Statistica Sinica Preprint No: SS-2025-0370	
Title	On Measure of Association for Heavy-tailed Random Variables
Manuscript ID	SS-2025-0370
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0370
Complete List of Authors	Vahed Maroufy, Mohsen Rezapour, Mahmoud Zarepour and Bahareh Afhami
Corresponding Authors	Vahed Maroufy
E-mails	vahed.maroufy@uth.tmc.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

Statistica Sinica

ON MEASURE OF ASSOCIATION FOR HEAVY-TAILED RANDOM VARIABLES

Vahed Maroufy[1], Mohsen Rezapour[1], Mahmoud Zarepour[2], and Bahareh Afhami[3]

[1] *Dept. of Biostatistics and Data Science, School of Public Health, University of Texas Health, Houston, Texas, U.S.A*

[2] *Dept. of Mathematics and Statistics, University of Ottawa, Canada*

[3] *Dept. of Statistics, Shahid Bahonar University, Kerman, Iran*

Abstract: Heavy-tailed high-dimensional data are common in practice, including disease prediction, gene expression analysis, and risk management. Dependence measures for symmetric α-stable (SαS) random vectors are commonly applied to heavy-tailed data. However, the existing measures of dependence for symmetric α-stable (SαS) random vectors do not imply independence at zero. To address this problem, we introduce a novel measure of dependence, *extended codifference*, for the SαS and non-symmetric α-stable heavy-tailed distribution family, that allows characterizing independence between heavy-tailed variables for $0 < \alpha < 2$. We propose an efficient non-parametric estimator that does not require estimation of tail indices and obtain its asymptotic distribution. Furthermore, we provide a guideline for selecting a suitable measure of dependence based on the properties of each measure of association. Finally, we provide several simulation studies for further illustration and apply extended codifference to clustering of single-cells based on their RNA-seq expression to identify cell types in adipose tissue.

Keywords: Measure of dependence, stable random variables, infinite variance, codifference.

Key words and phrases: Measure of dependence, infinite variance, codifference.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

1. Introduction

Heavy-tailed data vectors commonly appear in statistical data analyses, such as analysis of extreme values (Rohrbeck and Cooley, 2021), prediction modeling using big Electronic Health Records (EHR), and gene expression analysis using RNA-Seq data (Cantoni and Ronchetti, 2006; Zhu et al., 2019). For example, an annual extremal network for exploring the extremal dependence structure of environmental processes was proposed in Huang et al. (2019). Moreover, in a case-control study using EHR data, the prediction of specific outcomes of a rare disease may involve heavy-tailed biomarkers and patient features, including numerical lab results and vital signs. Also, in gene differential analysis, expression values as unique representatives of the biological state of a particular tissue or cell, can be noisy (Garel and Kodia, 2009; Kuznetsov et al., 2002) with a heavy-tailed distribution, of power-like (Pareto) form (Tsimring, 2014; Newman, 2005). A recent groundbreaking technology, measuring gene expression at the level of individual cells, is single-cell RNA-seq (scRNA-seq), which allows biologists to study heterogeneity of expression and detect unknown cell types (Hashimshony et al., 2012; Jaitin et al., 2014). This process produces high-dimensional, sparse numerical data matrices of expression values of thousands of genes for thousands of dissected individual cells. Clustering cells based on their gene expression patterns using these types of data is still a challenge under study by many researchers.

These studies often rely heavily on variable selection and classification methods for which measuring the dependence and association between the biomarkers and response (Fan and Lv, 2008) and correlation between genes' expression values are required

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

(Oyelade et al., 2016). However, heavy-tailed distributions often do not have finite second moments, and hence the variance-covariance matrix and standard measures of correlation, such as Pearson correlation, do not exist. Therefore, proper measures of dependence for heavy-tailed distributions are required to screen the biomarkers with significant association with the outcome of interest and detect clusters embedded in the feature subspace.

A standard family of distributions for modeling heavy-tailed data is the family of stable distributions. Stable distributions are invariant under linear transformation, have the Gaussian distribution as a particular case, and are capable of modeling heavy tails as well as skewness. When the central limit theorem is generalized to i.i.d random variables with infinite variance, the limiting distribution is stable (Feller, 1971). Stable distributions have been applied to different fields of study, including finance and pattern recognition (Fiche et al., 2013; Pele and Stanciulescu, 2015; Żak et al., 2017).

In the context of stable distributions, the classical covariance is not defined when the stability parameter $\alpha < 2$, because moments of order two or higher are infinite. To overcome this limitation, codifference was introduced as a robust alternative measure of dependence Samorodnitsky and Taqqu (1994). Codifference has several appealing properties: it reduces to covariance when $\alpha = 2$ (Gaussian case), and it remains well-defined for all $0 < \alpha < 2$. In addition, for $0 < \alpha < 1$, codifference equal to zero implies independence between random variables, making it a powerful tool in heavy-tailed data analysis. However, this independence characteristic does not generalize to $1 \leq \alpha < 2$.

Motivated by this limitation, this paper proposes a novel generalized dependence

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

measure that extends the codifference, ensuring that zero dependence implies independence for the entire range $0 < \alpha < 2$. This extension broadens the applicability of dependence measures for heavy-tailed data and provides a stronger theoretical foundation for modeling and inference in stable random vectors.

Among the measures of association for heavy-tailed random variables, the choice of a proper measure for an application depends on the purpose of the study. For example, testing independence between two variables on the one hand and evaluating the positive or negative direction of the dependence on the other hand may require two different measures, as some measures of association do not guarantee independence at value zero, although they can quantify associations. In addition, properties such as interchangeability and symmetry play crucial roles in determining the best choice of measure in applications. In this paper, we study the properties of the existing and novel measures of associations for SαS random vectors and provide a guideline for selecting a suitable measure for particular applications. Moreover, we study the asymptotic behavior of nonparametric estimators for some measures of association for heavy-tailed random variables. We also propose a novel extended codifference measure that implies independence at zero for $0 < \alpha < 2$, and applied it to test the hypothesis of independence of the components of a symmetric α-stable random vector via a simulation study (Definition 7).

The remainder of this paper is structured as follows. In Section 2, after briefly reviewing some measures of dependence for families of stable distributions, including the classical codifference definition, we define the extended codifference and prove that

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

zero value of this measure implies independence for all $0 < \alpha < 2$ (Definition 7 and Theorem 1). We also study the properties of different measures of dependence and provide a guideline for selecting a suitable measure of dependence to quantify positive or negative dependence. In Section 3, we propose a nonparametric estimator for extended codifference and signed symmetric covariation and study their asymptotic behaviors. We also illustrate and compare hypothesis testing of independence between two SαS random vectors for different measures of dependence. In Section 4, we present several simulation studies for more illustration and apply the codifference measure together with the spectral clustering method for single-cell clustering of adipose tissue, based on RNA-Seq expression values. Finally, we close the paper with a brief discussion section and give proofs in the appendix.

2. Measures of dependence for a SαS random vector

In this section, we introduce the class of spectral covariance functions, which includes many of the measures of association, and then review existing measures of dependence for a bivariate SαS random vector. We also introduce an extended measure of association that characterizes the independence of two SαS random variables with $0 < \alpha < 2$.

2.1 Overview of Measures of dependence for a SαS Random Vector

A distribution on $\mathbb{R}$ is stable if a linear combination of two independent random variables with this distribution has the same distribution, up to location and scale parameters. Stability parameter $\alpha \in (0, 2]$ measures tail thickness $\alpha = 2$, $\alpha = 1$, and $\alpha = 0.5$ corre-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

spond to the Gaussian, Cauchy and Lévy distribution, respectively. Stable distributions have infinite variance for $\alpha < 2$, and infinite mean for $\alpha \leq 1$. Most of the stable laws do not have a tractable closed-form density function. Hence, this class is characterized by its characteristic function. The following two definitions characterize SαS random vectors, and the class of spectral measures.

Definition 1. $\mathbf{X} := (X_1, \ldots, X_d)$ *is a SαS random vector in* $\mathbb{R}^d$ *with* $0 < \alpha < 2$ *(denoted by* $\mathbf{X} \in \mathcal{S}_\alpha$*), if and only if, there exists a unique symmetric finite measure* Γ *on the unit sphere* $SP_d = \{\mathbf{x} : \|\mathbf{x}\| = 1\,, \mathbf{x} \in \mathbb{R}^d\}$*, called spectral measure, such that*

$$\Phi_{\mathbf{X}}(\boldsymbol{\theta}) := \mathbb{E} \exp\{i\langle \boldsymbol{\theta}, \mathbf{X}\rangle\} = \exp\left\{-\int_{SP_d} |\langle \boldsymbol{\theta}, \mathbf{s}\rangle|^\alpha \, \Gamma(d\mathbf{s})\right\}. \tag{2.1}$$

where $\langle \boldsymbol{\theta}, \mathbf{s}\rangle$ *is Euclidean inner product defined on* $\mathbb{R}^d$*.*

Remark 1. *Expression* (2.1) *also applies in the Gaussian case when* $\alpha = 2$*; however, in this case, the spectral measure* Γ *is not uniquely defined. This point is illustrated by an example in Samorodnitsky and Taqqu (1994, p. 76).*

Definition 2. *(Damarackas and Paulauskas (2017)) Let* $f : SP_2 \to \mathbb{R}$ *be a function such that* $\int_{SP_2} f(s_1, s_2)\, \Gamma(d\mathbf{s})$ *is well-defined, and the following conditions are satisfied:*

$$f(s_1, s_2) = 0 \;\; \textit{if} \;\; s_1\, s_2 = 0\,, \quad f(s_1, s_2) = f(s_2, s_1) \;\; \textit{and} \;\; f(s_1, s_2) = f(-s_1, -s_2)\,. \tag{2.2}$$

Then, a class of spectral covariance functions is defined by

$$\rho(f; X_1, X_2) = \int_{SP_2} f(s_1, s_2)\, \Gamma(d\mathbf{s})\,. \tag{2.3}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

For a bivariate SαS random vector $X = (X_1, X_2)$ with $\alpha > 1$, the spectral covariance is defined by

$$\rho(X_1, X_2) = \int_{SP_2} s_1\, s_2\, \Gamma(d\mathbf{s})\,, \tag{2.4}$$

where Γ as in Definition 1. Analogous to the Pearson correlation coefficient, the spectral correlation coefficient (s.c.c.) for an α-stable random vector $\mathbf{X}$ is defined as,

$$\tilde{\rho}(X_1, X_2) = \int_{SP_2} s_1\, s_2\, \Gamma(d\mathbf{s}) \left(\int_{SP_2} s_1^2\, \Gamma(d\mathbf{s}) \int_{SP_2} s_2^2\, \Gamma(d\mathbf{s}) \right)^{-\frac{1}{2}}.$$

While the independence of two SαS random variables X_1 and X_2 guarantees that $\rho(X_1, X_2)$ is zero, the converse is not generally true.

