention $\binom{m}{\ell} = 0$ for $\ell > m$. Let $R^{(\ell)} \in \mathbb{R}^{p\times p}$ be the matrix defined in (30). Let $g_{j,\ell} = g_{j,\ell}(Y_{1:\ell}) = \binom{m_j}{\ell}^{1/2} f_{j,\ell}(Y_{1:\ell})$. For $u = (u_1, \ldots, u_p)^\intercal$ with $\|u\|_2 = 1$, (38) provides

$$\begin{aligned}
u^\intercal K_{W,f_{1:p}} u &= \mathbb{E}\left(\sum_{j=1}^{p}\sum_{k=1}^{p} W_{j,k} u_j u_k f_j(\boldsymbol{X}_j) f_k(\boldsymbol{X}_k) \right) \\
&= \sum_{j=1}^{p}\sum_{k=1}^{p} W_{j,k} u_j u_k \sum_{\ell=1}^{\ell^*} \binom{|G_j \cap G_k|}{\ell} \mathbb{E}\left[f_{j,\ell}(Y_{1:\ell}) f_{k,\ell}(Y_{1:\ell}) \right] \\
&= \sum_{\ell=1}^{\ell^*} \mathbb{E}\left[\sum_{j\in J^{(\ell)}} \sum_{k\in J^{(\ell)}} R^{(\ell)}_{j,k} W_{j,k} u_j u_k g_{j,\ell}(Y_{1:\ell}) g_{k,\ell}(Y_{1:\ell}) \right] \\
&\le \max_{1\le\ell\le\ell^*} \lambda_{\max}\Big(\big(R^{(\ell)} \circ W\big)_{J^{(\ell)},J^{(\ell)}} \Big) \sum_{\ell=1}^{\ell^*} \mathbb{E}\left[\sum_{j\in J^{(\ell)}} u_j^2 g_{j,\ell}^2(Y_{1:\ell}) \right] \\
&= \max_{1\le\ell\le\ell^*} \lambda_{\max}\Big(\big(R^{(\ell)} \circ W\big)_{J^{(\ell)},J^{(\ell)}} \Big) \sum_{j=1}^{p} \mathbb{E}\left[u_j^2 f_j^2(\boldsymbol{X}_j) \right] \\
&= \max_{1\le\ell\le\ell^*} \lambda_{\max}\Big(\big(R^{(\ell)} \circ W\big)_{J^{(\ell)},J^{(\ell)}} \Big),
\end{aligned}$$

where the second to the last equality follows from (36) and the fact that $g_{j,\ell} = 0$ for $\ell > m_j = |G_j|$. Similarly, for all $u = (u_1, \ldots, u_p)^\intercal$ with $\|u\|_2 = 1$,

$$u^\intercal K_{W,f_{1:p}} u \ge \min_{1\le\ell\le\ell^*} \lambda_{\min}\Big(\big(R^{(\ell)} \circ W\big)_{J^{(\ell)},J^{(\ell)}} \Big).$$

Thus, by (28) and (29),

$$\begin{aligned}
\rho^{NL}_{\max,\mathrm{symm}} &\le \max_{1\le\ell\le\ell^*} \lambda_{\max}\Big(\big(R^{(\ell)} \circ W\big)_{J^{(\ell)},J^{(\ell)}} \Big), \\
\rho^{NL}_{\min,\mathrm{symm}} &\ge \min_{1\le\ell\le\ell^*} \lambda_{\min}\Big(\big(R^{(\ell)} \circ W\big)_{J^{(\ell)},J^{(\ell)}} \Big).
\end{aligned} \tag{39}$$

To prove (39) holds with equality, we pick a specific $f_{1:p}$ for each ℓ as follows. Let h_0 be a function satisfying $\mathbb{E}[h_0(Y)] = 0$ and $\mathbb{E}[h_0^2(Y)] = 1$. For $j \in J^{(\ell)}$ define

$$h^{(\ell)}_{0,j}(\boldsymbol{X}_j) = \binom{|G_j|}{\ell}^{-1/2} \sum_{|S|=\ell, S\subseteq G_j} \prod_{i\in S} h_0(Y_i) \tag{40}$$

as symmetric functions of $\boldsymbol{X}_j$. For $\{j, k\} \subset J^{(\ell)}$ we have

$$\mathbb{E}\Big[h^{(\ell)}_{0,j}(\boldsymbol{X}_j) h^{(\ell)}_{0,k}(\boldsymbol{X}_k) \Big] = \binom{|G_j \cap G_k|}{\ell} \binom{|G_j|}{\ell}^{-1/2} \binom{|G_k|}{\ell}^{-1/2} I_{\{\ell\le|G_j\cap G_k|\}} = R^{(\ell)}_{j,k}.$$

