Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Dissecting Gene Expression Heterogeneity: Generalized Pearson Correlation Squares and the *K*-Lines Clustering Algorithm

Jingyi Jessica Li, Heather J. Zhou, Peter J. Bickel & Xin Tong

To cite this article: Jingyi Jessica Li, Heather J. Zhou, Peter J. Bickel & Xin Tong (2024) Dissecting Gene Expression Heterogeneity: Generalized Pearson Correlation Squares and the *K*-Lines Clustering Algorithm, Journal of the American Statistical Association, 119:548, 2450-2463, DOI: 10.1080/01621459.2024.2342639

To link to this article: https://doi.org/10.1080/01621459.2024.2342639

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2024, VOL. 119, NO. 548, 2450–2463: Applications and Case Studies
https://doi.org/10.1080/01621459.2024.2342639

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Dissecting Gene Expression Heterogeneity: Generalized Pearson Correlation Squares and the *K*-Lines Clustering Algorithm

Jingyi Jessica Li[a], Heather J. Zhou[a], Peter J. Bickel[b], and Xin Tong[c]

[a]Department of Statistics and Data Science, University of California, Los Angeles, Los Angeles, CA; [b]Department of Statistics, University of California, Berkeley, Berkeley, CA; [c]Department of Data Sciences and Operations, University of Southern California, Los Angeles, CA

ABSTRACT
Motivated by the pressing needs for dissecting heterogeneous relationships in gene expression data, here we generalize the squared Pearson correlation to capture a mixture of linear dependences between two real-valued variables, with or without an index variable that specifies the line memberships. We construct the generalized Pearson correlation squares by focusing on three aspects: variable exchangeability, no parametric model assumptions, and inference of population-level parameters. To compute the generalized Pearson correlation square from a sample without a line-membership specification, we develop a *K*-lines clustering algorithm to find *K* clusters that exhibit distinct linear dependences, where *K* can be chosen in a data-adaptive way. To infer the population-level generalized Pearson correlation squares, we derive the asymptotic distributions of the sample-level statistics to enable efficient statistical inference. Simulation studies verify the theoretical results and show the power advantage of the generalized Pearson correlation squares in capturing mixtures of linear dependences. Gene expression data analyses demonstrate the effectiveness of the generalized Pearson correlation squares and the *K*-lines clustering algorithm in dissecting complex but interpretable relationships. The estimation and inference procedures are implemented in the R package `gR2` (*https://github.com/lijy03/gR2*). Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

ARTICLE HISTORY
Received January 2022
Accepted April 2024

KEYWORDS
Asymptotic distribution; Mixture of linear dependences; Specified and unspecified generalized Pearson correlation squares

1. Introduction

In biomedical research, Pearson correlation and its rank-based variant Spearman correlation remain the most widely used association measures for describing the relationship between two scalar-valued variables, for example, two genes' expression levels. The reason underlying the two measures' popularity is 2-fold: linear and monotone relationships[1] are widespread in nature and interpretable to researchers. In many cases, however, the interesting relationship between two variables often depends on another hidden categorical variable.

For example, in a gene expression dataset of *Arabidopsis thaliana*, a plant model organism, many genes exhibit different linear dependences between root and shoot tissues (Li et al. 2008; Kim et al. 2012). Figure 1(A) shows pairwise relationships of flavin-monooxygenase (FMO) genes' expression levels, and all these relationships differ between root and shoot tissues. In particular, FMO GS-OX2 and FMO GS-OX5 exhibit a positive (sample-level Pearson) correlation in shoots (black dots) but a negative correlation in roots (gray circles). Imagine an idealistic, extreme scenario (Figure 1(B)) where two genes have a positive (population-level Pearson) correlation $\rho \in (0, 1)$ in the shoot tissue but a negative correlation $-\rho$ in the root tissue, and the two tissues are equally sampled; then the two genes would have a zero correlation if not conditional on the tissue.

Real scenarios are usually not so extreme, but many of them exhibit a mixture relationship composed of two linear dependences (Li 2002), and they may show the "Simpson's Paradox" where the overall correlation and the conditional correlations have opposite signs. Under such scenarios, Pearson correlation is a misleading measure, as it specifically looks for a single linear dependence. These scenarios often lack an index variable (e.g., the shoot/root tissue type) that segregates observations into distinct linear relationships. Moreover, numerous variable pairs (e.g., 10^8 gene pairs) often need to be examined to discover unknown, but interesting and interpretable associations. Therefore, an association measure is in much demand to capture such relationships that are decomposable into a (possibly unknown) number of linear dependences, in a powerful and efficient way.