Another measure of dependence is α-spectral covariance, defined by

$$\rho_\alpha(X_1, X_2) = \int_{SP_2} s_1^{<\alpha/2>}\, s_2^{<\alpha/2>}\, \Gamma(d\mathbf{s})\,, \tag{2.5}$$

where $x^{<a>} = \text{sign}(a)|x|^a$, for real constants a and $x \in \mathbb{R}$.

Definition 3. *Let (X_1, X_2) be SαS random vector with $1 < \alpha \leq 2$ and let Γ be the spectral measure as in Definition 1, the covariation of X_1 on X_2 is the real number*

$$[X_1, X_2]_\alpha = \int_{SP_2} s_1\, s_2^{<\alpha-1>}\, \Gamma(d\mathbf{s})\,, \tag{2.6}$$

(see Samorodnitsky and Taqqu (1994) for more details).

This measure is not symmetric; i.e., $[X_1, X_2]_\alpha \neq [X_2, X_1]_\alpha$, and it can be shown that

$$[X_1, X_2]_2 = \frac{1}{2}\,\text{Cov}(X_1, X_2)\,.$$

Moreover, if X_1 and X_2 are jointly SαS and independent, then $[X_1, X_2]_\alpha = 0$. However, the reverse does not hold in general. Furthermore, the covariation measure characterizes

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

the covariation norm, as shown in the following definition, which is an important norm function in defining certain measurements of dependence.

Definition 4. *The covariation norm of* $X \in \mathcal{S}_\alpha$, $0 < \alpha \leq 2$, *is defined by*

$$\|X\|_\alpha = ([X, X]_\alpha)^{\frac{1}{\alpha}} = (-\ln(\mathbb{E}\exp(iX)))^{\frac{1}{\alpha}} . \tag{2.7}$$

One disadvantage of all the measures of dependence reviewed above is that they do not generally imply independence at zero, except in certain restricted cases. For example, codifference given in (2.9), guarantees independence of two SαS random variables only for $0 < \alpha < 1$.

Another measure of association, similar to the *signs of correlation coefficient*, introduced in Garel and Kodia (2009) and revised by Kodia and Garel (2014), is *signed symmetric covariation (SCOV)* given in the following definition.

Definition 5. *Let* (X_1, X_2) *be a bivariate SαS random vector with stability index* $\alpha > 1$. *The signed symmetric covariation coefficient between* X_1 *and* X_2 *is defined as:*

$$scov(X_1, X_2) = \kappa(X_1, X_2) \left| \frac{[X_1, X_2]_\alpha \; [X_2, X_1]_\alpha}{\|X_1\|_\alpha^\alpha \; \|X_2\|_\alpha^\alpha} \right|^{\frac{1}{2}} \quad \textit{where} \tag{2.8}$$

$$\kappa(X_1, X_2) = \begin{cases} \operatorname{sign}([X_1, X_2]_\alpha) , & \textit{if } \operatorname{sign}([X_1, X_2]_\alpha) = \operatorname{sign}([X_2, X_1]_\alpha) , \\ -1 & \textit{otherwise.} \end{cases}$$

Codifference (Samorodnitsky and Taqqu, 1994) of two SαS random variables X_1 and X_2 is defined by

$$\tau_{X_1, X_2} = \|X_1\|_\alpha^\alpha + \|X_2\|_\alpha^\alpha - \|X_1 - X_2\|_\alpha^\alpha . \tag{2.9}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

By Property 2.10.4 in Samorodnitsky and Taqqu (1994), if X_1 and X_2 , are independent then $\tau_{X_1,X_2} = 0$, but the reverse holds only for $0 < \alpha < 1$.

If γ is a measure of dependence (including any measure discussed earlier), then the following definition outlines the desirable properties that such a measure should satisfy.

Definition 6. *Consider the following properties of the measure of dependence γ:*

i) $\gamma_{(X_1,X_2)} = 0$ *if and only if* X_1 *and* X_2 *are independent.*

ii) $\gamma_{(X_1,X_2)}$ *reduces with the classical covariance when* $\alpha = 2$.

iii) $\gamma_{(X_1,X_2)}$ *determines the direction of dependence when* X_1 *and* X_2 *are dependent.*

iv) If the dependence between X_1 *and* X_2 *is positive, then* $\gamma_{(X_1,X_2)} > 0$*; if the dependence is negative, then* $\gamma_{(X_1,X_2)} < 0$.

v) $\gamma_{(X_1,X_2)}$ *is invariant under sign changes; that is,* $\gamma_{(X_1,X_2)} = \gamma_{(-X_1,-X_2)}$.

vi) $\gamma_{(X_1,X_2)}$ *is symmetric; that is,* $\gamma_{(X_1,X_2)} = \gamma_{(X_2,X_1)}$.

Property (i) extends the equivalence between zero covariance and independence to SαS random vectors, similar to that of the Gaussian distribution ($\alpha = 2$). Property (ii) determines if $\gamma_{\mathbf{X}}$ is a natural extension of the regular covariance to the family of SαS distributions. We include property (iii) here since, unlike covariance, not all measures described above can determine the direction of dependence. The following theorem characterizes the properties of each of the five measures of dependence based on Definition 6.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

2.2 A New Measure of Dependence for a SαS Random Vector and Its Properties

Motivated by Definition 5, and building on codifference measure, we introduce an extended codifference that guarantees independence at zero for all $0 < \alpha < 2$.

Definition 7. *Extended codifference for SαS random vector* (X_1, X_2) *is defined by*

$$\tau^*_{X_1,X_2} = \|X_1\|^\alpha_\alpha + \|X_2\|^\alpha_\alpha - \frac{1}{2}\left(\|X_1 - X_2\|^\alpha_\alpha + \|X_1 + X_2\|^\alpha_\alpha\right). \tag{2.10}$$

Note that using (2.9), we can rewrite the extended codifference in terms of the characteristic function as follows

$$\tau^*_{X_1,X_2} = \ln\left(\frac{\left(\mathbb{E}\exp\left(i(X_1 - X_2)\right)\mathbb{E}\exp\left(i(X_1 + X_2)\right)\right)^{\frac{1}{2}}}{\mathbb{E}\exp\left(iX_1\right)\mathbb{E}\exp\left(iX_2\right)}\right). \tag{2.11}$$

The following result characterizes the independence of two SαS random variables using the extended codifference for $0 < \alpha < 2$.

Theorem 1. *A bivariate SαS random vector* (X_1, X_2), *with* $0 < \alpha < 2$, *has independent components if and only if* $\tau^*_{X_1,X_2} = 0$.

By definition, when $\alpha = 2$, the extended codifference can be rewritten as

$$\tau^*_{X_1,X_2} = -\ln e^{-\frac{1}{2}\sigma_1^2} - \ln e^{-\frac{1}{2}\sigma_2^2} - \frac{1}{2}\left(-\ln e^{-\frac{1}{2}(\sigma_1^2+\sigma_2^2-2\sigma_{12})} - \ln e^{-\frac{1}{2}(\sigma_1^2+\sigma_2^2+2\sigma_{12})}\right),$$

which is reduced to zero. In this situation, we can define a modified version of extended codifference for $p \in [0, 1]$, as

$$\tilde{\tau}_{X_1,X_2} = \|X_1\|^\alpha_\alpha + \|X_2\|^\alpha_\alpha - \left(p\,\|X_1 - X_2\|^\alpha_\alpha + (1-p)\,\|X_1 + X_2\|^\alpha_\alpha\right).$$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

It is straightforward to see that, for $p = \frac{1}{2}$, the $\tilde{\tau}_{X_1,X_2}$ coincides with the extended codifference, namely $\tau^*_{X_1,X_2}$, while for $p = 1$ it reduces to the codifference, namely τ_{X_1,X_2}. Moreover, when $\alpha = 2$, $\tilde{\tau}_{X_1,X_2}$ reduces to

$$-\ln e^{-\frac{1}{2}\sigma_1^2} - \ln e^{-\frac{1}{2}\sigma_2^2} - \left(-p \ln e^{-\frac{1}{2}(\sigma_1^2+\sigma_2^2-2\sigma_{12})} - (1-p) \ln e^{-\frac{1}{2}(\sigma_1^2+\sigma_2^2+2\sigma_{12})}\right),$$

which simplifies to $(2\,p-1)$ times the covariance. Therefore, for α close to 2, we may use $\tilde{\tau}_{X_1,X_2}$ with some $p \in (0,1) \setminus \{\frac{1}{2}\}$. Therefore, we can conclude the following result.

Remark 2. A bivariate SαS random vector (X_1, X_2), with $0 < \alpha \leq 2$, and for $p \in (0,1) \setminus \{\frac{1}{2}\}$, has independent components if and only if $\tilde{\tau}_{X_1,X_2} = 0$.

Here, we study some properties of measures of dependence introduced earlier and provide a guideline for selecting a suitable measure in applications, depending on the purpose of a study. Mainly, we show that the extended codifference should be employed to prove independence between two SαS random variables, while we suggest Spectral Covariance and SCOV (Garel and Kodia, 2009) to determine the direction and strength of dependence. We use the following definition to characterize each measure, where $\mathbf{X} = (X_1, X_2)$ and $\gamma_{\mathbf{X}}$ is the measure of dependence between X_1 and X_2.

Theorem 2. *For a bivariate SαS random vector with spectral measure* Γ,

a) spectral covariance (2.4) *satisfies (iii)-(v), but it does not satisfy (i) and (ii).*

b) α-spectral covariance (2.5) *satisfies (iii)-(v), but it does not satisfy (i) and (ii).*

c) for $1 < \alpha \leq 2$ covariation (2.6) *satisfies (iii), but it does not satisfy (i), (ii), (iv) and (v).*

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

d) for $0 < \alpha < 2$ *extended codifference* (2.10) *satisfies all properties except (ii) and (iii).*

e) for $0 < \alpha < 1$ *codifference* (2.9) *satisfies all properties except for (iii). Also, (i) does not hold for* $1 < \alpha < 2$.

Theorem 2 recommends extended codifference to show independence between two SαS random variables and recommends spectral covariance and α-spectral covariance to determine the positive or negative direction of dependence.

3. Nonparametric estimation of measures of dependence

This section proposes novel nonparametric estimators for codifference and extended codifference and an M-estimator for spectral covariance. We also study the asymptotic behavior of each estimator.