Thus, when $f_j(\boldsymbol{X}_j) = h_{0,j}^{(\ell)}(\boldsymbol{X}_j)$ for $j \in J^{(\ell)}$, we have $(K_{W,f_{1:p}})_{J^{(\ell)},J^{(\ell)}} = \big(R^{(\ell)} \circ W\big)_{J^{(\ell)},J^{(\ell)}}$. As this holds for every $\ell \le \ell^*$, (39) holds with equality by (28) and (29). We note that $\mathbb{E}[f_j(\boldsymbol{X}_j)f_k(\boldsymbol{X}_k)] \neq 0 = R_{j,k}^{(\ell)}$ typically holds for $j \notin J^{(\ell)}$ or $k \notin J^{(\ell)}$ as $\mathbb{E}[f_j^2(\boldsymbol{X}_j)] = 1$.

It remains to prove that the extreme eigenvalues in (39) are achieved with $\ell = 1$. For the maximum eigenvalue, we notice that by (30),

$$\begin{aligned} R_{j,k}^{(\ell)} &= \binom{|G_j \cap G_k|}{\ell}\binom{|G_j|}{\ell}^{-1/2}\binom{|G_k|}{\ell}^{-1/2} I_{\{|G_j\cap G_k|\ge \ell\}} \\ &= \frac{|G_j \cap G_k|(|G_j \cap G_k|-1)\cdots(|G_j\cap G_k| - l + 1)I_{\{|G_j\cap G_k|\ge\ell\}}}{\sqrt{|G_j|(|G_j|-1)\cdots(|G_j|-l+1)}\sqrt{|G_k|(|G_k|-1)\cdots(|G_k|-l+1)}} \\ &\le \frac{|G_j\cap G_k|}{\sqrt{|G_j|\cdot|G_k|}}, \end{aligned}$$

so that $0 \le R_{j,k}^{(\ell)}W_{j,k} \le R_{j,k}^{(1)}W_{j,k}$ for all $\{j,k\} \subseteq J^{(\ell)}$. Thus,

$$\lambda_{\max}\Big(\big(R^{(\ell)} \circ W\big)_{J^{(\ell)},J^{(\ell)}}\Big) \le \lambda_{\max}\Big(\big(R^{(1)} \circ W\big)_{J^{(\ell)},J^{(\ell)}}\Big) \le \lambda_{\max}\big(R^{(1)} \circ W\big). \tag{41}$$

due to the element-wise positiveness of $\big(R^{(\ell)} \circ W\big)_{J^{(\ell)},J^{(\ell)}}$. As (39) holds with equality, this completes the proof of (31).

The remaining of the proof is to characterize $\min_{1\le\ell\le\ell^*}\lambda_{\min}\big(\big(R^{(\ell)} \circ W\big)_{J^{(\ell)},J^{(\ell)}}\big)$ under Assumption C. As the result can be of independent interest, we state it in the following lemma and supply a proof immediately after the lemma.

Lemma 4. *Under Assumption C, we have*

$$\min_{1\le\ell\le\ell^*}\lambda_{\min}\big(\big(R^{(\ell)} \circ W\big)_{J^{(\ell)},J^{(\ell)}}\big) = \lambda_{\min}\big(R \circ W\big)$$

where $R^{(\ell)}$ are defined in (30), $R^{(1)} = R$ and $\ell^* = \max_{1\le j\le p}|G_j|$.

Proof of Lemma 4. Under Assumption C, we set

$$g_{0,j}^{(\ell-1)}(\boldsymbol{X}_j) = \binom{|G_j|-1}{\ell-1}^{-1/2}\sum_{|S|=\ell-1,\, S\subseteq G_{0,j}}\prod_{i\in S} h_0(Y_i), \quad j \in J^{(\ell)}, 2 \le \ell \le \ell^*,$$

with the h_0 in (40). Similar to the proof of the first part of (39) with equality, we have

$$\begin{aligned} &\mathbb{E}\Big[g_{0,j}^{(\ell-1)}(\boldsymbol{X}_j)g_{0,k}^{(\ell-1)}(\boldsymbol{X}_k)\Big] \\ &= \binom{|G_{0,j}\cap G_{0,k}|}{\ell-1}\binom{|G_j|-1}{\ell-1}^{-1/2}\binom{|G_k|-1}{\ell-1}^{-1/2} I_{\{|G_{0,j}\cap G_{0,k}|\ge\ell-1\}} \\ &= \binom{(|G_j\cap G_k|-1)_+}{\ell-1}\binom{|G_j|-1}{\ell-1}^{-1/2}\binom{|G_k|-1}{\ell-1}^{-1/2} I_{\{|G_j\cap G_k|\ge\ell\}}. \end{aligned}$$