In the literature of scalar-valued association measures (also known as dependence measures), many measures have been developed to capture dependent relationships more general than the linear dependence. The first type of measures aims to capture more general functional (i.e., one-to-one) relationships. For monotone relationships, the Spearman's rank correlation and the Kendall's τ are commonly used. For functional relationships more general than monotonicity, there are measures including the maximal correlation efficient, measures based on nonparametric estimation of correlation curves (Bjerve and Doksum 1993) or principal curves (Delicado and Smrekar 2009),

CONTACT Jingyi Jessica Li ✉ *jli@stat.ucla.edu* Department of Statistics and Data Science, University of California, Los Angeles, Los Angeles, CA.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

[1] Monotone relationships becomes linear after values of each variable are transformed into ranks.

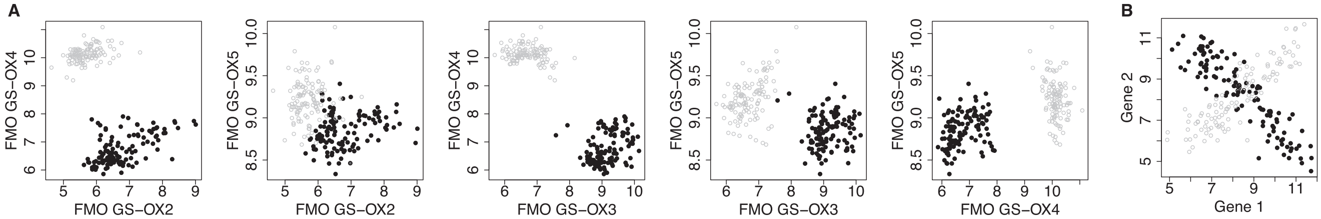

Figure 1. **A**: Pairwise expression levels of *A. thaliana* genes. **B**: A simulated toy example. Gray circles and black dots indicate data from root and shoot tissues, respectively.

generalized measures of correlation that deals with asymmetrically explained variances and nonlinear relationships (Zheng, Shi, and Zhang 2012), measures for detecting local monotone patterns using count statistics (Wang, Waterman, and Huang 2014), the G^2 statistic derived from a regularized likelihood ratio test for piecewise-linear relationships (Wang, Jiang, and Liu 2017), and the recently proposed Chatterjee's rank correlation ξ (Chatterjee 2021). The second type of measures aims to capture general dependence so that they only give zero values to independent random variable pairs. Examples include the maximal correlation coefficient and Chatterjee's rank correlation ξ, which also belong to the first type, and other measures including the Hoeffding's D, the mutual information, kernel-based measures such as the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al. 2005), the distance correlation (Székely, Rizzo, and Bakirov 2007; Székely and Rizzo 2009) and its generalization as the multiscale graph correlation (MGC) (Shen, Priebe, and Vogelstein 2019), the maximal information coefficient (Reshef et al. 2011), the Heller-Heller-Gorfine (HHG) association test statistic based on ranks of distances (Heller, Heller, and Gorfine 2012), and the semiparametric kernel independence test for handling excess zeros (Lee and Zhu 2021). Specifically, the following measures are not restricted to comparing real-valued random variables: the Hoeffding's D, the mutual information, the HSIC, the HHG test statistic, the distance correlation, and the MGC, among which the first four measures have the range $[0, \infty)$ instead of having absolute values under 1.

The aforementioned two types of measures have complementary advantages and disadvantages. Measures of the first type are generally interpretable but cannot capture the widespread nonfunctional (i.e., not one-to-one) relationships. In contrast, measures of the second type, though being versatile and having desirable theoretical properties, do not provide a straightforward interpretation of their captured relationships. As Figure 1 shows, many relationships are decomposable into a small number of linear dependences. Since the linear dependence is the simplest and most interpretable relationship, a mixture of a small number of linear dependences is also interpretable and often of great interest in biomedical research. For example, if researchers observe that a gene positively regulates a vital cancer gene in one cancer subtype but exhibits adverse regulatory effects in another subtype, different treatment strategies may be designed for the two cancer subtypes. However, mixtures of linear dependences remain challenging to capture: they are often missed by the first type of measures and cannot be distinguished from other less interpretable relationships by the second type of measures. Although mixtures of linear regression models have been of broad interest in fields including statistics, economics, social sciences, and machine learning for over 40 years (Quandt and Ramsey 1978; Murtaph and Raftery 1984; De Veaux 1989; Jacobs et al. 1991; Jones and McLachlan 1992; Wedel and DeSarbo 1994; Turner 2000; Hawkins, Allen, and Stromberg 2001; Hurn, Justel, and Robert 2003; Leisch 2008; Benaglia et al. 2009; Scharl, Grün, and Leisch 2009), they did not propose an association measure to capture mixtures of linear dependences, and neither do they trivially lead to a reasonable association measure, as we will explain below.

