# Extreme Nonlinear Correlation for Multiple Random Variables and Stochastic Processes with Applications to Additive Models

Zijian Guo[1] and Cun-Hui Zhang[2]
Rutgers University

*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 nested sums of iid square integrable random variables. This paper extends these results to pairwise Gaussian processes and vectors, nested sums of iid random variables, and permutation symmetric functions of sub-groups of iid random variables.

**Keywords:** Nonlinear Correlation; Gaussian Copula; Theoretical Restricted Eigenvalue; Theoretical Compatibility Condition; Additive Model; Symmetric Functions.

## 1 Introduction

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

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

where $f_1$ and $f_2$ are real functions. If $X_1$ and $X_2$ are bivariate normal, it was established in Lancaster (1957) and Yu (2008) 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

[1] Research partially supported by the NSF DMS 1811857.

[2] Research partially supported by the NSF Grants DMS-1513378, DMS-1721495 and IIS-1741390.

(iid) random variables with finite second moment. In a follow-up work, Bryc et al. (2005) proved $R(X_1, X_2) = \sqrt{m/n}$ for the nested sums without the second moment condition.

The current paper extends the above results to more than two random variables and Gaussian processes. Let $\lambda_{\min}$ and $\lambda_{\max}$ denote the largest and smallest eigenvalues of matrices or linear operators, and $\mathrm{Corr}_{\neq}(X_1, \ldots, X_p)$ the $p \times p$ off-diagonal covariance matrix of $p$ random variables with elements $\rho(X_j, X_k) I_{\{j \neq k\}}$. As

$$|\rho(X_1, X_2)| = \lambda_{\max}\big(\mathrm{Corr}_{\neq}(X_1, X_2)\big) = -\lambda_{\min}\big(\mathrm{Corr}_{\neq}(X_1, X_2)\big),$$

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 $X_1, \ldots, X_p$,

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

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

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

We note that for $p \geq 3$, $\rho_{\min}^{NL}$ is no longer determined by $\rho_{\max}^{NL}$, so that both quantities are needed to capture the extreme nonlinear correlation. Moreover, this extreme multivariate nonlinear correlation leads to the following further extension of the concept to stochastic processes: For $X_{\mathcal{T}} = \{X_t, t \in \mathcal{T}\}$ on an index set $\mathcal{T}$ equipped with a measure $\nu$,

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

where $\|h\|_{L_2(\nu)} = \big\{ \int_{\mathcal{T}} h^2(t) \nu(dt) \big\}^{1/2}$, $I_{\{s \neq t\}}$ is the indicator function for $s \neq t$ 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,

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

Clearly, (3) and (4) are respectively special cases of (5) and (6) with $\mathcal{T} = \{1, \ldots, p\}$ and $\nu$ being the counting measure.

The main assertion of this paper is that in a number of settings, the extreme nonlinear correlation is identical to its linear counterpart:

$$\rho^{NL}_{\max} = \rho^{L}_{\max} \ \text{ and } \ \rho^{NL}_{\min} = \rho^{L}_{\min}, \tag{7}$$

where $\rho^{L}_{\max}$ and $\rho^{L}_{\min}$ are defined by restricting the functions $f_j$ in (3) and (4) and $f_t$ in (5) and (6) to be the identity $f(x) = x$; e.g. in the more general stochastic process setting,

$$\rho^{L}_{\max} = \rho^{L}_{\max}(X_{\mathcal{T}}, \nu) = \sup_{\|h\|_{L_2(\nu)}=1} \int_{t\in\mathcal{T}} \int_{s\in\mathcal{T}} \rho\left(X_s, X_t\right) I_{\{s\neq t\}} h(s)h(t)\nu(ds)\nu(dt), \tag{8}$$

and

$$\rho^{L}_{\min} = \rho^{L}_{\min}(X_{\mathcal{T}}, \nu) = \inf_{\|h\|_{L_2(\nu)}=1} \int_{t\in\mathcal{T}} \int_{s\in\mathcal{T}} \rho\left(X_s, X_t\right) I_{\{s\neq t\}} h(s)h(t)\nu(ds)\nu(dt). \tag{9}$$

Thus, (7) asserts that the extreme nonlinear correlations match the boundary points of the spectrum of the off-diagonal correlation operator.

We will begin by proving (7) for Gaussian processes $X_{\mathcal{T}}$ on an arbitrary index set $\mathcal{T}$ equipped with a $\sigma$-finite measure $\nu$. Our analysis bears some resemblance to that of Lancaster (1957) 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 (7) under proper measurability and integrability conditions.

We shall say that random variables $X_1, \ldots, X_p$ are *hidden Gaussian* if $X_j = T_j(Z_j)$ for a Gaussian vector $Z = (Z_1, \ldots, Z_p)$ and some deterministic transformations $T_j, j \leq p$; $X_1, \ldots, X_p$ are *hidden pairwise Gaussian* if the Gaussian requirement on $Z$ is reduced to pairwise Gaussian. The equivalence of the nonlinear and linear extreme correlations (7) for the pairwise Gaussian process implies that for hidden pairwise Gaussian variables

$$\rho^{L}_{\min}(Z_1, \ldots, Z_p) \leq \rho^{NL}_{\min}(X_1, \ldots, X_p) \leq \rho^{NL}_{\max}(X_1, \ldots, X_p) \leq \rho^{L}_{\max}(Z_1, \ldots, Z_p).$$

That is to say, if the correlation structure among $X_1, X_2, \cdots, X_p$ is generated from a pairwise Gaussian distribution through marginal transformations (even in a hidden way), then their extreme nonlinear correlation is controlled within the spectrum of the off-diagonal correlation matrix of the underlying Gaussian distribution. When $Z_1, \ldots, Z_p$ are 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 multivariate nonlinear correlation 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 dramatically mitigates the curse of dimensionality in the more complex multiple nonparametric regression (Buja et al., 1989; Wood, 2017; Hastie and Tibshirani, 1986). Let $\|f\|_{L_2^{(0)}(\mathbb{P})}$ denote the semi-norm given by $\|f\|^2_{L_2^{(0)}(\mathbb{P})} = \mathrm{Var}(f(X_1, \dots, X_p))$. Our result on the minimum nonlinear correlation has two interesting implications in the analysis of high-dimensional additive models as follows. Firstly, as established in the literature (Meier et al., 2009; Koltchinskii and Yuan, 2010; Raskutti et al., 2012; Suzuki and Sugiyama, 2013; Tan and Zhang, 2017), regularized estimation in the additive model typically yields an error bound on the prediction error $\| \sum_{i=1}^{p} \widehat{f}_i - \sum_{i=1}^{p} f_i \|^2_{L_2^{(0)}(\mathbb{P})}$ under a certain restricted eigenvalue or compatibility condition on the design which would require a strictly positive lower bound for $1 + \rho^{NL}_{\min}$. The characterization of $\rho^{NL}_{\min}$ in the current paper will verify that the required theoretical restricted eigenvalue and compatibility conditions hold for a large class of non-trivial distributions. Secondly, when the minimum nonlinear correlation of $X_1, \dots, X_p$ is bounded away from zero, the squared loss for the estimation of individual $f_i$ can be derived from the prediction error bound via