3.1 Codifference estimator and its asymptotic behavior.

For the i.i.d. random sample $(X_{11}, X_{21}), \ldots, (X_{1n}, X_{2n})$, we suggest the following nonparametric estimators for τ_{X_1,X_2} Eq. (2.9) and $\tau^*_{X_1,X_2}$ Eq. (2.11), respectively,

$$\widehat{\tau}_{X_1,X_2} = \ln\left(\frac{\frac{1}{n}\sum_{l=1}^{n}\cos(X_{1l}-X_{2l})}{\frac{1}{n^2}\sum_{l,k=1}^{n}\cos(X_{1l})\cos(X_{2k})}\right), \quad \text{and} \tag{3.1}$$

$$\widehat{\tau}^*_{X_1,X_2} = \frac{1}{2}\ln\left(\frac{\frac{1}{n^2}\sum_{l,k=1}^{n}\cos(X_{1l}-X_{2l})\cos(X_{1k}+X_{2k})}{\left(\frac{1}{n^2}\sum_{l,k=1}^{n}\cos(X_{1l})\cos(X_{2k})\right)^2}\right). \tag{3.2}$$

We can also estimate the dispersion of a SαS random sample $X_1, \ldots, X_n$, using

$$\widehat{\tau}_{X,X} = -2\ln\left(\frac{1}{n}\sum_{j=1}^{n}\cos(X_j)\right) \quad \text{and}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

$$\widehat{\tau}^{*}_{X,X} = \frac{1}{2}\ln\left(\frac{\frac{1}{n}\sum_{l=1}^{n}\cos(2X_l)}{\left(\frac{1}{n}\sum_{l=1}^{n}\cos(X_l)\right)^4}\right).$$

Theorem 3 derives the asymptotic distributions for $\widehat{\tau}_{X,X}$ and $\widehat{\tau}^{*}_{X,X}$.

Theorem 3. *Suppose that $X_1, \ldots, X_n$ are i.i.d. SαS random variables, then*

$$\frac{\sqrt{n}}{\sigma_{1;\tau}}(\hat{\tau}_{X,X} - \tau_{X,X}) \xrightarrow{d} Z_1\,, \qquad 0 < \alpha \leq 2 \quad \textit{and} \tag{3.3}$$

$$\frac{\sqrt{n}}{\sigma_{2;\tau}}(\hat{\tau}^{*}_{X,X} - \tau^{*}_{X,X}) \xrightarrow{d} Z_2\,, \qquad 0 < \alpha < 2 \tag{3.4}$$

standard Gaussian random variables Z_1 and Z_2, and $\sigma_{i;\tau}$ $i = 1, 2$ are given in (**??**).

We estimate the off-diagonal components of the codifference and extended codifference matrix by using (3.1) and (3.2), respectively, and characterize their asymptotic distribution in the following theorem.

Theorem 4. *Suppose that $(X_{11}, X_{21})^\top, \ldots, (X_{1n}, X_{2n})^\top$ are i.i.d. SαS random vectors with spectral measure Γ, then*

$$\frac{\sqrt{n}}{\sigma_{3;\tau}}(\hat{\tau}_{X_1,X_2} - \tau_{X_1,X_2}) \xrightarrow{d} Z_1\,, \qquad 0 < \alpha \leq 2 \quad \textit{and} \tag{3.5}$$

$$\frac{\sqrt{n}}{\sigma_{4;\tau}}(\hat{\tau}^{*}_{X_1,X_2} - \tau^{*}_{X_1,X_2}) \xrightarrow{d} Z_2\,, \qquad 0 < \alpha < 2 \tag{3.6}$$

standard Gaussian random variables Z_1 and Z_2, and $\sigma_{3;\tau}$ and $\sigma_{4;\tau}$ are given in (**??**).

3.2 Codifference for non-symmetric α-stable random variables

In this subsection, we consider the measure of association for non-symmetric α-stable random variables with $0 < \alpha < 2$.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

An α-stable random vector $\mathbf{X}$ with $\alpha \neq 1$ has the characteristic function of form

$$\Psi_{\mathbf{X}}(\boldsymbol{\theta}) := \exp\left\{-\int_{SP_d} |\langle \boldsymbol{\theta}, \mathbf{s}\rangle|^{\alpha} \left(1 - i\ \beta \text{sign}(\langle \boldsymbol{\theta}, \mathbf{s}\rangle) \tan\frac{\pi\alpha}{2}\right) \Gamma(\mathbf{s})\right\}. \tag{3.7}$$

When $\alpha \neq 1$, using Example 2.3.6 in Samorodnitsky and Taqqu (1994), $\mathbf{X}$ has independent components if and only if $s_1\, s_2 = 0$, Γ-a.e. (almost everywhere with respect to Γ) and $\text{sign}(s_1 s_2) s_1 s_2 = 0$, Γ-a.e.. For a non-symmetric bivariate α-stable random vector, we can define codifference by the same formula as in (2.9), with a modification, i.e., we let $\|\mathbf{X}\|_{\alpha}^{\alpha} = \int_{SP_2} |(\boldsymbol{\theta}, \mathbf{s})|^{\alpha} \Gamma(\mathbf{s})$ and substitute $\|\mathbf{X}\|_{\alpha}^{\alpha}$ with $\|\mathbf{X}\|_{\alpha}^{\alpha} = -\ln \|\Psi_{\mathbf{X}}(\mathbf{1})\|$, where $\|\cdot\|$ is module of a complex number. We can estimate the covariation norm by using the following estimator,

$$\begin{aligned} \widehat{\|\mathbf{X}\|_{\alpha}^{\alpha}} &= -\frac{1}{2}\ln\left(\left(\frac{1}{n}\sum_{k=1}^{n}\cos(X_{1k})\right)^2 + \left(\frac{1}{n}\sum_{k=1}^{n}\sin(X_{1k})\right)^2\right) \\ &= -\frac{1}{2}\ln\left(\left(\frac{1}{n^2}\sum_{k,l=1}^{n}\cos(X_{1k} - X_{1l})\right)\right). \end{aligned} \tag{3.8}$$

With a similar argument as in Theorem 1 for the case $0 < \alpha < 2$, we can show

$$\int_{SP_2} \left(|s_1|^{\alpha} + |s_2|^{\alpha} - \frac{1}{2}\left(|s_1 - s_2|^{\alpha} + |s_1 + s_2|^{\alpha}\right)\right)\Gamma(\mathbf{s}) = 0 \tag{3.9}$$

holds if and only if $\text{sign}(s_1 s_2) s_1 s_2 = 0$, Γ-a.e.. Therefore, the bivariate α-stable random vector $\mathbf{X}$ has independent components if and only if equation (3.9) holds. For i.i.d. samples $\mathbf{X}_1 = (X_{11}, \ldots, X_{1n})$ and $\mathbf{X}_2 = (X_{21}, \ldots, X_{2n})$ from non-symmetric α-stable distributions, using (3.8) we can estimate the association between $\mathbf{X}_1$ and $\mathbf{X}_2$ by

$$\hat{\tau}^{*}_{\mathbf{X}_1,\mathbf{X}_2} = \ln\left(\frac{\left(\frac{1}{n^4}\sum_{k_1,l_1,k_2,l_2=1}^{n}\cos(\tilde{\mathbf{X}}_{k_1,l_1})\cos(\tilde{\mathbf{X}}_{k_2,l_2})\right)^{\frac{1}{2}}}{\frac{1}{n^2}\sum_{k,l=1}^{n}\cos(X_{1k} - X_{1l})\frac{1}{n^2}\sum_{k,l=1}^{n}\cos(X_{2k} - X_{2l})}\right),$$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

where $\tilde{\mathbf{X}}_{k,l} = X_{1k} - X_{1l} - (X_{2k} - X_{2l})$. Also, for sufficiently large sample size, by the strong law of large numbers, $\hat{\tau}^*_{\mathbf{X}_1,\mathbf{X}_2}$ is close to zero if, and only if, $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent. Hence for non-symmetric α-stable random vectors, we generate the random samples $(0, X_{11} - X_{12}, \ldots, X_{11} - X_{1n}, \ldots, X_{1n-1} - X_{1n}, 0)$ and $(0, X_{21} - X_{22}, \ldots, X_{21} - X_{2n}, \ldots, X_{2n-1} - X_{2n}, 0)$, and use codifference defined in (2.9) as dependence measure.

3.3 Codifference matrix

Suppose that $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^\top$, $i = 1, \ldots, n$, and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ is the observation matrix of i.i.d. random SαS random vectors. We define the codifference and extended codifference matrices by $\boldsymbol{\tau}_{\mathbf{X}} = \left(\tau_{X_i,X_j}\right)_{i,j=1,\ldots,p}$, and $\boldsymbol{\tau}^*_{\mathbf{X}} = \left(\tau^*_{X_i,X_j}\right)_{i,j=1,\ldots,p}$, and estimate them by $\widehat{\boldsymbol{\tau}}_{\mathbf{X}}$ and $\widehat{\boldsymbol{\tau}}^*_{\mathbf{X}}$, where

$$\widehat{\boldsymbol{\tau}}_{\mathbf{X}} = \left(\widehat{\tau}_{X_i,X_j}\right)_{i,j=1,\ldots,p}, \quad \widehat{\boldsymbol{\tau}}^*_{\mathbf{X}} = \left(\widehat{\tau}^*_{X_i,X_j}\right)_{i,j=1,\ldots,p}. \tag{3.10}$$

Let $\vartheta_{ij} = \tau_{X_i,X_j}(\vartheta^*_{ij} = \tau^*_{X_i,X_j})$, $\vartheta_1 = (\vartheta_{11}, \ldots, \vartheta_{pp})$, the diagonal elements of $\tau_{\mathbf{X}}$, $\boldsymbol{\vartheta}_2 = (\vartheta_{12}, \ldots, \vartheta_{1p}, \theta_{23}, \ldots, \vartheta_{2p}, \ldots, \vartheta_{p-1,p})$, the upper triangular components of $\tau_{\mathbf{X}}$, and $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2)$. Also define $\boldsymbol{\vartheta}^*$ similarly. The following theorem characterizes the asymptotic distributions of $\widehat{\boldsymbol{\vartheta}}_{\mathbf{X}}$ and $\widehat{\boldsymbol{\vartheta}}^*_{\mathbf{X}}$.

Theorem 5. *Suppose* $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ *is the observation matrix of sequence of i.i.d. SαS random vectors with characteristic function* $\Phi_{\mathbf{X}}$ *given in* (2.1)*. Then,*

$$\sqrt{n}\left(\widehat{\boldsymbol{\vartheta}}_{\mathbf{X}} - \boldsymbol{\vartheta}_{\mathbf{X}}\right) \xrightarrow{d} \mathcal{N}_{\frac{1}{2}(p^2+p)}(\mathbf{0}, \nabla g_1 \boldsymbol{\Sigma}_{1;\vartheta} (\nabla g_1)^\top) \quad \textit{and} \tag{3.11}$$

$$\sqrt{n}\left(\widehat{\boldsymbol{\vartheta}}^*_{\mathbf{X}} - \boldsymbol{\vartheta}^*_{\mathbf{X}}\right) \xrightarrow{d} \mathcal{N}_{\frac{1}{2}(p^2+p)}(\mathbf{0}, \nabla g_2 \boldsymbol{\Sigma}_{2;\vartheta} (\nabla g_2)^\top), \tag{3.12}$$

where $\boldsymbol{\Sigma}_{1;\vartheta}$, ∇g_1, $\boldsymbol{\Sigma}_{2;\vartheta}$, *and* ∇g_2 *are given in* (??), (??), (??), *and* (??).