In this work, we propose generalized Pearson correlation squares, for which the squared Pearson correlation is a special case, to capture a mixture of linear dependences. Our proposal addresses the practical need to screen variable pairs that exhibit complex yet interpretable relationships. These relationships can be decomposed into linear components, where at least one component demonstrates statistical significance. We consider two scenarios: the *specified scenario* where an index variable indicates the line membership of each observation, and the more common *unspecified scenario* where no index variable is available. Under the specified scenario, we aim for our new measure to quantify the "informativeness" of the index variable, indicating whether it specifies distinct linear relationships. To achieve a reasonable generalization of the Pearson correlation, our new measures adhere to an essential property embraced by most existing association measures—the exchangeability of the two variables.[2] Additionally, we seek to provide both population-level parameters and sample-level statistics for our measures, enabling statistical inference and the assessment of statistical significance for observed measure values.

First, under the unspecified scenario, can we directly use the existing work on mixtures of linear regression models? The answer is no because these models require a specification of the response and predictor variables; that is, they do not consider the two variables symmetric. Except for the degenerate case where only one linear component exists, that is, the linear model, these models do not lead to a measure exchangeable for the two variables.

Second, still under the unspecified scenario, how to assign observations to lines to ensure the exchangeability? A good assignment should be able to handle general cases where observations from each line do not follow a specific distribution

[2] Note that a linear model with group-specific slopes and intercepts (groups specified by the index variable Z) does not provide a desirable measure under the specified scenario. The reason is that the exchangeability is not satisfied: the linear model R^2 would be different if we swap the two variables X and Y; that is, the two linear models (written in R commands) `lm(Y ~ Z + X:Z)` and `lm(X ~ Z + Y:Z)` do not give the same R^2.

(as is required by model-based clustering) or have a spherical shape (as is required by the K-means clustering). To handle such cases, we propose the K-lines clustering algorithm in Section 2.2.1.

Third, how to define population-level measures to enable proper inference? A critical point is that the specified and unspecified scenarios need different population-level measures; otherwise, it would be impossible to construct unbiased estimators for both scenarios without distributional assumptions. We will elaborate on this point in Section 3.

Fourth, when an index variable is available, should we always use it to specify line memberships? Surprisingly, the answer is no because the index variable may be uninformative or irrelevant to the segregation of lines. In that case, it can be more informative to directly estimate line memberships from data by clustering. We will demonstrate this point using real data in Section 5.1.

This article is organized as follows. In Section 2, we define generalized Pearson correlation squares at the population and sample levels, under the line-membership specified and unspecified scenarios. For the unspecified scenario, we develop a K-lines clustering algorithm, following the subspace clustering literature (Vidal 2010). In Section 3, we derive the asymptotic distributions of the corresponding sample-level measures to enable efficient statistical inference. In Section 4, we conduct simulation studies under various settings to verify the asymptotic distributions and evaluate the finite-sample statistical power of the proposed measures. In Section 5, we demonstrate the use of the generalized Pearson correlation squares and the K-lines clustering algorithm for dissecting gene expression heterogeneity in two real datasets, followed by discussions in Section 6. Supplementary material includes all the proofs of lemmas and theorems, convergence properties of the K-lines algorithm, more simulation results, another real data application, real data description, more tables and figures, and additional references.

2. Generalized Pearson Correlation Squares and K-Lines Clustering Algorithm

The Pearson correlation is the most widely used measure to describe the relationship between two random variables $X, Y \in \mathbb{R}$. At the population level, the Pearson correlation of X and Y is defined as $\rho = \mathrm{cov}(X, Y)/\{\mathrm{var}(X)\mathrm{var}(Y)\}^{-1/2} \in [-1, 1]$, where $\mathrm{cov}(X, Y) = \mathbb{E}[\{X - \mathbb{E}(X)\}\{Y - \mathbb{E}(Y)\}]$, $\mathrm{var}(X) = \mathbb{E}\{[X - \mathbb{E}(X)]^2\}$, and $\mathrm{var}(Y) = \mathbb{E}\{[Y - \mathbb{E}(Y)]^2\}$ denote the covariance between X and Y, the variance of X, and the variance of Y, respectively. We say that X and Y are linearly dependent if $\rho \neq 0$.

At the sample level, the Pearson correlation $R = \{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})\}/\{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2\}^{1/2}$ is defined based on a sample $\{(X_i, Y_i)\}_{i=1}^{n}$ from the joint distribution of (X, Y), where $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$ and $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$. Motivated by the fact that R^2, the Pearson correlation square, is commonly used to describe the observed linear dependence in a bivariate sample, we develop generalized Pearson correlation squares to capture a mixture of linear dependences.