$$\sum_{i=1}^{p} \|\widehat{f}_i - f_i\|^2_{L_2^{(0)}(\mathbb{P})} \leq \frac{1}{1 + \rho^{NL}_{\min}} \left\| \sum_{i=1}^{p} \widehat{f}_i - \sum_{i=1}^{p} f_i \right\|^2_{L_2^{(0)}(\mathbb{P})}.$$

See Section 2.2 for more detailed discussions.

In addition to the extension of Lancaster (1957) to pairwise Gaussian processes and vectors, the current paper directly extends the results of Dembo, Kagan and Shepp (2001) and Bryc et al. (2005) by establishing (7) for nested sums $(X_1, X_2, \cdots, X_p)$ of iid random variables $Y_i$, with $X_j = \sum_{i=1}^{m_j} Y_i$ for some positive integers $m_j$, $1 \leq j \leq p$. Moreover, as a natural generalization of the nested sums, we consider groups of the iid variables as random vectors $\boldsymbol{X}_i = (Y_j, j \in G_i)$ where $G_i$ are sets of positive integers. We extend the first part of (7) by proving that

$$\max_{f_1, \dots, f_p} \rho^L_{\max}\big(f_1(\boldsymbol{X}_1), \dots, f_p(\boldsymbol{X}_p)\big) = \rho^L_{\max}\big(S_{G_1}, \dots, S_{G_p}\big) \tag{10}$$

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. Throughout the paper, such $f_i$ are called permutation symmetric functions or simply symmetric functions. We also establish the corresponding lower bound

$$\min_{f_1,\ldots,f_p} \rho^L_{\min}\big(f_1(\boldsymbol{X}_1),\ldots,f_p(\boldsymbol{X}_p)\big) = \rho^L_{\min}\big(S_{G_1},\ldots,S_{G_p}\big) \tag{11}$$

under a mild condition, including the case where $\cap_{j=1}^p G_j \neq \emptyset$.

**Paper Organization.** The rest of the paper is organized as follows. In Section 2, we study the extreme nonlinear correlation for pairwise Gaussian random processes or vectors and discuss the implications to additive models; In Section 3, we study the extreme multivariate nonlinear correlation of nested sums and also the more general symmetric functions.

# 2 Pairwise Gaussian Processes

In Section 2.1, we characterize the extreme nonlinear correlations (5) and (6) for pairwise Gaussian processes, and discuss the implications of the result in the multivariate setting, including Gaussian copulas and the more general hidden pairwise Gaussian distributions. In Section 2.2, we discuss applications of the result in 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.

## 2.1 Extreme Nonlinear Correlation for 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 (8), (9), (5) and (6).

**Assumption A:** (i) *The measure $\nu$ is $\sigma$-finite on $\mathcal{T}$.*

(ii) *The process $X_{\mathcal{T}}$ is standardized to $\mathbb{E}[X_t] = 0$ and $\mathbb{E}[X_t^2] = 1$, the kernel $K(s,t) = \mathbb{E}\big[X_s X_t\big] I_{\{s\neq t\}}$ is measurable as a function of $(s,t)$ in the product space $\mathcal{T}\times\mathcal{T}$, and the extreme linear correlations in* (8) *and* (9) *are both finite.*

We note that there is no loss of generality to assume that $X_{\mathcal{T}}$ is standardized as (8) and (9) involve only the correlation between $X_s$ and $X_t$. We also note that the extreme linear correlations in (8) and (9) are both finite if and only if the linear operator $h \to Kh = \int K(\cdot,s)h(s)\nu(ds)$ is bounded in $L_2(\nu)$.

**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 integer $m \geq 1$, the kernel $K_f(s,t) = \mathbb{E}\big[f_s(X_s)f_t(X_t)\big]I_{\{s\neq t\}}$ is measurable as a function of $(s,t)$ on $\mathcal{T}\times\mathcal{T}$, and the linear operator $K_f : h \to \int K_f(\cdot,s)h(s)\nu(ds)$ is bounded.*

We note that 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 $\mathcal{F}_{\mathcal{T}}$ contains 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,*

$$\rho_{\max}^{NL} = \sup_{f_{\mathcal{T}} \in \mathcal{F}_{\mathcal{T}}} \frac{\int_{t \in \mathcal{T}} \int_{s \in \mathcal{T}} \mathbb{E}\big[f_s(X_s), f_t(X_t)\big] I_{\{s \neq t\}} \nu(ds)\nu(dt)}{\int \mathbb{E}\big[f_t^2(X_t)\big]\nu(dt)}, \tag{12}$$

*and*

$$\rho_{\min}^{NL} = \inf_{f_{\mathcal{T}} \in \mathcal{F}_{\mathcal{T}}} \frac{\int_{t \in \mathcal{T}} \int_{s \in \mathcal{T}} \mathbb{E}\big[f_s(X_s), f_t(X_t)\big] I_{\{s \neq t\}} \nu(ds)\nu(dt)}{\int \mathbb{E}\big[f_t^2(X_t)\big]\nu(dt)}. \tag{13}$$

A proof of Lemma 1 can be found in the Appendix. The more explicit expressions established in the lemma would facilitate the Hermite 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 of the off-diagonal correlation kernel.