For $j = k$, $\mathrm{Var}\big(g_{0,j}^{(\ell-1)}(\boldsymbol{X}_j)\big) \le 1$ as $|G_{0,j}| \le |G_j| - 1$. It follows that, for $|G_j \cap G_k| \ge 1$,

$$R_{j,k}\mathbb{E}\Big[g_{0,j}^{(\ell-1)}(\boldsymbol{X}_j)g_{0,k}^{(\ell-1)}(\boldsymbol{X}_k)\Big]\begin{cases} = R_{j,k}^{(\ell)}, & j \neq k \text{ in } J^{(\ell)}, \\ \le R_{j,k}^{(\ell)}, & j = k \in J^{(\ell)}. \end{cases}$$

For the case $|G_j \cap G_k| = 0$, the above relationship trivially holds with both sides equal to zero. It follows that when $\lambda_{\min}(R \circ W) \le 0$,

$$\begin{aligned}
&\lambda_{\min}\big((R^{(\ell)} \circ W)_{J^{(\ell)},J^{(\ell)}}\big)\\
&= \min_{\|\boldsymbol{u}_{J^{(\ell)}}\|_2=1} \sum_{j\in J^{(\ell)}} \sum_{k\in J^{(\ell)}} u_j u_k R^{(\ell)}_{j,k} W_{j,k}\\
&\ge \min_{\|\boldsymbol{u}\|_2=1} \sum_{j\in J^{(\ell)}} \sum_{k\in J^{(\ell)}} u_j u_k R_{j,k} W_{j,k} \mathbb{E}\Big[g^{(\ell-1)}_{0,j}(\boldsymbol{X}_j) g^{(\ell-1)}_{0,k}(\boldsymbol{X}_k)\Big]\\
&\ge \min_{\|\boldsymbol{u}\|_2=1} \lambda_{\min}(R\circ W)\mathbb{E}\Bigg[\sum_{j\in J^{(\ell)}} \Big(u_j g^{(\ell-1)}_{0,j}(\boldsymbol{X}_j)\Big)^2\Bigg]\\
&\ge \lambda_{\min}(R\circ W),
\end{aligned}$$

where the last inequality holds due to the fact that

$$\mathbb{E}\Big[(g^{(\ell-1)}_{0,j}(\boldsymbol{X}_j))^2\Big] = \binom{|G_{0,j}|}{\ell-1}\binom{|G_j|-1}{\ell-1}^{-1} I_{\{|G_{0,j}|\ge \ell-1\}} \le 1.$$

In the general case, we notice that $\lambda_{\min}(R\circ(W - cI_{p\times p})) \le 0$ and $W_{j,k} - cI_{\{j=k\}} \ge 0$ for all $1 \le j,k \le p$ and $c = \min_{1\le j\le p} W_{j,j}$, so that

$$\begin{aligned}
\min_{f_{1:p}\in\mathcal{F}_{1:p}} \lambda_{\min}\big(K_{W,f_{1:p}}\big) - c &= \min_{f_{1:p}\in\mathcal{F}_{1:p}} \lambda_{\min}\big(K_{1,f_{1:p}} \circ (W - cI_{p\times p})\big)\\
&= \lambda_{\min}\big(R\circ(W - cI_{p\times p})\big)\\
&= \lambda_{\min}\big(R\circ W\big) - c
\end{aligned}$$

as the diagonal of $K_{1,f_{1:p}}$ and R are both $I_{p\times p}$. □

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank the Associate Editor and two referees whose constructive comments led to the examples in Section 2.3.

Appendix

We prove Lemmas 1 and 2 in this Appendix.