We define the *line-membership specified scenario* as the case where we also observe an index random variable $Z \in \{1, \ldots, K\}$ that specifies the linear dependence between X and Y, and K is the number of linear dependences. In parallel, we define the *line-membership unspecified scenario* as the case where no index variable is available. In the special case of $K = 1$, we have $Z \equiv 1$.

There may exist more than one index variable, and correspondingly there could be multiple specified scenarios. For example, in the *A. thaliana* gene expression dataset (Table S4 in the supplementary material), there are four index variables (`condition`, `treatment`, `replicate`, and `tissue`) that correspond to four different specified scenarios. As we will show in Section 5.1, only the `tissue` specification leads to a reasonable separation of linear relationships (Figures 1 and S5–S8 in the supplementary material). Hence, a specified scenario is not always preferred to the unspecified scenario when the goal is to capture an informative mixture of linear dependences.

2.1. Line-Membership Specified Scenario

2.1.1. Population-Level Generalized Pearson Correlation Square (Specified)

As the index variable Z is observable under the line-membership specified scenario, we denote $p_{kS} = \mathbb{P}(Z = k)$, $k = 1, \ldots, K$. Conditional on $Z = k$, the population-level Pearson correlation between X and Y is $\rho_{kS} = \mathrm{cov}(X, Y \mid Z = k)/\{\mathrm{var}(X \mid Z = k)\mathrm{var}(Y \mid Z = k)\}^{1/2}$, if $\mathrm{var}(X \mid Z = k) > 0$ and $\mathrm{var}(Y \mid Z = k) > 0$; otherwise, $\rho_{kS} = 0$. In the special case of $K = 1$, $\rho_{1S}^2 = \rho^2 = \mathrm{cov}^2(X, Y)/\{\mathrm{var}(X)\mathrm{var}(Y)\}$ is the population-level Pearson correlation square that indicates the population-level strength of a linear dependence. Motivated by this, we combine $\rho_{1S}^2, \ldots, \rho_{KS}^2$ into one measure to indicate the overall strength of K linear dependences.

Definition 2.1. At the population level, when the line membership variable Z is specified, the *generalized Pearson correlation square* between X and Y is defined as

$$\rho_{GS}^2 = \mathbb{E}_Z\left(\rho_{ZS}^2\right) = \mathbb{E}_Z\left\{\frac{\mathrm{cov}^2(X, Y \mid Z)}{\mathrm{var}(X \mid Z)\mathrm{var}(Y \mid Z)}\right\} = \sum_{k=1}^{K} p_{kS}\, \rho_{kS}^2, \tag{2.1}$$

a weighted sum of $\rho_{1S}^2, \ldots, \rho_{KS}^2$, that is, the strengths of the K linear dependences, with weights as $p_{1S}, \ldots, p_{KS}$. Note that the subindex G stands for generalized, and S stands for specified.

2.1.2. Sample-Level Generalized Pearson Correlation Square (Specified)

To estimate ρ_{GS}^2, we consider a sample $\{(X_i, Y_i, Z_i)\}_{i=1}^{n}$ from the joint distribution of $(X, Y, Z) \in \mathbb{R}^2 \times \{1, \ldots, K\}$.

Definition 2.2. At the sample level, when observations $\{(X_i, Y_i)\}_{i=1}^{n}$ have line memberships $\{Z_i\}_{i=1}^{n}$ specified, the *generalized Pearson correlation square* is defined as

$$R_{GS}^2 = \sum_{k=1}^{K} \widehat{p}_{kS}\, \widehat{\rho}_{kS}^2, \tag{2.2}$$

where the subindex G stands for generalized, S stands for specified, $\widehat{p}_{kS} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(Z_i = k)$,

$$\widehat{\rho}_{kS}^2 = \frac{\left\{\sum_{i=1}^n (X_i - \bar{X}_{kS})(Y_i - \bar{Y}_{kS})\mathbb{I}(Z_i = k)\right\}^2}{\left\{\sum_{i=1}^n (X_i - \bar{X}_{kS})^2 \mathbb{I}(Z_i = k)\right\}\left\{\sum_{i=1}^n (Y_i - \bar{Y}_{kS})^2 \mathbb{I}(Z_i = k)\right\}},$$

$\bar{X}_{kS} = n_{kS}^{-1} \sum_{i=1}^n X_i \mathbb{I}(Z_i = k)$, $\bar{Y}_{kS} = n_{kS}^{-1} \sum_{i=1}^n Y_i \mathbb{I}(Z_i = k)$, and $n_{kS} = \sum_{i=1}^n \mathbb{I}(Z_i = k)$.

The R_{GS}^2 measure is a weighted sum of the R^2's of all line components, that is, $\widehat{\rho}_{1S}^2, \ldots, \widehat{\rho}_{KS}^2$. Note that X and Y are exchangeable in R_{GS}^2. We next define the counterpart of R_{GS}^2 under the more common scenario in which no index variable Z is observable.