**Lemma 2.** *Let $\rho_{\max}^{L}$ and $\rho_{\min}^{L}$ be as in* (8) *and* (9) *respectively. Under Assumption A,*

$$\rho_{\min}^{L} \leq \int_{t \in \mathcal{T}} \int_{s \in \mathcal{T}} K^m(s,t) h(s) h(t) \nu(ds)\nu(dt) \leq \rho_{\max}^{L}. \tag{14}$$

*for any integer $m \geq 2$ 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 $K^m(s,t)$ is controlled inside that of $K(s,t)$. The proof of the Lemma, given in the Appendix, utilize an interesting construction of the Schur power kernel $K^m$ with iid copies of $X_{\mathcal{T}}$. Such a proof technique is of independent interest.

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,*

$$\rho_{\max}^{NL} = \rho_{\max}^{L} \quad \textit{and} \quad \rho_{\min}^{NL} = \rho_{\min}^{L},$$

*where $\rho_{\max}^{NL}$ and $\rho_{\min}^{NL}$ are the extreme nonlinear correlations in* (5) *and* (6) *respectively, and $\rho_{\max}^{L}$ and $\rho_{\min}^{L}$ are their linear counterpart in* (8) *and* (9) *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 $K(s,t) = \mathbb{E}\big[X_s, X_t\big]I_{s\neq t}$ be as in Assumption A. As $(X_s, X_t)$ is bivariate normal with $\text{Var}(X_s) = \text{Var}(X_t) = 1$, $\mathbb{E}[H_m(X_s)H_n(X_t)]I_{\{s\neq t\}} = K^m(s,t)I_{\{m=n\}}$ as in Lancaster (1957). It follows that $\mathbb{E}\big[f_s(X_s)f_t(X_t)\big]I_{s\neq t} = \sum_{m=1}^{\infty} K^m(s,t)a_m(s)a_m(t)$ and that by Lemma 2

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

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

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

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

We state in the rest of the subsection some corollaries as immediate consequences of Theorem 1 and Lemma 1.

**Corollary 1.** *Let $\{X_t, 0 \le t \le 1\}$ be a Gaussian process with Lebesgue measurable off-diagonal correlation $K(s,t) = \rho(X_s, X_t)I_{\{s\neq t\}}$ as a function in $(s,t) \in [0,1]^2$. Let $K$ denote the linear operator $h \to \int_0^1 K(\cdot, s)h(s)ds$. Then, for all bounded continuous functions $f(x,t)$,*

$$\lambda_{\min}(K)\int_0^1 \text{Var}\big(f(X_t,t)\big)dt \le \text{Var}\left(\int_0^1 f(X_t,t)dt\right) \le \lambda_{\max}(K)\int_0^1 \text{Var}\big(f(X_t,t)\big)dt.$$

*Equivalently, the extreme nonlinear correlations in* (5) *and* (6) *with* $\mathcal{T} = [0,1]$ *and the Lebesgue measure* $\nu(dx) = dx$ *are given by*

$$\rho_{\max}^{NL}(X_{[0,1]}) = \lambda_{\max}(K) \quad \textit{and} \quad \rho_{\min}^{NL}(X_{[0,1]}) = \lambda_{\min}(K).$$

**Corollary 2.** *Let* $X_1, X_2, \cdots, X_p$ *be pairwise Gaussian random variables with a correlation matrix* $\Sigma \in \mathbb{R}^{p\times p}$. *Then, for all functions* $f_j$ *satisfying* $\mathbb{E}f_j(X_j) = 0$ *and* $\mathbb{E}f^2(X_j) < \infty$,

$$\lambda_{\min}(\Sigma) \cdot \sum_{j=1}^{p} \mathbb{E}f_j^2(X_j) \le \mathbb{E}\left(\sum_{j=1}^{p} f_j(X_j)\right)^2 \le \lambda_{\max}(\Sigma) \cdot \sum_{j=1}^{p} \mathbb{E}f_j^2(X_j). \tag{15}$$

*Equivalently, the extreme nonlinear correlations in* (3) *and* (4) *are given by*

$$\rho_{\max}^{NL}(X_1, \ldots, X_p) = \lambda_{\max}(\Sigma) - 1 \quad \textit{and} \quad \rho_{\min}^{NL}(X_1, \ldots, X_p) = \lambda_{\min}(\Sigma) - 1.$$

Finally, we state in the following corollary the implication of Theorem 1 on Gaussian copula and other hidden pairwise Gaussian variables: the extreme (nonlinear) correlations of such random variables are controlled by the spectrum limits of the off-diagonal covariance matrix of the underlying Gaussian distribution.

**Corollary 3.** *Suppose* $(X_1, X_2, \cdots, X_p)$ *follows a hidden Gaussian distribution in the sense of* $X_j = T_j(Z_j)$ *for a Gaussian vector* $(Z_1, \ldots, Z_p) \sim N(0, \Sigma^z)$ *and some deterministic functions* $T_j$ *with* $0 < \mathrm{Var}(T_j(Z_j)) < \infty$. *Then,*

$$\lambda_{\min}(\Sigma^z) - 1 \le \rho_{\min}^{NL}(X_1, \ldots, X_p) \le \rho_{\max}^{NL}(X_1, \ldots, X_p) \le \lambda_{\max}(\Sigma^z) - 1.$$

*Moreover, the Gaussian assumption on* $(Z_1, \ldots, Z_p)$ *can be weakened to pairwise Gaussian.*

## 2.2 Implications in Additive Models

In high-dimensional additive regression models, the restricted eigenvalue and compatibility conditions are crucial elements of the theory of regularized estimation. These conditions are closely related to the extreme nonlinear correlation as we discuss here.

In the additive regression model, the relationship between the response variable $Y$ and design variables $X_1, \ldots, X_p$ is given by

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

where $\varepsilon \sim N(0, \sigma^2)$ is the noise variable independent of $\{X_1, \ldots, X_p\}$. Let $\mathcal{I} = \{j : f_j \neq 0\}$ be the unknown index set of real signals and $\kappa_0$ and $\xi_0$ be positive constants, the theoretical restricted eigenvalue and compatibility conditions can be defined as

$$\inf \left\{ \frac{|\mathcal{I}|^{1-q/2} \left\| \sum_{j=1}^{p} f_j(X_j) \right\|^2_{L_2^{(0)}(P)}}{\sum_{j \in \mathcal{J}} \left\| f_j(X_j) \right\|^q_{L_2^{(0)}(P)}} : \frac{\sum_{j \in \mathcal{I}} \left\| f_j(X_j) \right\|_{L_2^{(0)}(P)}}{\sum_{j \in \mathcal{I}^c} \left\| f_j(X_j) \right\|_{L_2^{(0)}(P)}} > \xi_0 \right\} \geq \kappa_0 \tag{16}$$

with the convention $0/0 = 0$, with the left-hand side being the restricted eigenvalue for $q = 2$ and $\mathcal{J} \supseteq \mathcal{I}$ and compatibility coefficient for $q = 1$ and $\mathcal{J} = \mathcal{I}$. The above definition generalizes both the restricted eigenvalue condition (Bickel et al., 2009) and the compatibility condition (van de Geer and Bühlmann, 2009) introduced in the high-dimensional regression.