Proof of Lemma 1. Let $g_t(x) = h(t)f_t(x)\big/\big\{\mathbb{E}\big[f_t^2(X_t)\big]\big\}^{1/2}$ with h satisfying $\|h\|^2_{L_2(\nu)} = 1$. As $\int_{\mathcal{T}} \mathbb{E}[g_t^2(X_t)]\nu(dt) = \|h\|^2_{L_2(\nu)} = 1$, $f_{\mathcal{T}} \in \mathcal{F}_{\mathcal{T}}$ implies $g_{\mathcal{T}} \in \mathcal{F}_{\mathcal{T}}$, so that by (5)

$$\begin{aligned}
\rho^{NL}_{\max} &= \sup_{f_{\mathcal{T}}\in\mathcal{F}_{\mathcal{T}}} \sup_{\|h\|_{L_2(\nu)}=1} \int_{s\in\mathcal{T}} \int_{t\in\mathcal{T}} \rho\left(f_s(X_s), f_t(X_t)\right) W_{s,t} h(s)h(t)\nu(ds)\nu(dt)\\
&\le \sup_{g_{\mathcal{T}}\in\mathcal{F}_{\mathcal{T}}} \frac{\int_{s\in\mathcal{T}}\int_{t\in\mathcal{T}} \mathbb{E}\big[g_s(X_s), g_t(X_t)\big] W_{s,t}\nu(ds)\nu(dt)}{\int \mathbb{E}\big[g_t^2(X_t)\big]\nu(dt)}.
\end{aligned}$$

On the other hand, letting $h(t) = \{\mathbb{E}[f_t^2(X_t)] / \int_{t\in\mathcal{T}} \mathbb{E}[f_t^2(X_t)]\nu(dt)\}^{1/2}$, we have

$$\begin{aligned}
\rho_{\max}^{NL} &\geq \int_{s\in\mathcal{T}}\int_{t\in\mathcal{T}} \rho\left(f_s(X_s), f_t(X_t)\right) W_{s,t} h(s)h(t)\nu(ds)\nu(dt) \\
&= \frac{\int_{s\in\mathcal{T}}\int_{t\in\mathcal{T}} \mathbb{E}[f_s(X_s), f_t(X_t)] W_{s,t}\nu(ds)\nu(dt)}{\int_{t\in\mathcal{T}} \mathbb{E}[f_t^2(X_t)]\nu(dt)}.
\end{aligned}$$

for all $f_{\mathcal{T}} \in \mathcal{F}_{\mathcal{T}}$. Thus, (5) and (14) are equivalent. We omit the proof of the equivalence between (6) and (15) as it can be established by the same argument. □

Proof of Lemma 2. Let h be a function on $\mathcal{T}$ with $\|h\|_{L_2(\nu)} = 1$ and $B_n \subset \mathcal{T}$ be as in Assumption A. Let $\{X_t^{(i)}, t \in \mathcal{T}\}_{1\leq i\leq m-1}$ be iid copies of $X_{\mathcal{T}}$. Because $\left|\mathbb{E}[X_s X_t]\right| \leq \mathbb{E}[|X_s^{(i)} X_t^{(i)}|] \leq 1$ and $|K_W(s,t)| \leq W_{s,t}$, by Cauchy–Schwarz

$$\begin{aligned}
&\mathbb{E}\int_{B_n}\int_{B_n}\left(\prod_{i=1}^{m-1}|X_s^{(i)}X_t^{(i)}|\right)\left|K_W(s,t)h(s)h(t)\right|\nu(ds)\nu(dt) \\
&\leq \left(\int_{B_n}\int_{B_n} W_{s,t}^2\nu(ds)\nu(dt)\right)^{1/2} < \infty.
\end{aligned}$$

Thus the exchange of expectation and integration is allowed in the following derivation:

$$\begin{aligned}
&\int_{B_n}\int_{B_n}\left(\mathbb{E}[X_sX_t]\right)^{m-1}K_W(s,t)h(s)h(t)\nu(ds)\nu(dt) \\
&= \mathbb{E}\int_{B_n}\int_{B_n}\left(\prod_{i=1}^{m-1}(X_s^{(i)}X_t^{(i)})\right)K_W(s,t)h(s)h(t)\nu(ds)\nu(dt) \\
&= \mathbb{E}\int_{B_n}\int_{B_n} K_W(s,t)\left\{h(s)\prod_{i=1}^{m-1}X_s^{(i)}\right\}\left\{h(t)\prod_{i=1}^{m-1}X_t^{(i)}\right\}\nu(ds)\nu(dt) \\
&\leq \rho_{\max}^{L}\int \mathbb{E}\left[\left(I\{t\in B_n\}h(t)\prod_{i=1}^{m-1}X_t^{(i)}\right)^2\right]\nu(dt) \\
&= \rho_{\max}^{L}\int_{B_n} h^2(t)\nu(dt).
\end{aligned}$$