2.2. Line-Membership Unspecified Scenario

Under the line-membership unspecified scenario, we investigate a mixture of K linear dependences between X and Y without observing any index variable Z. For this scenario, we start with formulating the sample-level measure as a counterpart of R_{GS}^2 because generalization from the specified scenario to the unspecified scenario is more straightforward at the sample level than the population level. We consider a sample $\{(X_i, Y_i)\}_{i=1}^n$ from the joint distribution of $(X, Y) \in \mathbb{R}^2$.

2.2.1. K-Lines Clustering Algorithm

As no line-membership information is available, we will first assign each (X_i, Y_i) to a line. Because we would like X and Y to be exchangeable in the new measure, a reasonable way is to assign (X_i, Y_i) to the closest line in the perpendicular distance. We use a shorthand notation $\beta = (\theta, c)^\mathsf{T}, \theta \in [0, 2\pi]$ and $c \in \mathbb{R}$, to denote the line $\{(x, y)^\mathsf{T} : \cos\theta \cdot x + \sin\theta \cdot y - c = 0\} \subset \mathbb{R}^2$. The perpendicular distance from $(x, y)^\mathsf{T}$ to β is

$$d_\perp\left((x, y)^\mathsf{T}, \beta\right) = |\cos\theta \cdot x + \sin\theta \cdot y - c| \,. \tag{2.3}$$

Then we define the *sample-level unspecified line centers* as the K lines that minimize the average squared perpendicular distance of data points to their closest line.

Definition 2.3. Let $B_K = \{\beta_1, \ldots, \beta_K\}$ be a multiset of K lines with possible repeats. We define the average within-cluster squared perpendicular distance as

$$W(B_K, P_n) = \frac{1}{n}\sum_{i=1}^n \min_{\beta \in B_K} d_\perp^2\left((X_i, Y_i)^\mathsf{T}, \beta\right), \tag{2.4}$$

where P_n is the empirical measure that places mass n^{-1} at each of $(X_1, Y_1), \ldots, (X_n, Y_n)$. Then we define the multiset of *sample-level unspecified line centers* as

$$\widehat{B}_{KU} \in \arg\min_{B_K} W(B_K, P_n)\,, \tag{2.5}$$

where the subindex U stands for unspecified. We write each solution to (2.5) as $\widehat{B}_{KU} = \{\widehat{\beta}_{1U}, \ldots, \widehat{\beta}_{KU}\}$, where $\widehat{\beta}_{kU} = (\widehat{\theta}_{kU}, \widehat{c}_{kU})^\mathsf{T}$ is the kth line center.

To find $\widehat{B}_{KU}$, we propose the *K-lines clustering algorithm*, which is inspired by the well-known K-means algorithm (Lloyd 1982). The K-means algorithm cannot account for within-cluster correlation structures but can only identify spherical clusters under a distance metric, for example, the Euclidean distance. In contrast, the K-lines algorithm finds clusters that exhibit strong within-cluster correlations; it is specifically designed for applications where two real-valued variables have distinct correlations in different hidden clusters. We note that the K-lines algorithm is a special case of subspace clustering (Vidal 2010).

As an iterative procedure, the K-lines clustering algorithm includes two alternating steps in each iteration. The *recentering step* uses the current cluster assignment (i.e., line memberships) to update each cluster line center, which minimizes the within-cluster sum of squared perpendicular distances of data points to the line center. The *assignment step* updates the cluster assignment based on the current cluster line centers: assign every data point to its closest cluster line center in the perpendicular distance. The two steps alternate until the algorithm converges. Figure 2 illustrates the K-lines clustering algorithm.

The recentering step updates each cluster center using the *major axis regression*, which minimizes the sum of squares of the perpendicular distances from points to the regression line. The major axis regression line is the first principal component of the two variables' sample covariance matrix (Jolliffe 1982; Smith 2009). Given the cluster assignment in the $(t-1)$th iteration: $\mathcal{C}_1^{(t-1)}, \ldots, \mathcal{C}_K^{(t-1)}$, the updated kth cluster center is

$$\begin{aligned}\widehat{\beta}_{kU}^{(t)} &= \arg\min_\beta \sum_{i \in \mathcal{C}_k^{(t-1)}} d_\perp^2\left((X_i, Y_i)^\mathsf{T}, \beta\right) \\ &= \left(\widehat{\theta}_{kU}, \widehat{u}_{12,k}\bar{X}_{kU} - \widehat{u}_{11,k}\bar{Y}_{kU}\right)^\mathsf{T}\,, \end{aligned} \tag{2.6}$$