Regarding the analysis of high-dimensional additive models, the condition (16) with $q = 2$ has been used in Koltchinskii and Yuan (2010); Suzuki and Sugiyama (2013) as a key assumption. The condition (16) with $q = 1$ and $\mathcal{J} = \mathcal{I}$ has been used in Tan and Zhang (2017) to establish the prediction accuracy of the high-dimensional sparse additive models. Despite the importance of (16), it has been typically imposed as a condition but without verifying its validity other than in some very special cases such as the class of densities on $[0, 1]^p$ uniformly bounded away from 0 and $\infty$. The result of the current paper on extreme multivariate nonlinear correlation will shed light on the restricted eigenvalue or theoretical compatibility condition for additive models, in the sense that the condition (16) is satisfied with $\kappa_0$ being the minimum eigenvalue of the correlation matrix. Such a result is stated in the following corollary, as a consequence of combining Corollary 2 and Lemma 1.

**Corollary 4.** *Suppose $(X_1, X_2, \cdots, X_p)$ follows a hidden Gaussian distribution with $X_j = T_j(Z_j)$ for a pairwise Gaussian vector $(Z_1, \ldots, Z_p)$ with $\mathrm{Corr}(Z_1, \ldots, Z_p) = \Sigma^z$ and some deterministic functions $T_j$ with $0 < \mathrm{Var}(T_j(Z_j)) < \infty$. Then, the condition (16) holds with $\kappa_0 = \lambda_{\min}(\Sigma^z)$. In particular, if $\lambda_{\min}(\Sigma^z) > 0$ is a positive constant, then the theoretical restricted value condition $(q = 2, \mathcal{J} \supseteq \mathcal{I})$ and compatibility condition $(q = 1, \mathcal{J} = \mathcal{I})$ hold.*

The above corollary implies that the condition (16) holds for the Gaussian copula model, where the variable $X_j = F_j(Z_j) \in [0, 1]$ for $1 \leq j \leq p$ is generated by the underlying pairwise Gaussian random variables $(Z_1, Z_2, \cdots, Z_p)$ and $\{F_j\}_{1 \leq j \leq p}$ are the corresponding cumulative distribution function. To the best of the authors' knowledge, this is a new connection of the theoretical restricted eigenvalue and compatibility conditions to the minimum eigenvalue of the correlation matrix.

In addition to verifying the important condition (16), we can also apply the minimum multivariate nonlinear correlation to connect the rate of convergence for estimating the individual components $f_j$ to the prediction error established in the literature (Meier et al., 2009; Koltchinskii and Yuan, 2010; Raskutti et al., 2012; Suzuki and Sugiyama, 2013; Tan and Zhang, 2017).

**Corollary 5.** *Under the same assumption as Corollary 5,*

$$\sum_{i=1}^{p} \|\widehat{f}_i - f_i\|^2_{L_2^{(0)}(\mathbb{P})} \leq \frac{1}{\lambda_{\min}(\Sigma^z)} \left\| \sum_{i=1}^{p} \widehat{f}_i - \sum_{i=1}^{p} f_i \right\|^2_{L_2^{(0)}(\mathbb{P})}. \tag{17}$$

# 3 Extreme Nonlinear Correlation for Symmetric Functions of iid Random Variables

In this section, we move beyond the pairwise Gaussianality 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 Bryc et al. (2005). In Section 3.2, we consider class of symmetric functions defined on groups of iid random variables and establish the extreme nonlinear correlation in the much broader setting.

## 3.1 Extreme Nonlinear Correlation for Partial 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

$$X_j = S_{m_j} = \sum_{i=1}^{m_j} Y_i \quad \text{for } j = 1, \ldots, p. \tag{18}$$

Here, non-degenerate means that the distribution of the random variable is not concentrated at a 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. (2005) 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 $p$. 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 $(X_1, X_2, \cdots, X_p)$ be nested sums of $Y_i$ with sample sizes $1 \le m_1 \le \cdots \le m_p$ as defined in* (18)*. Then,*

$$\rho^{NL}_{\max}(X_1, \ldots, X_p) = \lambda_{\max}(R), \quad \rho^{NL}_{\min}(X_1, \ldots, X_p) = \lambda_{\min}(R), \tag{19}$$

*where $R = (R_{j,k})_{p\times p}$ is the matrix with elements $R_{jk} = I_{\{j\ne k\}}(m_j \wedge m_k)/\sqrt{m_j m_k}$. If $Y$ has a finite second moment, then $R \in \mathbb{R}^{p\times p}$ is the off-diagonal correlation matrix of the nested sums $X_j = S_{m_j}$, $1 \le j \le p$, so that* (7) *holds with $(X_1, \ldots, X_p)$,*

$$\rho^{NL}_{\max}(X_1, \ldots, X_p) = \rho^{L}_{\max}(X_1, \ldots, X_p), \quad \rho^{NL}_{\min}(X_1, \ldots, X_p) = \rho^{L}_{\min}(X_1, \ldots, X_p).$$

*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 $G_j = \{1, 2, \cdots, m_j\}$ and $\cap_{j=1}^p G_j = G_1 \ne \emptyset$, it follows from Theorem 3 in the next subsection that

$$\rho^{NL}_{\max}(X_1, \ldots, X_p) \le \lambda_{\max}(R), \quad \rho^{NL}_{\min}(X_1, \ldots, X_p) \ge \lambda_{\min}(R).$$

It remains to prove that $\lambda_{\max}(R)$ and $\lambda_{\min}(R)$ are attainable by functions $f_j(X_j)$. This would be simple under the second moment condition on $Y$ as we may simply set $f_j(X_j) = X_j$. In the case of $\mathbb{E}[Y^2] = \infty$, we prove that $R$ is in the closure of the off-diagonal correlation matrices generated by $(f_j(X_j), j \le p)$. This will be done below by proving

$$\lim_{t\to 0+} \rho\big(\sin(tX_j - m_j c_t), \sin(tX_j - m_k c_t)\big) = R_{j,k}, \quad 1 \le j < k \le p, \tag{20}$$

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

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

Note that in our proof below, we need to take the limit in (20) along a subsequence of $t \to 0+$ to avoid $\mathbb{P}\{\sin(tY - c_t) = 0\} = 1$ if the situation arises. This would always be feasible as $\mathbb{P}\{\sin(tY - c_t) = 0\} = 1$ can be achieved only in a countable set of $t$.