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

Remark 3. These asymptotic distributions are quite complicated to use in applications. Instead, one can show that the bootstrap estimators are consistent since the root-n consistency with the Gaussian limit holds (see Singh (1981)). Therefore, we suggest using bootstrapping to construct confidence regions or to test hypotheses. We should remind the reader that the regular bootstrap is not asymptotically valid when limits are stable laws (see Knight (1989)).

3.4 Estimator for signed symmetric covariation

Signed symmetric covariation coefficient (Definition 5) for bivariate SαS random vector $\mathbf{X}$, is a standardized measure of dependence for which Kodia and Garel (2014) proposed a nonparametric estimator and showed that it is asymptotically consistent. However, the asymptotic distribution of this estimator can not be obtained explicitly. Instead, we propose an M-estimator for *scov* and study its asymptotic behavior. For this purpose, we need to estimate $\beta_1 := \frac{[X_1,X_2]_\alpha}{\|X_2\|_\alpha^\alpha}$ and $\beta_2 := \frac{[X_2,X_1]_\alpha}{\|X_1\|_\alpha^\alpha}$.

Suppose (X_1, X_2) is a bivariate SαS random vector, then by Theorem 4.1.2 in Samorodnitsky and Taqqu (1994), we have $\mathbb{E}(X_1|X_2) = \frac{[X_1,X_2]_\alpha}{\|X_2\|_\alpha^\alpha} X_2$ a.e. Therefore, we can estimate β_1 and β_2 by solving

$$\underset{\beta_1,\beta_2}{\operatorname{argmin}} \sum_{j=1}^{n} \rho\left(\frac{X_{2j}}{X_{1j}} - \beta_1, \frac{X_{1j}}{X_{2j}} - \beta_2\right), \tag{3.13}$$

where ρ is a differentiable convex function a.e. and guarantees a unique solution. Following Sohrabi and Zarepour (2018), we consider the multivariate loss function $\rho(x_1, x_2) = \rho_1(x_1) + \rho_2(x_2)$, satisfying the following standard conditions:

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

Assumption A1: $\rho : \mathbb{R}^2 \to \mathbb{R}$ is a convex and twice differentiable function, and take $\psi_j(\cdot) = \rho_j'(\cdot)$, and $\psi_j'(\cdot) = \rho_j''(\cdot)$. We can simply use Huber's loss function here.

Assumption A2: $\mathbb{E}\left(\psi_1\left(\frac{X_{21}}{X_{11}}\right)\right) = \mathbb{E}\left(\psi_2\left(\frac{X_{11}}{X_{21}}\right)\right) = 0$ and $\mathbb{E}\left(\psi_1^2\left(\frac{X_{21}}{X_{11}}\right)\right) < \infty$, $\mathbb{E}\left(\psi_2^2\left(\frac{X_{11}}{X_{21}}\right)\right) < \infty$, $0 < \mathbb{E}\left|\psi_1'\left(\frac{X_{21}}{X_{11}}\right)\right| < \infty$, and $0 < \mathbb{E}\left|\psi_2'\left(\frac{X_{11}}{X_{21}}\right)\right| < \infty$.

Assumption A3: $\psi_j(\cdot)$ has Lipschitz-continuous derivative $\psi_j'(\cdot)$; i.e., there exists a real constant $k \geq 0$ such that for all x and y, $|\psi_j'(x) - \psi_j'(y)| \leq k|X_1 - X_2|$.

The following theorem gives the asymptotic distribution of the M-estimator $\hat{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top$.

Theorem 6. *Let $\hat{\boldsymbol{\beta}}_n$ be the M-estimator of the parameter $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top$ obtained by the Eq. (3.13) for i.i.d. bivariate SαS random vectors (X_{1i}, X_{2i}), $i = 1, \ldots, n$. Then, $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \mathbf{W}$, where $\mathbf{W}$ is a bivariate Gaussian random vector with zero mean and variance-covariance $\mathbf{c}^{-1}\boldsymbol{\Sigma}_M\mathbf{c}$ with $\mathbf{c}^\top = diag\{c_1, c_2\}$, $c_1 = \mathbb{E}\left(\psi_1'\left(\frac{X_{21}}{X_{11}}\right)\right)$, $c_2 = \mathbb{E}\left(\psi_2'\left(\frac{X_{11}}{X_{21}}\right)\right)$ and*

$$\boldsymbol{\Sigma}_M = \begin{pmatrix} V\left(\psi_1\left(\frac{X_{21}}{X_{11}}\right)\right) & Cov\left(\psi_1\left(\frac{X_{21}}{X_{11}}\right), \psi_2\left(\frac{X_{11}}{X_{21}}\right)\right) \\ Cov\left(\psi_1\left(\frac{X_{21}}{X_{11}}\right), \psi_2\left(\frac{X_{11}}{X_{21}}\right)\right) & V\left(\psi_2\left(\frac{X_{11}}{X_{21}}\right)\right) \end{pmatrix}. \quad (3.14)$$

Note that the results can be extended to higher dimensions.

Remark 4. The variance-covariance matrix of the asymptotic distribution of $\hat{\boldsymbol{\beta}}_n$ does not have a closed-form expression in general. We can use the following estimators to

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

obtain $\hat{\mathbf{\Sigma}}_M$, an estimator for the matrix $\mathbf{\Sigma}_M$.

$$\widehat{V}\left(\psi_1\left(\frac{X_{21}}{X_{11}}\right)\right) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\psi_1\left(\frac{X_{2i}}{X_{1i}}\right) - \overline{\psi_1\left(\frac{X_2}{X_1}\right)}\right)^2,$$

$$\widehat{Cov}\left(\psi_1\left(\frac{X_{21}}{X_{11}}\right), \psi_2\left(\frac{X_{11}}{X_{21}}\right)\right)$$
$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\psi_1\left(\frac{X_{2i}}{X_{1i}}\right) - \overline{\psi_1\left(\frac{X_2}{X_1}\right)}\right)\left(\psi_2\left(\frac{X_{1i}}{X_{2i}}\right) - \overline{\psi_2\left(\frac{X_1}{X_2}\right)}\right),$$

$$\widehat{V}\left(\psi_2\left(\frac{X_{11}}{X_{21}}\right)\right) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\psi_2\left(\frac{X_{1i}}{X_{2i}}\right) - \overline{\psi_2\left(\frac{X_1}{X_2}\right)}\right)^2.$$

where $$\overline{\psi_j\left(\frac{X_2}{X_1}\right)} = \frac{1}{n}\sum_{i=1}^{n}\psi_j\left(\frac{X_{2i}}{X_{1i}}\right), \quad j = 1, 2\,.$$

Therefore, the statistics $n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})^\top \hat{\mathbf{\Sigma}}_M^{-1}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$, for sufficiently large n, has the same distribution as the T^2-Hotelling random variable.

The M-estimator of the $scov(\mathbf{X})$ given in (2.8) obtained from (3.13) in terms of $\hat{\beta}_i$, $i = 1, 2$, is given by

$$\widehat{scov(\mathbf{X})} = \kappa(\hat{\boldsymbol{\beta}})\,|\hat{\beta}_1\,\hat{\beta}_2|^{\frac{1}{2}}\,, \tag{3.15}$$

where $\kappa(\hat{\boldsymbol{\beta}}) = \text{sign}(\hat{\beta}_1)$ if $\text{sign}(\hat{\beta}_1) = \text{sign}(\hat{\beta}_2)$ and equals -1, otherwise.

Based on the definition of scov, and using Theorem 2, the bivariate SαS random vector (X_1, X_2) has positive dependence if $\text{sign}(\hat{\beta}_1) = \text{sign}(\hat{\beta}_2) = 1$, and negative dependence if $\text{sign}(\hat{\beta}_1)\,\text{sign}(\hat{\beta}_2) = -1$.

Building on the results in Theorem 6, which establishes the joint asymptotic normality of the estimator vector $(\hat{\beta}_1, \hat{\beta}_2)^\top$, we can also study the asymptotic behavior of

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

a nonlinear function of the parameters. Specifically, we are interested in the product $\beta_1\,\beta_2$. By applying the delta method to Theorem 6, we obtain

$$\sqrt{n}\left(\hat{\beta}_1\hat{\beta}_2 - \beta_1\beta_2\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\beta_1\beta_2}),$$

where $\sigma^2_{\beta_1\beta_2} = \beta_2^2\,\Upsilon_{11} + 2\,\beta_1\,\beta_2\,\Upsilon_{12} + \beta_1^2\,\Upsilon_{22}$, and Υ_{ij} denote the entries of the asymptotic covariance matrix of the estimator vector $(\hat{\beta}_1, \hat{\beta}_2)^\top$, which is $\Upsilon = \mathbf{c}^{-1}\mathbf{\Sigma}_M\mathbf{c}$, in which $\mathbf{\Sigma}_M$ and $\mathbf{c}$ are given in Theorem 6.

Remark 5. We consider the hypothesis test $H_0:\ \beta_1\beta_2 = 0$ versus $H_1:\ \beta_1\beta_2 \neq 0$. A corresponding Wald test statistic is

$$T_n = \frac{\hat{\beta}_1\hat{\beta}_2 - 0}{\sqrt{\widehat{\sigma^2_{\beta_1\beta_2}}/n}}, \tag{3.16}$$

where $\widehat{\sigma^2_{\beta_1\beta_2}} = \hat{\beta}_2^2\hat{\Upsilon}_{11} + 2\hat{\beta}_1\hat{\beta}_2\hat{\Upsilon}_{12} + \hat{\beta}_1^2\hat{\Upsilon}_{22}$. Under H_0 and for large n, the statistic T_n has an approximate standard normal distribution (see supplementary material for further details).

4. Simulation Study

In this section, we provide numerical illustrations and applications of the measures introduced in Sections 2 and 3. Specifically, we present two simulation examples. In Example 1, we generate independent and identically distributed data from a SαS distribution, while Example 2 applies the signed symmetric covariation coefficient (Subsection 3.4) to simulated data from the sub-Gaussian family of SαS distributions. The purpose of the examples is multifold. We illustrate the asymptotic distribution of the

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

components of the codifference matrix, as established in Theorems 3 and 4. In addition, we estimate (β_1, β_2) using Eq. (3.13) with several standard optimization algorithms. We then perform hypothesis testing using five statistical tests to assess whether a bivariate SαS random vector has independent components, and compare their performance.