where $\cos\widehat{\theta}_{kU} = \widehat{u}_{12,k}$, $\sin\widehat{\theta}_{kU} = -\widehat{u}_{11,k}$, and $(\widehat{u}_{11,k}, \widehat{u}_{12,k})^\mathsf{T}$ is the first eigenvector of the sample covariance matrix

$$\left|\mathcal{C}_k^{(t-1)}\right|^{-1}\begin{bmatrix}\sum_{i\in\mathcal{C}_k^{(t-1)}}(X_i - \bar{X}_{kU})^2 & \sum_{i\in\mathcal{C}_k^{(t-1)}}(X_i - \bar{X}_{kU})(Y_i - \bar{Y}_{kU}) \\ \sum_{i\in\mathcal{C}_k^{(t-1)}}(X_i - \bar{X}_{kU})(Y_i - \bar{Y}_{kU}) & \sum_{i\in\mathcal{C}_k^{(t-1)}}(Y_i - \bar{Y}_{kU})^2\end{bmatrix},$$

with $\bar{X}_{kU} = |\mathcal{C}_k^{(t-1)}|^{-1}\sum_{i\in\mathcal{C}_k^{(t-1)}} X_i$ and $\bar{Y}_{kU} = |\mathcal{C}_k^{(t-1)}|^{-1} \sum_{i\in\mathcal{C}_k^{(t-1)}} Y_i\,.$

Similar to the K-means clustering algorithm, the K-lines clustering algorithm is not guaranteed to find the global minimizer, $\arg\min_{B_K} W(B_K, P_n)$. Empirically, we run the K-lines clustering algorithm for M times with random initializations and obtain M multisets of unspecified line centers $B_K^{(1)}, \ldots, B_K^{(M)}$. Then we set $\widehat{B}_{KU} \in \arg\min_{B_K \in \{B_K^{(1)}, \ldots, B_K^{(M)}\}} W(B_K, P_n)$. Regarding the effects of M, see Section B.3 in the supplementary material. In the R package gR2, the default setting is $M = 30$ if $n \geq 50$, and $M = \lfloor 1500/n \rfloor$ if $n < 50$.

Remark 2.1. **Normalization.** Regarding whether $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ should be separately normalized before the K-lines algorithm is applied, the decision is problem-specific and depending on the scales of X and Y, same as for K-means clustering.

Remark 2.2. **Computational complexity.** The complexity of the K-lines algorithm is $O(nKT)$, the same as Lloyd's implementation of the K-means algorithm for a given initialization with T iterations. The reason is that the K-lines algorithm and Loyld's K-means algorithm have only two differences: (a) calculation of