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

$$\begin{aligned}\Big|\mathbb{E}\big[\sin(Y')\cos(Y')\big]\Big| &= \Big|\mathbb{E}\big[\sin(Y')(1 - \cos(Y'))\big]\Big| \\ &\le \mathbb{E}[\sin^2(Y')] + \sqrt{\mathbb{E}[\sin^2(Y')]\mathbb{P}\{|Y| > 1/t\}}.\end{aligned} \tag{21}$$

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

$$\lim_{t\to 0+} \rho\big(\sin(S'_{a:m}), \sin(S'_{b:n})\big) = \frac{(m - b + 1)}{(m - a + 1)^{1/2}(n - b + 1)^{1/2}}. \tag{22}$$

This implies (20) with $a = b = 1$, $m = m_j$ and $n = m_k$, but the more general $a$ and $b$ would provide extension to sums of arbitrary subgroups of $Y_i$ later in Corollary 6.

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

$$f_{a,m} = \sum_{u=a}^{m} f_{a,m,u} \quad \text{where} \quad f_{a,m,u} = \Big(\prod_{i=a}^{u-1}\cos(Y'_i)\Big)\sin(Y'_u)\cos(S'_{(u+1):m}).$$

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 $b \le u \wedge v \le u \vee v \le m$,

$$\begin{aligned}
& f_{a,m,u}f_{b,n,v} \\
= \;& \Big(\prod_{i=a}^{u-1}\cos(Y'_i)\Big)\sin(Y'_u)\cos(S'_{(u+1):m})\Big(\prod_{i=b}^{v-1}\cos(Y'_i)\Big)\sin(Y'_v)\cos(S'_{(v+1):n}) \\
= \;& \sin(Y'_{u\wedge v})\cos(Y'_{u\wedge v})\sin(Y'_{u\vee v})g(Y'_i, a \le i \le n, i \ne u \wedge v)
\end{aligned}$$

for a certain function $g$ bounded by 1. Thus, as a consequence of (21)

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

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

$$\begin{aligned}
& \mathbb{E}\big[f_{a,m,u}f_{b,n,u}\big] \\
= \;& \mathbb{E}\big[\sin^2(Y'_u)\big]\mathbb{E}\Big[\Big(\prod_{i=a}^{b-1}\cos(Y'_i)\Big)\Big(\prod_{i=b}^{u-1}\cos^2(Y'_i)\Big)\cos\big(S'_{(u+1):m}\big)\cos\big(S'_{(u+1):n}\big)\Big].
\end{aligned}$$

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 (22) and completes the proof. □

### 3.2 Extreme Nonlinear Correlation for Symmetric Functions of Groups of Variables

In this section, we consider a broader setting than nested sums considered in Section 3.1. 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 of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$.

As $\{\boldsymbol{X}_j\}_{1\le j\le p}$ are vectors, we adjust the definition of the extreme nonlinear correlation in (3) and (4) as follows:

$$\begin{aligned}\rho^{NL}_{\max,\,\text{symm}} &= \rho^{NL}_{\max,\,\text{symm}}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_p) \\ &= \sup_{f_j\in\mathcal{F}_j,1\le j\le p} \lambda_{\max}\Big(\text{Corr}_{\neq}\big(f_1(\boldsymbol{X}_1),\ldots,f_p(\boldsymbol{X}_p)\big)\Big),\end{aligned} \tag{23}$$

where $\mathcal{F}_j = \{f_j : 0 < \text{Var}(f_j(\boldsymbol{X}_j)) < \infty, f_j(y_1,\ldots,y_{m_j}) \text{ symmetric}\}$, and correspondingly

$$\begin{aligned}\rho^{NL}_{\min,\,\text{symm}} &= \rho^{NL}_{\min,\,\text{symm}}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_p) \\ &= \inf_{f_j\in\mathcal{F}_j,1\le j\le p} \lambda_{\min}\Big(\text{Corr}_{\neq}\big(f_1(\boldsymbol{X}_1),\ldots,f_p(\boldsymbol{X}_p)\big)\Big).\end{aligned} \tag{24}$$

Formally, the symmetry of $f_j$ means $f_j(y_1,\ldots,y_{m_j}) = f_j(y_{i_1},\ldots,y_{i_{m_j}})$ for all permutations $i_1,\ldots,i_{m_j}$ of $1,\ldots,m_j$; i.e. symmetry means permutation invariance. 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 $\cap_{j=1}^p G_j \neq \emptyset$, 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 $\emptyset\neq G_1\subset G_2\subset\cdots\subset G_p$, 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 $\rho^{NL}_{\max,\,\text{symm}}$ and $\rho^{NL}_{\min,\,\text{symm}}$ be the extreme symmetric nonlinear correlations of $(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_p)$ as defined in (23) and (24). Let $R^{(\ell)}\in\mathbb{R}^{p\times p}$ be the matrix with elements*

$$R^{(\ell)}_{j,k} = \binom{|G_j\cap G_k|}{\ell}\binom{|G_j|}{\ell}^{-1/2}\binom{|G_k|}{\ell}^{-1/2} I_{\{|G_j\cap G_k|\ge\ell\}} I_{\{j\neq k\}}, \tag{25}$$

*for $1\le\ell\le\ell^*$, where $\ell^* = \max_{1\le j<k\le p}|G_j\cap G_k|$. Then, with $R = R^{(1)}$, we establish*

$$\rho^{NL}_{\max,\text{symm}} = \lambda_{\max}(R), \quad \rho^{NL}_{\min,\,\text{symm}} = \min_{1\le\ell\le\ell^*}\lambda_{\min}\big(R^{(\ell)}\big). \tag{26}$$

*If in addition Assumption C holds, then*

$$\rho^{NL}_{\min,\,\text{symm}} = \lambda_{\min}\big(R\big). \tag{27}$$

The connection between Theorem 2 and Theorem 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 $G_j = \{1, 2, \cdots, m_j\}$, and the corresponding index sets $G_j$ satisfy the 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 a pair of nested sums to a pair of symmetric functions.