Moreover, as the exchange of expectation and integration is allowed,

$$\begin{aligned}
&\int_{B_n}\int_{B_n}\left(\mathbb{E}[X_sX_t]\right)^{m-1}K_W(s,t)h(s)h(t)\nu(ds)\nu(dt) \\
&= \mathbb{E}\int_{B_n}\int_{B_n} K_W(s,t)\left\{h(s)\prod_{i=1}^{m-1}X_s^{(i)}\right\}\left\{h(t)\prod_{i=1}^{m-1}X_t^{(i)}\right\}\nu(ds)\nu(dt) \\
&\geq \rho_{\min}^{L}\int \mathbb{E}\left[\left(I\{t\in B_n\}h(t)\prod_{i=1}^{m-1}X_t^{(i)}\right)^2\right]\nu(dt) \\
&= \rho_{\min}^{L}\int_{B_n} h^2(t)\nu(dt).
\end{aligned}$$

As the operator K_W is bounded by Assumption A, $\rho_{\max}^L$ and $\rho_{\min}^L$ are both finite, so that the inequalities still hold as $B_n \to \mathcal{T}$. □

References

[1] Peter J Bickel, Ya'acov Ritov, Alexandre B Tsybakov, Simultaneous analysis of lasso and dantzig selector, Ann. Statist. 37 (4) (2009) 1705–1732.

[2] Richard C Bradley, et al., Basic properties of strong mixing conditions. a survey and some open questions, Probab. Surv. 2 (2005) 107–144.

[3] Leo Breiman, Jerome H Friedman, Estimating optimal transformations for multiple regression and correlation, J. Amer. Statist. Assoc. 80 (391) (1985) 580–598.

[4] Wlodzimierz Bryc, Amir Dembo, Abram Kagan, On the maximum correlation coefficient, Theory Probab. Appl. 49 (1) (2005) 132–138.

[5] Andreas Buja, Trevor Hastie, Robert Tibshirani, Linear smoothers and additive models, Ann. Statist. (1989) 453–510.

[6] Amir Dembo, Abram Kagan, Lawrence A Shepp, Remarks on the maximum correlation coefficient, Bernoulli 7 (2) (2001) 343–350.

[7] Bradley Efron, Charles Stein, The jackknife estimate of variance, Ann. Statist. (1981) 586–596.

[8] Jerome H Friedman, Werner Stuetzle, Projection pursuit regression, J. Amer. Statist. Assoc. 76 (376) (1981) 817–823.

[9] Sara A van de Geer, Peter Bühlmann, On the conditions used to prove oracle results for the lasso, Electron. J. Stat. 3 (2009) 1360–1392.

[10] Trevor Hastie, Robert Tibshirani, Generalized additive models, Statist. Sci. 1 (3) (1986) 297–318.

[11] Wassily Hoeffding, A class of statistics with asymptotically normal distribution, Ann. Math. Stat. (1948) 293–325.

[12] Wassily Hoeffding, The Strong Law of Large Numbers for U-Statistic, Technical Report, North Carolina State University. Dept. of Statistics, 1961.

[13] Vladimir Koltchinskii, Ming Yuan, Sparsity in multiple kernel learning, Ann. Statist. 38 (6) (2010) 3660–3695.

[14] Henry Oliver Lancaster, Some properties of the bivariate normal distribution considered in the form of a contingency table, Biometrika 44 (1/2) (1957) 289–292.

[15] Wenbo V Li, Comparison results for the lower tail of gaussian seminorms, J. Theor. Probab. 5 (1) (1992) 1–31.

[16] Lukas Meier, Sara Van de Geer, Peter Bühlmann, High-dimensional additive modeling, Ann. Statist. 37 (6B) (2009) 3779–3821.

[17] Garvesh Raskutti, Martin J Wainwright, Bin Yu, Minimax-optimal rates for sparse additive models over kernel classes via convex programming, J. Mach. Learn. Res. 13 (Feb) (2012) 389–427.

[18] Taiji Suzuki, Masashi Sugiyama, Fast learning rate of multiple kernel learning: Trade-off between sparsity and smoohntess, Ann. Statist. (2013) 1381–1405.

[19] Zhiqiang Tan, Cun-Hui Zhang, Doubly penalized estimation in additive regression with high-dimensional data, 2017, arXiv preprint arXiv:1704.07229.

[20] Simon N Wood, Generalized Additive Models: An Introduction with R, Chapman and Hall/CRC, 2017.

[21] Yaming Yu, On the maximal correlation coefficient, Statist. Probab. Lett. 78 (9) (2008) 1072–1075.