The first three tests, $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathcal{A}_3$, evaluate independence based on the product of parameters β_1 and β_2:

$$\mathbb{H}_0 : \ \beta_1 \beta_2 = 0 \quad \text{vs} \quad \mathbb{H}_1 : \ \beta_1 \beta_2 \neq 0.$$

The remaining two tests, $\mathcal{A}_4$ and $\mathcal{A}_5$, are based on the codifference and extended codifference, defined in Eqs. (2.9) and (2.10), respectively:

$$\mathbb{H}_0 : \ \tau_{X_1,X_2} = 0 \quad \text{vs} \quad \mathbb{H}_1 : \ \tau_{X_1,X_2} \neq 0\,,$$

$$\mathbb{H}_0 : \ \tau^*_{X_1,X_2} = 0 \quad \text{vs} \quad \mathbb{H}_1 : \ \tau^*_{X_1,X_2} \neq 0\,.$$

The rejection regions for the tests are as follows:

$\mathcal{A}_1$ Based on Bonferroni's inequality, reject $\mathbb{H}_0$ if $|\hat{\beta}_1| < a_1$ or $|\hat{\beta}_2| < a_2$, with

$$a_1 = t_{\frac{\alpha}{4},n-1}\sqrt{\frac{1}{n}\widehat{V}\big(\psi_1(X_2/X_1)\big)}, \quad a_2 = t_{\frac{\alpha}{4},n-1}\sqrt{\frac{1}{n}\widehat{V}\big(\psi_2(X_1/X_2)\big)},$$

where $t_{\alpha,\nu}$ is the α-quantile of the Student's t distribution with ν degrees of freedom.

$\mathcal{A}_2$ Based on Hotelling's T^2, reject $\mathbb{H}_0$ if $|\hat{\beta}_1| < a_1^*$ or $|\hat{\beta}_2| < a_2^*$, with

$$a_1^* = \sqrt{\widehat{V}(\psi_1(X_2/X_1))\frac{2(n-1)}{n(n-2)}F_{\alpha,2,n-2}}, \quad a_2^* = \sqrt{\widehat{V}(\psi_2(X_1/X_2))\frac{2(n-1)}{n(n-2)}F_{\alpha,2,n-2}},$$

where F_{α,ν_1,ν_2} denotes the $100(1-\alpha)\%$ quantile of the F distribution.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

$\mathcal{A}_3$ Using the Wald-type statistic T_n (Remark 5), reject $\mathbb{H}_0$ if $|T_n| > z_{1-\alpha/2}$, where z_α is the α-quantile of the standard normal distribution.

$\mathcal{A}_4$ Based on the codifference, reject $\mathbb{H}_0$ if $\frac{\sqrt{n}}{\sigma_{1;xy}}\hat{\tau}_{X_1,X_2} > z_{1-\alpha/2}$.

$\mathcal{A}_5$ Based on the extended codifference, reject $\mathbb{H}_0$ if $\frac{\sqrt{n}}{\sigma_{2;xy}}\hat{\tau}^*_{X_1,X_2} > z_{1-\alpha/2}$.

In Examples 1 and 2, for the tests $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathcal{A}_3$, we employ the Huber loss functions ρ_j, $j = 1, 2$, as specified in Theorem 6:

$$\rho_j(x) = \begin{cases} \frac{1}{2}\,x^2, & \text{if } |x| \le d_j, \\ d_j\,|x| - \frac{1}{2}\,d_j^2, & \text{if } |x| > d_j, \end{cases}$$

where d_j is a positive constant (e.g., $d_j = 1.5$).

Example 1. *In this example, we randomly generate independent and identically distributed (i.i.d.) with independent components* (X_{1i}, X_{2i}), $i = 1, 2, \ldots, 50$, *from an* $S\alpha S$ *with stability index* $\alpha = 0.7$ *and scale parameters* $\boldsymbol{\sigma} = (0.5,\, 0.75)^\top$. *The simulation is repeated 10,000 times.*

We then compute the corresponding estimates of the codifference and extended codifference matrices, denoted by $\widehat{\tau}_{X_1,X_2}$ *and* $\widehat{\tau}^*_{X_1,X_2}$, *respectively. These estimators are used to examine the asymptotic behavior of both the off-diagonal and diagonal elements of the codifference and extended codifference matrices, as described in Eqs.* (3.1) *and* (3.2).

Supplementary Figures **??** *and* **??** *presents histograms of the diagonal and off-diagonal entries of* $\widehat{\tau}_{X_1,X_2}$ *and* $\widehat{\tau}^*_{X_1,X_2}$, *for two different values of* $\alpha = 0.7, 1.1$, *based on 10,000 simulation iterations. The results show that the off-diagonal elements are*

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

close to zero, consistent with Theorem 1, which establishes that asymptotic independence implies vanishing off-diagonal codifference terms.

Table 1 presents estimated values of the parameters (β_1, β_2) *for various values of the stability index and scale parameters, obtained using Eq.* (3.13) *and several standard optimization algorithms. Specifically, we employ the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm, the Limited-memory BFGS (L-BFGS) variant, the Nelder–Mead simplex algorithm, and the Conjugate Gradient (CG) method. These methods are widely used for solving unconstrained nonlinear optimization problems and are available in most scientific computing libraries. BFGS and L-BFGS are quasi-Newton methods that iteratively approximate the Hessian matrix; CG is a gradient-based method particularly suited for large-scale problems. In contrast, the Nelder–Mead algorithm is derivative-free and effective when gradient information is unavailable or unreliable. Table 1 illustrates (i) the results across various computational algorithms are consistent, indicating the robustness of our novel parameter estimates; (ii) all estimated values of* β_1 *and* β_2 *(and products) are close to 0, indicating evidence towards the independence assumption.*

Table 2 presents the p-values for the statistical tests $\mathcal{A}_1, \ldots, \mathcal{A}_5$ *in Section 4. In all cases, the p-values exceed* 0.05*; thus, none of the null hypotheses are rejected. These results provide no statistical evidence against the independence of* β_1 *and* β_2*, nor against a zero codifference and extended codifference between* X_1 *and* X_2*. Overall, the simulation results suggest that the components of the S*α*S random vector are likely independent.*

Example 2. *Suppose that* $\mathbf{G}_i := (G_{1i}, G_{2i})$*, for* $i = 1, 2, \ldots, 500$*, be zero-mean jointly Gaussian random vectors with covariance entries defined by* $\gamma_{kj} := \mathbb{E}(G_{ki}\, G_{ji})$*, (*$k, j =$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

Table 1: Estimation of $(\hat{\beta}_1, \hat{\beta}_2)$ for various stability indices and scale parameters $\boldsymbol{\sigma} = (\sigma_1, \sigma_2)^\top$ in i.i.d. SαS samples

α	σ_1	σ_2	BFGS	L-BFGS	Nelder-Mead	CG
0.2	0.5	0.75	(0.00104, 0.00800)	(0.00274, 0.00855)	(0.02915, 0.03011)	(0.00042, 0.00765)
0.2	1.5	1.2	(0.00038, 0.00735)	(0.00178, 0.00831)	(0.02835, 0.03122)	(0.00056, 0.00757)
0.6	0.5	0.75	(0.00369, 0.00455)	(0.00374, 0.00458)	(0.00370, 0.00459)	(0.00371, 0.00459)
0.6	1.5	1.2	(0.00246, 0.00646)	(0.00252, 0.00645)	(0.00275, 0.00651)	(0.00251, 0.00649)
0.9	0.5	0.75	(0.00419, 0.00364)	(0.00418, 0.00366)	(0.00418, 0.00366)	(0.00418, 0.00363)
0.9	1.5	1.2	(0.00241, 0.00620)	(0.00242, 0.00619)	(0.00241, 0.00619)	(0.00241, 0.00619)
1.1	0.5	0.75	(0.00433, 0.00361)	(0.00433, 0.00361)	(0.00437, 0.00361)	(0.00433, 0.00361)
1.1	1.5	1.2	(0.00244, 0.00595)	(0.00243, 0.00595)	(0.00244, 0.00596)	(0.00244, 0.00595)
1.6	0.5	0.75	(0.00521, 0.00287)	(0.00521, 0.00287)	(0.00521, 0.00287)	(0.00521, 0.00287)
1.6	1.5	1.2	(0.00294, 0.00502)	(0.00294, 0.00502)	(0.00294, 0.00502)	(0.00294, 0.00502)
1.9	0.5	0.75	(0.00629, 0.00237)	(0.00629, 0.00237)	(0.00630, 0.00236)	(0.00629, 0.00237)
1.9	1.5	1.2	(0.00365, 0.00384)	(0.00365, 0.00384)	(0.00366, 0.00384)	(0.00365, 0.00384)
2	0.5	0.75	(0.00670, 0.00213)	(0.00670, 0.00213)	(0.00670, 0.00213)	(0.00670, 0.00213)
2	1.5	1.2	(0.00389, 0.00354)	(0.00390, 0.00354)	(0.00388, 0.00353)	(0.00390, 0.00354)

1, 2). *Fix* $1 < \alpha < 2$ *and let* $A_i \sim S_{\alpha/2}\left(\left(\cos\frac{\pi\alpha}{4}\right)^{2/\alpha}, 1, 0\right)$ *be a positive* $\alpha/2$*-stable random variable, independent of* $\mathbf{G}_i$*. Define the sub-Gaussian random vector* $\mathbf{X}_i := (X_{1i}, X_{2i}) := A_i^{1/2}\,\mathbf{G}_i$.

In this example, we generate 1000 iterations from the model and use the Huber loss function, as described in Section 4, with $\gamma_{11} = 0.5$ *and* $\gamma_{22} = 0.75$.

The histograms of the off-diagonal and diagonal components of the sample matrices

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

Table 2: P-values for hypothesis tests with null hypotheses $\mathcal{A}_i$, $i = 1, 2, \ldots, 4$, conducted on i.i.d. SαS samples for various stability indices and scale parameters. The significance level for the tests is set at $\alpha = 0.05$.