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

$$f_j(\boldsymbol{X}_j) = \sum_{i\in G_j} h_0(Y_i) \quad \text{for } 1 \le j \le p, \tag{28}$$

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 (22) 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 second moment condition.

**Corollary 6.** *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,*

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

*under Assumption C. Consequently,* (7) *holds for $X_j = S_{G_j}$ when $\mathbb{E}[Y^2] < \infty$.*

The proof of Theorem 3 relies on the Hoeffding (1948, 1961) decomposition of symmetric functions of random variables, stated as Lemma 3 below; See Lemma 1 in Hoeffding (1961), the decomposition lemma in Efron and Stein (1981), and Lemma 1 in Dembo et al. (2001).

**Lemma 3.** *Let $\boldsymbol{Y} = (Y_1, \cdots, Y_m)$ with iid components $Y_i$ and $f_0(\boldsymbol{Y}) = f_0(Y_1, \cdots, Y_m)$ 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*

$$f_{0,k}(y_1, \ldots, y_k) = \mathbb{E}\left[ f_0(\boldsymbol{Y}) - \sum_{j=1}^{k-1} \sum_{1\le i_1<\cdots<i_j\le m} f_{0,j}(Y_{i_1}, \ldots, Y_{i_j}) \middle| (Y_1, \ldots, Y_k) = (y_1, \ldots, y_k) \right].$$

*Then, the following expansion holds,*

$$f_0(\boldsymbol{Y}) = \sum_{\ell=1}^{m} \sum_{1\le i_1<\cdots<i_\ell\le m} f_{0,\ell}(Y_{i_1}, \ldots, Y_{i_\ell}), \tag{29}$$

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

$$\mathbb{E}\Big[f_{0,\ell}(Y_{i_1}, \ldots, Y_{i_\ell})\Big|\{Y_{i_1}, \ldots, Y_{i_\ell}\}\backslash Y_{i_s}\Big] = 0. \tag{30}$$

*Consequently,*

$$\mathbb{E}\big[f_0^2(\boldsymbol{Y})\big] = \sum_{\ell=1}^{m}\binom{m}{\ell}\mathbb{E}\Big[f_{0,\ell}^2(Y_1, \ldots, Y_\ell)\Big]. \tag{31}$$

*Proof of Theorem 3.* Let $G^{(\ell)} = \{(i_1, \cdots, i_\ell) : i_1 < \cdots < i_\ell, i_s \in G \text{ for } 1 \le s \le \ell\}$ for all subsets of positive integers. Since $f_j(\boldsymbol{X}_j)$ are symmetric functions of $\{Y_i\}_{i \in G_j}$ with finite second moment, (29) gives

$$f_j(\boldsymbol{X}_j) = \sum_{\ell=1}^{m_j}\sum_{(i_1,\cdots,i_\ell)\in G_j^{(\ell)}} f_{j,\ell}(Y_{i_1}, \ldots, Y_{i_\ell}). \tag{32}$$

We first apply (30) 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

$$\begin{aligned}\mathbb{E}f_j(\boldsymbol{X}_j)f_k(\boldsymbol{X}_k) &= \mathbb{E}\sum_{\ell=1}^{|G_j\cap G_k|}\sum_{(i_1,\ldots,i_\ell)\in(G_j\cap G_k)^{(\ell)}} f_{j,\ell}(Y_{i_1}, \ldots, Y_{i_\ell})f_{k,\ell}(Y_{i_1}, \ldots, Y_{i_\ell}) \\ &= \sum_{\ell=1}^{\ell^*}\binom{|G_j \cap G_k|}{\ell}\mathbb{E}\Big[f_{j,\ell}(Y_1, \ldots, Y_\ell)f_{k,\ell}(Y_1, \ldots, Y_\ell)\Big] \qquad (33)\end{aligned}$$

with the convention $\binom{m}{\ell} = 0$ for $\ell > m$. Let $R^{(\ell)} \in \mathbb{R}^{p\times p}$ be the matrix defined in (25). Let $g_{j,\ell} = g_{j,\ell}(Y_1, \ldots, Y_\ell) = \binom{m_j}{\ell}^{1/2}f_{j,\ell}(Y_1, \ldots, Y_\ell)$. By (33), we have

$$\begin{aligned}&\mathbb{E}\left(\sum_{1\le j\neq k\le p} f_j(\boldsymbol{X}_j)f_k(\boldsymbol{X}_k)\right) \\ = \;& \sum_{1\le j\neq k\le p}\sum_{\ell=1}^{\ell^*}\binom{|G_j \cap G_k|}{\ell}E\left[f_{j,\ell}(Y_1, \ldots, Y_\ell)f_{k,\ell}(Y_1, \ldots, Y_\ell)\right] \\ = \;& \sum_{\ell=1}^{\ell^*}\mathbb{E}\left[\sum_{j=1}^{p}\sum_{k=1}^{p}R_{j,k}^{(\ell)}g_{j,\ell}(Y_1, \ldots, Y_\ell)g_{k,\ell}(Y_1, \ldots, Y_\ell)\right] \\ \le \;& \max_{1\le\ell\le\ell^*}\lambda_{\max}\big(R^{(\ell)}\big)\sum_{\ell=1}^{\ell^*}\mathbb{E}\left[\sum_{j=1}^{p}g_{j,\ell}^2(Y_1, \ldots, Y_\ell)\right]\end{aligned}$$

$$
\begin{aligned}
&= \max_{1\le\ell\le\ell^*} \lambda_{\max}\big(R^{(\ell)}\big) \sum_{j=1}^{p}\sum_{\ell=1}^{\ell^*} \mathbb{E}\left[g_{j,\ell}^2(Y_1,\ldots,Y_\ell)\right] \\
&\le \max_{1\le\ell\le\ell^*} \lambda_{\max}\big(R^{(\ell)}\big) \sum_{j=1}^{p} \mathbb{E}\left[f_j^2(X_j)\right], \qquad (34)
\end{aligned}
$$