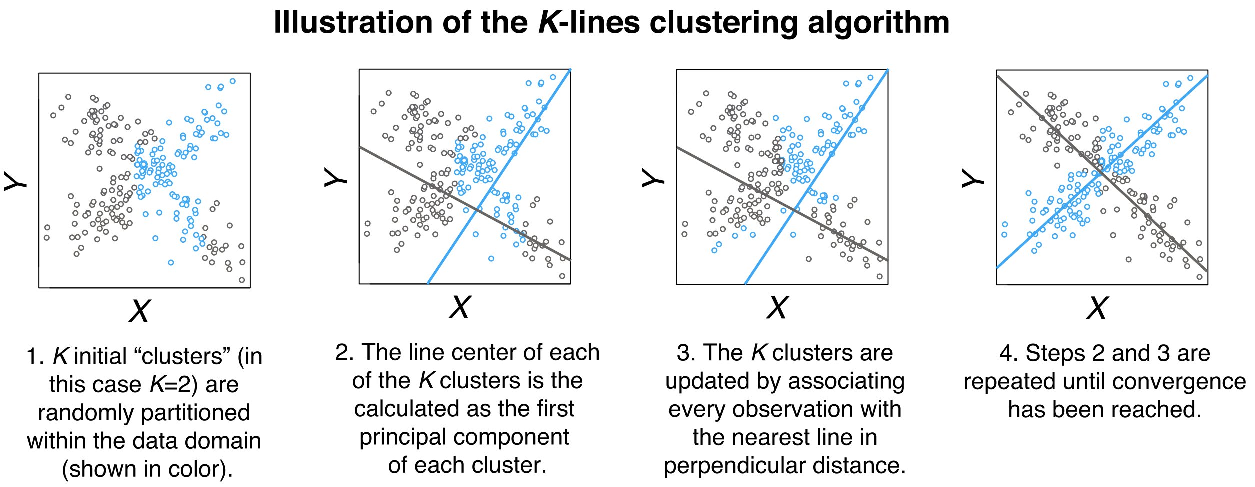

Figure 2. An illustration of the *K*-lines clustering algorithm.

```
Initialization: Assign random initial clusters $\mathcal{C}_1^{(0)}, \dots, \mathcal{C}_K^{(0)}$, such that $\cup_{k=1}^K \mathcal{C}_k^{(0)} = \{1, \dots, n\}$.
The algorithm proceeds by alternating between two steps. In the $t$th iteration, $t = 1, 2, \dots$
Recentering step: Calculate the cluster line centers $\widehat{\beta}_{1U}^{(t)}, \dots, \widehat{\beta}_{KU}^{(t)}$ based on the cluster assignment $\mathcal{C}_1^{(t-1)}, \dots, \mathcal{C}_K^{(t-1)}$ by (2.6).
Assignment step: Update the cluster assignment for $k = 1, \dots, K$
$\mathcal{C}_k^{(t)} = \left\{ i : d_\perp\left((X_i, Y_i)^\top, \widehat{\beta}_{kU}^{(t)}\right) \le d_\perp\left((X_i, Y_i)^\top, \widehat{\beta}_{sU}^{(t)}\right), \text{ for all } s = 1, \dots, K \right\}$.
Stop the iteration when the cluster assignment no longer changes.
Output: Cluster assignment $\mathcal{C}_1, \dots, \mathcal{C}_K$; Sample-level unspecified line centers $\widehat{\beta}_{1U}, \dots, \widehat{\beta}_{KU}$.
```

Algorithm 1: *K*-lines clustering algorithm

K cluster centers, whose complexity is $O(2^2 n + 2^3) = O(n)$ for finding the first principal components in K-lines versus $O(n)$ for calculating the arithmetic means in K-means; (b) calculation of distances from data points to cluster centers, whose complexity is $O(n)$ in both K-means and K-lines.

Remark 2.3. **Data-driven choice of K.** When users do not have prior knowledge about the value of K, how to choose K becomes an important question in practice. Some methods for choosing K in K-means clustering can be adapted. For example, the elbow method, though not theoretically principled, is visually appealing to practitioners and widely used. It employs a scree plot whose horizontal axis displays a range of K values, and whose vertical axis shows the average within-cluster sum of squared distances corresponding to each K. For our K-lines algorithm, it is reasonable to use a scree plot to show how $W(B_K, P_n)$, the average within-cluster squared perpendicular distance defined in (2.4), decreases as K increases.

Alternatively, when it is reasonable to assume that $(X, Y) \mid (Z = k)$ follows a bivariate Gaussian distribution for all $k = 1, \dots, K$, one may use the Akaike information criterion (AIC) to choose K. Specifically, AIC is defined as

$$\mathrm{AIC}(K) = 2(6K-1) - 2\sum_{i=1}^{n} \log p\left(X_i, Y_i \mid \{\widehat{p}_{kU}, \widehat{\mu}_{kU}, \widehat{\Sigma}_{kU}\}_{k=1}^{K}\right)$$

$$= 2(6K-1) - \pi^{-1}\sum_{i=1}^{n} \log\left[\sum_{k=1}^{K} \widehat{p}_{kU}\left|\widehat{\Sigma}_{kU}\right|^{-1/2} \exp\left\{-\frac{1}{2}\left((X_i, Y_i)^\top - \widehat{\mu}_{kU}\right)^\top \widehat{\Sigma}_{kU}^{-1}\left((X_i, Y_i)^\top - \widehat{\mu}_{kU}\right)\right\}\right], \qquad (2.7)$$

where the first term is $2(6K-1)$ because there are 6 parameters for each component and the component proportions sum to 1; in the second term, $\widehat{p}_{kU} = |\mathcal{C}_k|/n$, $\widehat{\mu}_{kU} = \left(\bar{X}_{kU}, \bar{Y}_{kU}\right)^\top$,

$$\widehat{\Sigma}_{kU} = |\mathcal{C}_k|^{-1} \begin{bmatrix} \sum_{i \in \mathcal{C}_k}(X_i - \bar{X}_{kU})^2 & \sum_{i \in \mathcal{C}_k}(X_i - \bar{X}_{kU})(Y_i - \bar{Y}_{kU}) \\ \sum_{i \in \mathcal{C}_k}(X_i - \bar{X}_{kU})(Y_i - \bar{Y}_{kU}) & \sum_{i \in \mathcal{C}_k}(Y_i - \bar{Y}_{kU})^2 \end{bmatrix}.$$

We will demonstrate the elbow method and the AIC method in Section 4.2.

2.2.2. Sample-Level Generalized Pearson Correlation Square (Unspecified)

Powered by the K-lines algorithm, we introduce the sample surrogate indices $\widehat{Z}_i \in \{1, \dots, K\}$, $i = 1, \dots, n$, based on which we then define the sample-level generalized Pearson correlation square for this line-membership unspecified scenario.