α	σ_1	σ_2	$\mathcal{A}_1$	$\mathcal{A}_2$	$\mathcal{A}_3$	$\mathcal{A}_4$	$\mathcal{A}_5$
0.2	0.5	0.75	0.51582	0.35678	0.69732	0.86106	0.85744
0.2	1.5	1.2	0.50984	0.35372	0.69618	0.94708	0.95818
0.6	0.5	0.75	0.31633	0.29006	0.70262	0.82698	0.78369
0.6	1.5	1.2	0.31804	0.29186	0.70456	0.94171	0.93867
0.9	0.5	0.75	0.33664	0.30749	0.70172	0.85628	0.79709
0.9	1.5	1.2	0.34200	0.31184	0.70283	0.87077	0.87242
1.1	0.5	0.75	0.34948	0.31709	0.70186	0.89102	0.83491
1.1	1.5	1.2	0.35608	0.32331	0.70215	0.82016	0.80129
1.6	0.5	0.75	0.37185	0.33446	0.70434	0.93520	0.85016
1.6	1.5	1.2	0.38433	0.34566	0.70380	0.76512	0.68864
1.9	0.5	0.75	0.38052	0.34017	0.70607	0.95240	0.84732
1.9	1.5	1.2	0.39830	0.35532	0.70394	0.77235	0.68768
2	0.5	0.75	0.38254	0.34167	0.70620	0.95577	0.83667
2	1.5	1.2	0.40138	0.35765	0.70363	0.77634	0.69182

$\widehat{\tau}_{X_1,X_2}$ and $\widehat{\tau}^*_{X_1,X_2}$ are displayed in Supplementary Figure **??**, *for $\alpha = 0.2$ and $\gamma_{12} = -0.3$. The distributional shape indicates strong evidence that these components follow an approximately normal distribution.*

Supplementary Figure **??** *illustrates the confidence region for the estimated values*

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

of (β_1, β_2)*, constructed using Hotelling's* T^2 *statistic and based on the BFGS optimization algorithm, under the setting* $\alpha = 0.2$*. The Hotelling's* T^2*-based confidence ellipses are centered at the origin, with both* β_1 *and* β_2 *lying close to zero. This result is consistent with the null hypothesis of independence, indicating no significant evidence of association between the components.*

To numerically verify Theorem 4, we estimate (β_1, β_2) *from simulated data and examine the distribution of* $\widehat{\tau}_{X_1,X_2}$ *and* $\widehat{\tau}^*_{X_1,X_2}$*. Histograms in Supplementary Figure* **??** *suggest that the components are approximately normally distributed, providing empirical support for the theorem.*

From Example 2.7.4 in Samorodnitsky and Taqqu (1994), we obtain $[X_1, X_2]_\alpha = 2^{-\alpha/2}\,\gamma_{12}\,\gamma_{22}^{(\alpha-2)/2}$*, and* $[X_2, X_1]_\alpha = 2^{-\alpha/2}\,\gamma_{12}\,\gamma_{11}^{(\alpha-2)/2}$*. Furthermore,* $\|X_1\|_\alpha^\alpha = 2^{-\alpha/2}\,\gamma_{11}^{\alpha/2}$*, and* $\|X_2\|_\alpha^\alpha = 2^{-\alpha/2}\,\gamma_{22}^{\alpha/2}$*. Consequently, we have*

$$\beta_1 = \frac{[X_1, X_2]_\alpha}{\|X_1\|_\alpha^\alpha} = \frac{\gamma_{12}}{\gamma_{11}}, \qquad \beta_2 = \frac{[X_1, X_2]_\alpha}{\|X_2\|_\alpha^\alpha} = \frac{\gamma_{12}}{\gamma_{22}}. \tag{4.1}$$

As a result, β_1 *and* β_2 *are independent of* α*, and we therefore expect their estimators to remain stable as* α *varies. This behavior is confirmed in Table 3, that reports estimates of* (β_1, β_2) *obtained using the methods described in Table 1 of Example 1. Similar to Table 1, all estimation methods yield closely aligned results, providing strong support for the validity of the parameter estimates. To further evaluate these estimators, we compute their errors relative to the theoretical values given in* (4.1)*. The results, shown in Table 4, demonstrate that the proposed methods produce parameter estimates with minimal error.*

Overall, the tables indicate that estimation accuracy remains stable across different

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

values of α and γ_{12}, confirming the robustness of the proposed procedures.

Table 5 reports the p-values for the statistical tests $\mathcal{A}_i$, $i = 1, \ldots, 5$ (as defined in Section 4), evaluated at the 0.05 significance level across different stability indices α and scale parameters. The corresponding codifference measures, τ_{X_1,X_2} and $\tau^*_{X_1,X_2}$ (computed using Eqs. (4.2) and (4.3)), are reported in parentheses, offering additional insight into the dependence structure of the variables.

The first three tests, $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathcal{A}_3$, may fail to detect dependence under certain sub-Gaussian settings. Specifically, when the marginals are dependent, but $\gamma_{12} = 0$, these tests do not reject the null hypothesis of independence. This occurs because the test parameters β_i reduce to the correlation parameter (see Eq. (4.1)), which vanishes even though the underlying sub-Gaussian vector remains dependent. As a result, these tests may incorrectly indicate independence.

Hypothesis $\mathcal{A}_4$ shows reduced performance for $\alpha > 1$, while $\mathcal{A}_5$ becomes unstable as α approaches 2. In such cases, we recommend using the approach described in Eq. (2.12) for a more reliable assessment of dependence. The limitations of $\mathcal{A}_4$ reflect the fact that a zero codifference does not generally imply independence for SαS vectors (see Theorem 2). Likewise, $\mathcal{A}_5$ fails to produce valid p-values near $\alpha = 2$, as the extended codifference is not defined for this value (see Remark 2).

According to Proposition 2.5.2 in Samorodnitsky and Taqqu (1994), the joint characteristic functions of $X_1 + X_2$ and $X_1 - X_2$ are given by

$$\mathbb{E}\big(\exp(i(X_1 + X_2))\big) = \exp\Big(-\big|\frac{1}{2}(\gamma_{11} + \gamma_{22} + 2\gamma_{12})\big|^{\frac{\alpha}{2}}\Big),$$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

$$\mathbb{E}\big(\exp(i(X_1 - X_2))\big) = \exp\Big(-\big|\frac{1}{2}(\gamma_{11} + \gamma_{22} - 2\gamma_{12})\big|^{\frac{\alpha}{2}}\Big).$$

These expressions demonstrate that X_1 *and* X_2 *may remain dependent even when* $\gamma_{12} = 0$*, highlighting that zero covariance does not imply independence in sub-Gaussian or more general stable distributions.*

The codifference measures are computed as

$$\tau_{X_1,X_2} = \Big|\frac{1}{2}\gamma_{11}\Big|^{\frac{\alpha}{2}} + \Big|\frac{1}{2}\gamma_{22}\Big|^{\frac{\alpha}{2}} - \Big|\frac{1}{2}(\gamma_{11} + \gamma_{22} - 2\gamma_{12})\Big|^{\frac{\alpha}{2}}, \tag{4.2}$$

$$\tau^*_{X_1,X_2} = \Big|\frac{1}{2}\gamma_{11}\Big|^{\frac{\alpha}{2}} + \Big|\frac{1}{2}\gamma_{22}\Big|^{\frac{\alpha}{2}} - \frac{1}{2}\Big|\frac{1}{2}(\gamma_{11} + \gamma_{22} - 2\gamma_{12})\Big|^{\frac{\alpha}{2}} - \frac{1}{2}\Big|\frac{1}{2}(\gamma_{11} + \gamma_{22} + 2\gamma_{12})\Big|^{\frac{\alpha}{2}}. \tag{4.3}$$

These codifference and extended codifference measures complement the p-values, providing a robust evaluation of the dependence structure among the variables.

5. Single-Cell Clustering

Cell or gene clustering has an important role in identifying putative cell types in complex tissues using gene expression patterns. For example, Single-cell RNA sequencing, as a groundbreaking method for this purpose, allows scientists to gather large catalogs detailing the transcriptomes of thousands of individual cells, with the main focus of clustering individual cells using unsupervised clustering methods (Kiselev et al., 2019) based on gene expression patterns.

There are multiple computational packages for clustering of single cells, including Scanpy (Wolf et al., 2017) and Seurat (Satija et al., 2015). All the existing packages

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

Table 3: Estimation of (β_1, β_2) based on sub-Gaussian samples

α	γ_{12}	BFGS	L-BFGS	Nelder-Mead	CG
0.2	-0.3	$(-0.59908, -0.39840)$	$(-0.59908, -0.39840)$	$(-0.59908, -0.39840)$	$(-0.59908, -0.39840)$
0.2	0	$(0.00210, 0.00317)$	$(0.00210, 0.00317)$	$(0.00206, 0.00318)$	$(0.00210, 0.00317)$
0.2	0.3	$(0.59768, 0.39927)$	$(0.59768, 0.39927)$	$(0.59765, 0.39928)$	$(0.59768, 0.39927)$
0.6	-0.3	$(-0.59908, -0.39840)$	$(-0.59908, -0.39840)$	$(-0.59908, -0.39840)$	$(-0.59908, -0.39840)$
0.6	0	$(0.00210, 0.00317)$	$(0.00210, 0.00317)$	$(0.00206, 0.00318)$	$(0.00210, 0.00317)$
0.6	0.3	$(0.59768, 0.39927)$	$(0.59768, 0.39927)$	$(0.59765, 0.39928)$	$(0.59768, 0.39927)$
0.9	-0.3	$(-0.59908, -0.39840)$	$(-0.59908, -0.39840)$	$(-0.59908, -0.39840)$	$(-0.59908, -0.39840)$
0.9	0	$(0.00210, 0.00317)$	$(0.00210, 0.00317)$	$(0.00206, 0.00318)$	$(0.00210, 0.00317)$
0.9	0.3	$(0.59768, 0.39927)$	$(0.59768, 0.39927)$	$(0.59765, 0.39928)$	$(0.59768, 0.39927)$
1.1	-0.3	$(-0.59908, -0.39840)$	$(-0.59908, -0.39840)$	$(-0.59908, -0.39840)$	$(-0.59908, -0.39840)$
1.1	0	$(0.00210, 0.00317)$	$(0.00210, 0.00317)$	$(0.00206, 0.00318)$	$(0.00210, 0.00317)$
1.1	0.3	$(0.59768, 0.39927)$	$(0.59768, 0.39927)$	$(0.59765, 0.39928)$	$(0.59768, 0.39927)$
1.6	-0.3	$(-0.59908, -0.39840)$	$(-0.59908, -0.39840)$	$(-0.59908, -0.39840)$	$(-0.59908, -0.39840)$
1.6	0	$(0.00210, 0.00317)$	$(0.00210, 0.00317)$	$(0.00206, 0.00318)$	$(0.00210, 0.00317)$
1.6	0.3	$(0.59768, 0.39927)$	$(0.59768, 0.39927)$	$(0.59765, 0.39928)$	$(0.59768, 0.39927)$
1.9	-0.3	$(-0.59908, -0.39840)$	$(-0.59908, -0.39840)$	$(-0.59908, -0.39840)$	$(-0.59908, -0.39840)$
1.9	0	$(0.00210, 0.00317)$	$(0.00210, 0.00317)$	$(0.00206, 0.00318)$	$(0.00210, 0.00317)$
1.9	0.3	$(0.59768, 0.39927)$	$(0.59768, 0.39927)$	$(0.59765, 0.39928)$	$(0.59768, 0.39927)$

mainly employ a sequence of computational methods, including standardization, principal component analysis (PCA), and several clustering Algorithms, including Lieden and Louvain (for Scanpy). In addition, these packages have a primary screening process that removes genes with sparse expression values, and those include extreme values or