where the last inequality is true since $\max_{1\le\ell\le\ell^*}\lambda_{\max}\big(R^{(\ell)}\big)\ge 0$ and (31). Regarding the lower bound, we can use a similar argument,

$$
\begin{aligned}
\mathbb{E}\left(\sum_{1\le j\ne k\le p} f_j(\boldsymbol{X}_j)f_k(\boldsymbol{X}_k)\right) &\ge \min_{1\le\ell\le\ell^*} \lambda_{\min}\big(R^{(\ell)}\big) \sum_{\ell=1}^{\ell^*} \mathbb{E}\left[\sum_{j=1}^{p} g_{j,\ell}^2(Y_1,\ldots,Y_\ell)\right] \\
&= \min_{1\le\ell\le\ell^*} \lambda_{\min}\big(R^{(\ell)}\big) \sum_{j=1}^{p}\sum_{\ell=1}^{\ell^*} \mathbb{E}\left[g_{j,\ell}^2(Y_1,\ldots,Y_\ell)\right] \\
&\ge \min_{1\le\ell\le\ell^*} \lambda_{\min}\big(R^{(\ell)}\big) \sum_{j=1}^{p} \mathbb{E}\left[f_j^2(X_j)\right], \qquad (35)
\end{aligned}
$$

where the last inequality is true since $\min_{1\le\ell\le\ell^*}\lambda_{\min}\big(R^{(\ell)}\big)\le 0$ and (31). By Lemma 1, the above inequalities (34) and (35) imply

$$
\rho^{NL}_{\max,\text{symm}} \le \max_{1\le\ell\le\ell^*}\lambda_{\max}\big(R^{(\ell)}\big) \;\text{ and }\; \rho^{NL}_{\min,\text{symm}} \ge \min_{1\le\ell\le\ell^*}\lambda_{\min}\big(R^{(\ell)}\big). \qquad (36)
$$

Hence, it suffices to focus on bounding the limits of the spectrum of $R^{(\ell)}$ for $1\le\ell\le\ell^*$. Note that for the case $|G_j\cap G_k|\ge\ell$,

$$
\begin{aligned}
\frac{\binom{|G_j\cap G_k|}{\ell}}{\binom{|G_j|}{\ell}^{1/2}\binom{|G_k|}{\ell}^{1/2}} &= \frac{|G_j\cap G_k|(|G_j\cap G_k|-1)\cdots(|G_j\cap G_k|-l+1)}{\sqrt{|G_j|(|G_j|-1)\cdots(|G_j|-l+1)}\sqrt{|G_k|(|G_k|-1)\cdots(|G_k|-l+1)}} \\
&\le \frac{|G_j\cap G_k|}{\sqrt{|G_j|\cdot|G_k|}}.
\end{aligned}
\qquad (37)
$$

The above inequality implies that $0\le R^{(\ell)}_{j,k}\le R^{(1)}_{j,k}$ for $1\le\ell\le\ell^*$ and $1\le j,k\le p$. Due to the element-wise positiveness of $R^{(\ell)}$, we have

$$
\lambda_{\max}\big(R^{(\ell)}\big) = \max_{\|\boldsymbol{u}\|_2=1}\sum_{j=1}^{p}\sum_{k=1}^{p} R^{(\ell)}_{j,k}|u_j||u_k| \le \max_{\|\boldsymbol{u}\|_2=1}\sum_{j=1}^{p}\sum_{k=1}^{p} R^{(1)}_{j,k}|u_j||u_k| = \lambda_{\max}\big(R^{(1)}\big), \qquad (38)
$$

where the inequality follows from the fact that $R^{(\ell)}_{j,k}\le R^{(1)}_{j,k}$ for $1\le j,k\le p$. Together with (36), we have

$$
\rho^{NL}_{\max,\text{symm}} \le \lambda_{\max}(R). \qquad (39)
$$

Let $h_0$ be a function satisfying $\mathbb{E}[h_0(Y)] = 0$ and $\mathbb{E}[h_0^2(Y)] = 1$. Define

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

Note that

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

Hence, for $j\ne k$,

$$\mathbb{E}\Big[h_{0,j}^{(\ell)}(\boldsymbol{X}_j)h_{0,k}^{(\ell)}(\boldsymbol{X}_k)\Big] = R_{j,k}^{(\ell)}. \tag{40}$$

It follows that $\rho^{NL}_{\min,\,\mathrm{symm}} \le \lambda_{\min}\big(R^{(\ell)}\big)$ for all $1\le \ell\le \ell^*$. Consequently, together with (36), we establish $\rho^{NL}_{\min,\,\mathrm{symm}} = \min_{1\le\ell\le\ell^*}\lambda_{\min}\big(R^{(\ell)}\big)$. Similarly, the equality in (39) can be achieved by applying (40) with $\ell = 1$. Thus, we have established (26).

The remaining of the proof is to characterize $\min_{1\le\ell\le\ell^*}\lambda_{\min}\big(R^{(\ell)}\big)$ 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*

$$\min_{1\le\ell\le\ell^*}\lambda_{\min}\big(R^{(\ell)}\big) = \lambda_{\min}\big(R\big)$$

*where $R$ is defined in* (25) *and* $\ell^* = \max_{1\le j<k\le p}|G_j\cap G_k|$.

*Proof of Lemma 4.* Under Assumption C, we set

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

Similar to (40), for $j\ne k$

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

It follows that, for the case $|G_j\cap G_k|\ge 1$,

$$R_{j,k}\mathbb{E}\Big[g_{0,j}^{(\ell-1)}(\boldsymbol{X}_j)g_{0,k}^{(\ell-1)}(\boldsymbol{X}_k)\Big] = R_{j,k}^{(\ell)};$$

For the case $|G_j \cap G_k| = 0$, the above equality trivially holds with both sides equal to zero. Hence, we have

$$R_{j,k}\mathbb{E}\Big[g_{0,j}^{(\ell-1)}(\boldsymbol{X}_j)g_{0,k}^{(\ell-1)}(\boldsymbol{X}_k)\Big] = R_{j,k}^{(\ell)},\ \forall 1 \le j,k \le p.$$

As $R$ is an off-diagonal correlation matrix, $\lambda_{\min}(R) \le 0$. It follows that

$$\begin{aligned}
\lambda_{\min}\big(R^{(\ell)}\big) &= \min_{\|\boldsymbol{u}\|_2=1} \sum_{j,k} u_j u_k R_{j,k}\mathbb{E}\Big[g_{0,j}^{(\ell-1)}(\boldsymbol{X}_j)g_{0,k}^{(\ell-1)}(\boldsymbol{X}_k)\Big] \\
&= \min_{\|\boldsymbol{u}\|_2=1} \mathbb{E}\bigg[\sum_{j,k}\Big(u_j g_{0,j}^{(\ell-1)}(\boldsymbol{X}_j)\Big)\Big(u_k g_{0,k}^{(\ell-1)}(\boldsymbol{X}_k)\Big)R_{j,k}\bigg] \\
&\ge \min_{\|\boldsymbol{u}\|_2=1} \lambda_{\min}(R)\mathbb{E}\bigg[\sum_{j}\Big(u_j g_{0,j}^{(\ell-1)}(\boldsymbol{X}_j)\Big)^2\bigg] \\
&\ge \lambda_{\min}(R) \max_{\|\boldsymbol{u}\|_2=1} \mathbb{E}\bigg[\sum_{j}\Big(u_j g_{0,j}^{(\ell-1)}(\boldsymbol{X}_j)\Big)^2\bigg] \\
&\ge \lambda_{\min}(R),
\end{aligned}$$

where the last inequality holds due to the fact that

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

□

# 4 Appendix

We prove Lemmas 1 and 2 in this Appendix.

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

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

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

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

for all $f_{\mathcal{T}} \in \mathcal{F}_{\mathcal{T}}$. Thus, (5) and (12) are equivalent. We omit the proof of the equivalence between (6) and (13) as it can be established by the same argument. □

*Proof of Lemma 2.* As $\rho^L_{\max}$ and $\rho^L_{\min}$ are extreme eigenvalues Let $h$ be a function on $\mathcal{T}$ with $\|h\|_{L_2(\nu)} = 1$. Let $B_n \subseteq B_{n+1}, n \geq 1$, be a sequence of subsets of $\mathcal{T}$ with $\nu(B_n) < \infty$ and $\cup_{n=1}^\infty B_n = \mathcal{T}$. Let $h_{B_n}(t) = h(t) I_{\{t \in B_n\}}$. We have

$$\mathbb{E}\int\int \Big| K(s,t) X_s X_t h_{B_n}(s) h_{B_n}(t) \Big| \nu(ds)\nu(dt) \leq \nu(B_n)\mathbb{E}\int X_t^2 h_{B_n}^2(t)\nu(dt) \leq \nu(B_n) < \infty.$$

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

$$\begin{aligned}
& \int\int |K^m(s,t)||h_{B_n}(s)h_{B_n}(t)|\nu(ds)\nu(dt) \\
\leq & \int\int K^2(s,t)|h_{B_n}(s)h_{B_n}(t)|\nu(ds)\nu(dt) \\
= & \int\int K(s,t)\mathbb{E}\big[X_s X_t\big]|h_{B_n}(s)h_{B_n}(t)|\nu(ds)\nu(dt) \\
= & \mathbb{E}\int\int K(s,t)X_s X_t|h_{B_n}(s)h_{B_n}(t)|\nu(ds)\nu(dt) \\
\leq & \rho^L_{\max}\int \mathbb{E}\big[X_t^2\big]h_B^2(t)\nu(dt) \\
\leq & \rho^L_{\max}\int h^2(t)\nu(dt).
\end{aligned}$$

Thus, by the monotone convergence theorem

$$\lim_{n\to\infty}\left| \int\int K^m(s,t)\Big(h(s)h(t) - h_{B_n}(s)h_{B_n}(t)\Big)\nu(ds)\nu(dt)\right| = 0$$

It follows that

$$\int\int K^m(s,t)h(s)h(t)\nu(ds)\nu(dt) \leq \rho^L_{\max}\int h^2(t)\nu(dt).$$

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

$$\begin{aligned}
& \int\int K^m(s,t)h_{B_n}(s)h_{B_n}(t)\nu(ds)\nu(dt) \\
= & \mathbb{E}\int\int K(s,t)\Big\{\prod_{i=1}^{m-1} X_s^{(i)}h_{B_n}(s)\Big\}\Big\{\prod_{i=1}^{m-1} X_t^{(i)}h_{B_n}(t)\Big\}\nu(ds)\nu(dt) \\
\geq & \rho^L_{\min}\int \mathbb{E}\Big\{\prod_{i=1}^{m-1} X_t^{(i)}h_{B_n}(t)\Big\}^2\nu(dt) \\
= & \rho^L_{\min}\int h_{B_n}^2(t)\nu(dt),
\end{aligned}$$

where $\{X_t^{(i)}, t \in \mathcal{T}\}$ are iid copies of $X_{\mathcal{T}}$. The conclusion follows by letting $n \to \infty$. □

## References

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Wlodzimierz Bryc, Amir Dembo, and Abram Kagan. On the maximum correlation coefficient. *Theory of Probability & Its Applications*, 49(1):132–138, 2005.

Andreas Buja, Trevor Hastie, and Robert Tibshirani. Linear smoothers and additive models. *The Annals of Statistics*, 17(2):453–510, 1989.

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

Bradley Efron and Charles Stein. The jackknife estimate of variance. *The Annals of Statistics*, pages 586–596, 1981.

Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statistical Science*, 1 (3):297–318, 1986.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, pages 293–325, 1948.

Wassily Hoeffding. The strong law of large numbers for u-statistics. Technical report, North Carolina State University. Dept. of Statistics, 1961.

Vladimir Koltchinskii and Ming Yuan. Sparsity in multiple kernel learning. *The Annals of Statistics*, 38(6):3660–3695, 2010.

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

Lukas Meier, Sara Van de Geer, and Peter Bühlmann. High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779–3821, 2009.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 13(Feb):389–427, 2012.

Taiji Suzuki and Masashi Sugiyama. Fast learning rate of multiple kernel learning: Trade-off between sparsity and smoothness. *The Annals of Statistics*, pages 1381–1405, 2013.

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

Sara A van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

Simon N Wood. *Generalized additive models: an introduction with R*. Chapman and Hall/CRC, 2017.

Yaming Yu. On the maximal correlation coefficient. *Statistics & Probability Letters*, 78(9): 1072–1075, 2008.