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

Table 4: Error in estimation of (β_1, β_2) based on sub-Gaussian samples

α	γ_{12}	BFGS	L-BFGS	Nelder-Mead	CG
0.2	−0.3	(0.00092, 0.00160)	(0.00092, 0.00160)	(0.00092, 0.00160)	(0.00092, 0.00160)
0.2	0	(0.00210, 0.00317)	(0.00210, 0.00317)	(0.00210, 0.00317)	(0.00210, 0.00317)
0.2	0.3	(−0.00232, −0.00073)	(−0.00232, −0.00073)	(−0.00235, −0.00072)	(−0.00232, −0.00073)
0.6	−0.3	(0.00092, 0.00160)	(0.00092, 0.00160)	(0.00092, 0.00160)	(0.00092, 0.00160)
0.6	0	(0.00210, 0.00317)	(0.00210, 0.00317)	(0.00206, 0.00318)	(0.00210, 0.00317)
0.6	0.3	(−0.00232, −0.00073)	(−0.00232, −0.00073)	(−0.00235, −0.00072)	(−0.00232, −0.00073)
0.9	−0.3	(0.00092, 0.00160)	(0.00092, 0.00160)	(0.00092, 0.00160)	(0.00092, 0.00160)
0.9	0	(0.00210, 0.00317)	(0.00210, 0.00317)	(0.00206, 0.00318)	(0.00210, 0.00317)
0.9	0.3	(−0.00232, −0.00073)	(−0.00232, −0.00073)	(−0.00235, −0.00073)	(−0.00232, −0.00073)
1.1	−0.3	(0.00092, 0.00160)	(0.00092, 0.00160)	(0.00092, 0.00160)	(0.00092, 0.00160)
1.1	0	(0.00210, 0.00317)	(0.00210, 0.00317)	(0.00206, 0.00318)	(0.00210, 0.00317)
1.1	0.3	(−0.00232, −0.00073)	(−0.00232, −0.00073)	(−0.00235, −0.00072)	(−0.00232, −0.00073)
1.6	−0.3	(0.00092, 0.00160)	(0.00092, 0.00160)	(0.00092, 0.00160)	(0.00092, 0.00160)
1.6	0	(0.00210, 0.00317)	(0.00210, 0.00317)	(0.00207, 0.00318)	(0.00210, 0.00317)
1.6	0.3	(−0.00232, −0.00073)	(−0.00232, −0.00073)	(−0.00235, −0.00072)	(−0.00232, −0.00073)
1.9	−0.3	(0.00092, 0.00160)	(0.00092, 0.00160)	(0.00092, 0.00160)	(0.00092, 0.00160)
1.9	0	(0.00210, 0.00317)	(0.00210, 0.00317)	(0.00206, 0.00318)	(0.00210, 0.00317)
1.9	0.3	(−0.00232, −0.00073)	(−0.00232, −0.00073)	(−0.00235, −0.00072)	(−0.00232, −0.00073)

have heavy tails. Hence, they include a fraction of the genes in the analysis.

However, removing genes with extreme expression causes loss of information, and the PCA method ignores the information in extreme observations, as it only works for well-behaved normal data. In this section, we apply spectral clustering using our novel

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

Table 5: P-values for hypothesis tests with null hypotheses $\mathcal{A}_i$, $i = 1, 2, \ldots, 4$

α	γ_{12}	$\mathcal{A}_1$	$\mathcal{A}_2$	$\mathcal{A}_3$	$\mathcal{A}_4$	$\mathcal{A}_5$
0.2	-0.3	0.00000	0.00000	0.00067	0.00000 (0.78489)	0.00000 (0.83416)
0.2	0	0.38921	0.34707	0.70926	0.00000 (0.82304)	0.00000 (0.82304)
0.2	0.3	0.00000	0.00000	0.00057	0.00000 (0.88343)	0.00000 (0.83416)
0.6	-0.3	0.00000	0.00000	0.00067	0.00000 (0.42796)	0.00000 (0.55951)
0.6	0	0.38921	0.34707	0.70926	0.00000 (0.53636)	0.00000 (0.53636)
0.6	0.3	0.00000	0.00000	0.00057	0.00000 (0.69106)	0.00000 (0.55951)
0.9	-0.3	0.00000	0.00000	0.00067	0.00475 (0.21351)	0.00000 (0.39476)
0.9	0	0.38921	0.34707	0.70926	0.00000 (0.36967)	0.00000 (0.36967)
0.9	0.3	0.00000	0.00000	0.00057	0.00000 (0.57600)	0.00000 (0.39476)
1.1	-0.3	0.00000	0.00000	0.00067	0.18946 (0.09155)	0.00000 (0.30110)
1.1	0	0.38921	0.34707	0.70926	0.00001 (0.27737)	0.00000 (0.27737)
1.1	0.3	0.00000	0.00000	0.00057	0.00000 (0.51065)	0.00000 (0.30110)
1.6	-0.3	0.00000	0.00000	0.00067	0.00280 (-0.15339)	0.00033 (0.11292)
1.6	0	0.38921	0.34707	0.70926	0.01864 (0.09955)	0.00039 (0.09955)
1.6	0.3	0.00000	0.00000	0.00057	0.00000 (0.37923)	0.00034 (0.11292)
1.9	-0.3	0.00000	0.00000	0.00067	0.00000 (-0.26682)	0.16522 (0.02559)
1.9	0	0.38921	0.34707	0.70926	0.50699 (0.02193)	0.13998 (0.02193)
1.9	0.3	0.00000	0.00000	0.00057	0.00000 (0.31801)	0.17020 (0.02559)

codifference dependence measure to a single-cell RNA-Seq data collected from Adipose tissue (Rajbhandari et al., 2019) and compare our results with those of the Scanpy

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

package (Wolf et al., 2017). Our method is extremely fast and can be applied to the whole data without pre-processing, standardization, or removal of the sparse genes.

The Adipose tissue single-cell RNA-seq data includes 36661 individual cells and 27334 genes; it is highly sparse, and RNA-Seq expression of some of the cells is heavy-tailed. Figure 1 (Left panel) presents a histogram of the percentage of non-zero expression values for 1000 selected cells with an estimated $0 < \alpha < 2$. We use the Hill estimator (Hill, 1975) to estimate the tail index. The majority of these cells are sparse with non-zero expression values for less than 10% of all genes. Also, Figure 1 (Middle and Right panels) presents the histograms for two cells with estimated $\alpha = 0.361$ and 1.368, demonstrating that they are heavy-tailed with different index values.

We applied spectral clustering (Von Luxburg, 2007) using the codifference matrix as the adjacency matrix and compared the result with the Scanpy package implemented using both Leiden and Louvain clustering algorithms, presented via diffusion map in Figure 2. Although Scanpy presents moderately clearer identification of some clusters, the results of our clustering using the codifference measure are quite comparable. The remaining noise in the spectral clustering can be due to multiple factors.

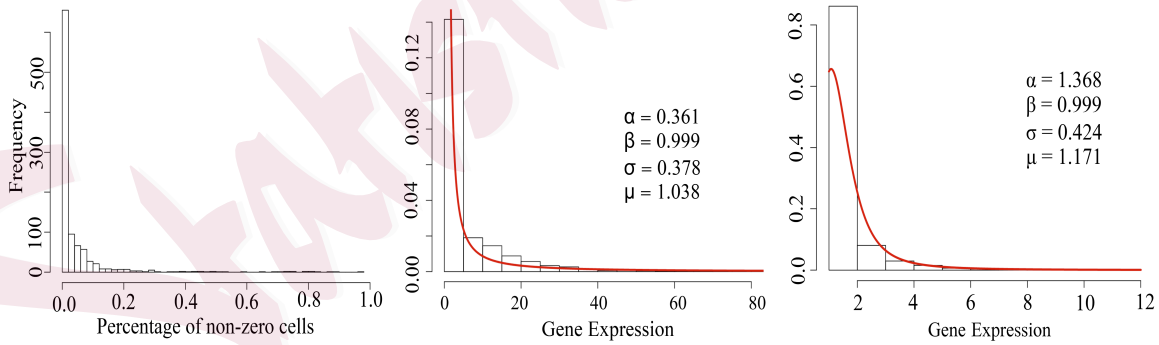
First, Scanpy uses a small fraction of highly variable genes for the initial PCA dimensionality reduction, while we do not exclude any gene from clustering analysis. This is particularly important in identifying genes that are differentially regulated in small subsets of cells. Moreover, this can provide biologists with a potential new sub-cluster of cells that differ from one another only in the expression of some of these genes. Second, we use the data in its full dimensionality, but Scanpy reduces the dimension

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

significantly, which reduces variation in the data.

In summary, this example contributes to two important fronts: scientific and computational. Scientifically, our approach preserves a greater portion of the original signal, particularly low-expression genes that may nonetheless play critical roles in cellular differentiation or disease pathways. Computationally, our method demonstrates stable clustering behavior directly on raw data, whereas traditional pipelines often require aggressive filtering and pre-processing before producing meaningful results.

Figure 1: (Left) Presents the histogram of percentage of genes with non-zero expression values within a single cell, for 1000 individual cells with $\alpha < 2$. (Middle, Right) present histograms of the expression values for the entire gene population in two single cells, (Left) *Adipocyte* and (Right) *Macrophage*. The tail indexes are estimated using the Hill method, and other parameters are estimated using *StableEstim* R packages.

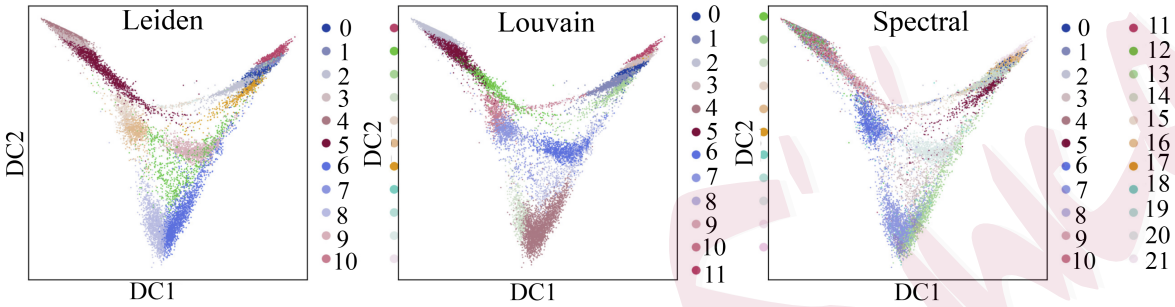


6. Discussion

In this paper, we studied multiple measures of dependence for heavy-tailed data and proposed extended codifference as an alternative measure of dependence for heavy-

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

Figure 2: Diffusion maps presenting clustering maps based on Scanpy (left and middle panels) and based on codifference measure (right panel).



tailed random variables. We also proposed non-parametric estimators for codifference and extended codifference, which are much easier to compute as they do not require estimating the α index and spectral measure.

This work can be extended and improved in multiple directions. One possible direction is to develop a linear dimensionality reduction method that, unlike PCA, takes the tail behavior into account. This will help reduce the remaining noise in the data before clustering, while using the information hidden in the extreme expression values in the clustering. Another possible direction is to study measures of association between random vectors in the domain of attraction of a stable distribution with different tail indices. This class is a large class of heavy-tailed distributions and includes the class of regularly varying distributions.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

Supplementary Material

The online Supplementary Material contains selected graphs for the simulation example, along with proofs and detailed derivations of the theorems and remarks.

Acknowledgement

Dr. Maroufy and Rezapour were in part supported by Dr. Maroufy's faculty start-up funds from The University of Texas Health Science Center at Houston. Dr. Zarepur was supported by NSERC grant RGPIN/ 2018-04008.

References

Abu Alfeilat, H. A., Hassanat, A. B., Lasassmeh, O., Tarawneh, A. S., Alhasanat, M. B., Eyal Salman, H. S., and Prasath, V. S., (2019). Effects of distance measure choice on k-nearest neighbor classifier performance: a review. *Big data,* 7(4), 221-248.

Alparslan U.-U., Nolan J.-P., (2016). Measure of dependence for stable distributions. *Extremes,* 19, 303-323.

Cambanis, S., Soltani, A.R., (1984). Prediction of stable processes: Spectral and moving average representations. *Z. Wahrsch. Verw. Gebiete,* 66, 593-612.

Cantoni, E., and Ronchetti, E., (2006). A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditures. *Journal of Health Economics,* 25(2) 198-213.

Cooley, D., and Thibaud, E., (2019). Decompositions of dependence for high-dimensional extremes. *Biometrika,* 106(3), 587-604.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

REFERENCES

Davis, R.A., Mikosch, T. and Pfaffel, O. (2016). Asymptotic theory for the sample covariance matrix of a heavy-tailed multivariate time series. *Stoch. Proc. Appl.* 126, 767-799.

Damarackas, J., Paulauskas, V. (2017). Spectral covariance and limit theorems for random fields with infinite variance. *Journal of Multivariate Analysis.* 153, 156-175.

de Haan, L., Peng, L., (1998). Comparison of tail index estimators. *Statistica Neerlandica* 52, 60-70.

Fan, J., and Lv, J., (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 70(5), 849-911.

Feller W., (1971). *An introduction to probability theory and its applications.* Vol. 2. 2nd ed. New York: Wiley.

Fiche, A., Cexus, J. C., Martin, A., and Khenchaf, A., (2013). Features modeling with an α-stable distribution: Application to pattern recognition based on continuous belief functions. *Information Fusion,* 14(4), 504-520.

Forsythe, G. E., and Golub, G. H. (1965). On the stationary values of a second-degree polynomial on the unit sphere. *Journal of the Society for Industrial and Applied Mathematics*, 13(4), 1050-1068.

Garel, B., Kodia, B., (2009). Signed symmetric covariation coefficient for alpha-stable dependence modeling. *Comptes Rendus de l'Académie des Sciences - Series I*, 347-352.

HILL, B. (1975). A simple approach to inference about the tail of a distribution. *Annals of Statistics* 3, 1163-1174.

Hashimshony, Tamar, Wagner, Florian, Sher, Noa, and Yanai, Itai. (2012). Cel-seq: single-cell RNA-seq by multiplexed linear amplification. *Cell reports,* 2(3), 666-673.

Heiny, J. and Mikosch, T. (2019). The eigenstructure of the sample covariance matrices of high-dimensional

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

stochastic volatility models with heavy tails. *Bernoulli*, 25(4B), 3590-3622.

Heiny, J., Mikosch, T., and Yslas, J. (2021). Point process convergence for the off-diagonal entries of sample covariance matrices. *The Annals of Applied Probability,* 31(2), 538-560.

Huang, W. K., Cooley, D. S., Ebert-Uphoff, I., Chen, C., and Chatterjee, S., (2019). New Exploratory Tools for Extremal Dependence: χ Networks and Annual Extremal Networks. *Journal of Agricultural, Biological and Environmental Statistics,* 24(3), 484-501.

Jaitin, Diego Adhemar, Kenigsberg, Ephraim, KerenShaul, Hadas, Elefant, Naama, Paul, Franziska, Zaretsky, Irina, Mildner, Alexander, Cohen, Nadav, Jung, Steffen, Tanay, Amos, et al. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science,* 343(6172) 776-779.

Jiang, Y., Cooley, D., and Wehner, M. F. (2020). Principal Component Analysis for Extremes and Application to US Precipitation. *Journal of Climate,* 33(15), 6441-6451.

Kiselev, V. Y., Andrews, T. S., and Hemberg, M., (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics,* 20(5), 273-282.

Knight, K. (1989). On the bootstrap of the sample mean in the infinite variance case. *The Annals of Statistics,* 17(3), 1168-1175.

Kodia, B., Garel, B. (2014). Estimation and Comparison of Signed Symmetric Covariation Coefficient and Generalized Association Parameter for Alpha-stable Dependence Modeling. *Communications in Statistics - Theory and Methods,* 43, 24, 5156-5174.

Kokoszka, P.S., Taqqu, M.S., (1993). Asymptotic dependence of moving average type self-similar stable random fields. *Nagoya Mathematical Journal*, 130, 85-100.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

REFERENCES

Kuznetsov, V. A. Knott, G. D. and Bonner, R. F. (2002). General statistics of stochastic process of gene expression in eukaryotic cells. *Genetics* 161, 1321-1332.

Lev S Tsimring (2014). Noise in Biology. *Rep. Prog. Phys.* 77: 026601.

Newman M.E.J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics* 46, 323-351.

Nolan, J.-P., (2016). *Stable Distributions - Models for heavy-tailed Data* Springer New York.

Samorodnitsky, G., and Taqqu, M. S. (1994), *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Chapman & Hall/CRC, Florida.

Oshlack, Alicia, Robinson, Mark D, Young, Matthew D, et al. (2010). From RNA-seq reads to differential expression result. . *Genome biol,* 11(12):220.

Oyelade, J., Itunuoluwa I., Funke O., Olufemi A., Efosa U., Faridah A., Moses A., and Ezekiel A. (2016). Clustering algorithms: their application to gene expression data. *Bioinformatics and Biology insights,* 10, BBI-S38316.

Pele, D. T., and Stanciulescu, V. N., (2015). On a Class of Alpha-stable Distributions and Its Applications in Estimating Market Risk. *The Review of Finance and Banking,* 7(2), 007-015.

Prabhakaran, S., Azizi, E., Carr, A., & Pe'er, D. (2016). Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. *In International Conference on Machine Learning* 1070-1079.

Press, S. J., (1972). Multivariate stable distributions. *Journal of Multivariate Analysis*, 2, 444-462.

Rajbhandari, P., Arneson, D., Hart, S.K., Ahn, I.S., Diamante, G., Santos, L.C., Zaghari, N., Feng, A.C., Thomas, B.J., Vergnes, L. and Lee, S.D., (2019). Single cell analysis reveals immune cell–adipocyte crosstalk regulating the transcription of thermogenic adipocytes. *Elife,* 8, p.e49501.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

REFERENCES

Rahimi, , A., and Recht, B. (2007). Random Features for Large-Scale Kernel Machines. *In NIPS* Vol. 3 No. 4, 1177–1184

Resnick, S., Greenwood, P., (1979). A bivariate stable characterization and domain of attraction. *Journal of Multivariate Analysis,* 9, 206-221.

Rohrbeck, C., and Cooley, D., (2021). Simulating flood event sets using extremal principal components. arXiv preprint arXiv:2106.00630.

Rosadi, D., (2006). Order identification for Gaussian moving averages using the codifference function. *Journal of Statistical Computation and Simulation*, 76, 6, 553-559.

Russell, B. T., Cooley, D. S., Porter, W. C., Reich, B. J., and Heald, C. L. (2016). Data mining to investigate the meteorological drivers for extreme ground level ozone events. *Annals of Applied Statistics,* 10(3), 1673-1698.

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., and Regev, A., (2015). Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, 33(5), 495-502.

Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *The Annals of Statistics*, 9, 1187-1195.

Sohrabi, M., Zarepour, M. (2018). Bootstrapping the mean vector for the observations in the domain of attraction of a multivariate stable law. *Statistics*, 52, 50-63.

Spjøtvoll, E., (1972). A Note on a Theorem of Forsythe and Golub. *SIAM Journal on Applied Mathematics,* 23(3), 307-311.

Székely, Gábor J and Rizzo, Maria L and Bakirov, Nail K (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35, 6, 2769-2794.

Xu, C., and Su, Z., (2015). Identification of cell types from single-cell transcriptomes using a novel clustering

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

method. *Bioinformatics,* 31(12), 1974-1980.

Vallejos, Catalina A., John C. Marioni and Richardson, Sylvia. (2015). Basics: Bayesian analysis of single-cell sequencing data. *PLoS Computational Biology,* 11(6) e1004333.

Von Luxburg, U., (2007). A tutorial on spectral clustering. *Statistics and computing,* 17(4), 395-416.

Williams, C., and Seeger, M. (2001). Using the Nyström method to speed up kernel machines. In Proceedings of the 14th annual conference on neural information processing systems (No. CONF, pp. 682-688).

Wolf, F. A., Angerer, P., and Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome biology,* 19(1), 1-5.

Zhu, A., Ibrahim, J. G., and Love, M. I., (2019). Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics,* 35(12), 2084-2092.

Żak, G., Teuerle, M., Wyłomańska, A., and Zimroz, R., (2017). Measures of dependence for α-stable distributed processes and its application to diagnostics of local damage in presence of impulsive noise. *Shock and Vibration,* 2017 (1), 1-9.

Department of Biostatistics and Data Science, School of Public Health, University of Texas Health Center at Houston

E-mail: (vahed.maroufy@uth.tmc.edu)

ORCID: https://orcid.org/0000-0002-6653-7586

Alumni of the Department of Biostatistics and Data Science, School of Public Health, University of Texas Health Center at Houston

E-mail: (mohsenrzp@gmail.com)

Dept. of Mathematics and Statistics, University of Ottawa

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0370

E-mail: (Mahmoud.Zarepour@uottawa.ca)

Dept. of Statistics, Shahid Bahonar University, Kerman

E-mail: (bafhami@gmail.com.)