Definition 2.4. Suppose that Algorithm 1 outputs K unspecified line centers $\widehat{\beta}_{1U}, \dots, \widehat{\beta}_{KU}$. Also suppose that the probability that (X_i, Y_i) is equally close to more than one line center is zero. For each (X_i, Y_i), we define its *sample surrogate index*

$$\widehat{Z}_i = \operatorname*{arg\,min}_{k \in \{1, \dots, K\}} d_\perp\left((X_i, Y_i)^\top, \widehat{\beta}_{kU}\right), \; i = 1, \dots, n. \qquad (2.8)$$

Definition 2.5. At the sample level, when observations $\{(X_i, Y_i)\}_{i=1}^n$ have line memberships unspecified, the *generalized Pearson correlation square* is defined as

$$R_{GU}^2 = \sum_{k=1}^{K} \widehat{p}_{kU} \cdot \widehat{\rho}_{kU}^2, \qquad (2.9)$$

where G stands for generalized, U stands for unspecified, $\widehat{p}_{kU} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(\widehat{Z}_i = k)$,

$$\widehat{\rho}^2_{kU} = \frac{\left\{\sum_{i=1}^{n}\left(X_i - \bar{X}_{kU}\right)\left(Y_i - \bar{Y}_{kU}\right)\mathbb{I}(\widehat{Z}_i = k)\right\}^2}{\left\{\sum_{i=1}^{n}\left(X_i - \bar{X}_{kU}\right)^2 \mathbb{I}(\widehat{Z}_i = k)\right\}\left\{\sum_{i=1}^{n}\left(Y_i - \bar{Y}_{kU}\right)^2 \mathbb{I}(\widehat{Z}_i = k)\right\}},$$

$\bar{X}_{kU} = n_{kU}^{-1}\sum_{i=1}^{n} X_i \mathbb{I}(\widehat{Z}_i = k)$, $\bar{Y}_{kU} = n_{kU}^{-1}\sum_{i=1}^{n} Y_i \mathbb{I}(\widehat{Z}_i = k)$, and $n_{kU} = \sum_{i=1}^{n} \mathbb{I}(\widehat{Z}_i = k)$.

Remark 2.4. Besides R^2_{GU}, the cluster-specific Pearson correlation squares $\widehat{\rho}^2_{1U}, \ldots, \widehat{\rho}^2_{KU}$ are useful for identifying the clusters (subgroups of data points) that exhibit distinct, strong linear dependences.

2.2.3. Population-Level Generalized Pearson Correlation Square (Unspecified)

Analogous to the definition of sample-level unspecified line centers (Definition 2.3), we define the *population-level unspecified line centers* as the K lines that minimize the expected squared perpendicular distance of (X, Y) to its closest line.

Definition 2.6. We define the expected within-cluster squared perpendicular distance as

$$W(B_K, P) = \mathbb{E}\left\{\min_{\beta \in B_K} d^2_{\perp}\left((X, Y)^{\mathsf{T}}, \beta\right)\right\}, \qquad (2.10)$$

where P is the joint probability measure of (X, Y). Then we define a multiset of *population-level unspecified line centers*, $B_{KU} = \{\beta_{1U}, \ldots, \beta_{KU}\}$, where $\beta_{kU} = (\theta_{kU}, c_{kU})^{\mathsf{T}}$ is the kth line center, as

$$B_{KU} \in \underset{B_K}{\arg\min}\, W(B_K, P), \qquad (2.11)$$

where the subindex U stands for unspecified.

Provided that B_{KU} is uniquely determined, we define a random surrogate index $\widetilde{Z} \in \{1, \ldots, K\}$ as the index of the line center to which (X, Y) is closest.

Definition 2.7. Suppose that the unspecified line centers $\beta_{1U}, \ldots, \beta_{KU}$ at the population level are unique. Also, suppose that the probability that (X, Y) is equally close to multiple line centers is zero. We define a *random surrogate index* as

$$\widetilde{Z} = \underset{k \in \{1,\ldots,K\}}{\arg\min}\, d_{\perp}\left((X, Y)^{\mathsf{T}}, \beta_{kU}\right). \qquad (2.12)$$

Motivated by ρ^2_{GS}, we define the population-level generalized Pearson correlation square for the line-membership unspecified scenario, based on $(X, Y, \widetilde{Z})$.

Definition 2.8. At the population level, when no line membership variable is specified (the "unspecified scenario"), the *generalized Pearson correlation square* between X and Y is

$$\rho^2_{GU} = \sum_{k=1}^{K} p_{kU}\, \rho^2_{kU}, \qquad (2.13)$$

where the subindex G stands for generalized, U stands for unspecified, $p_{kU} = \mathbb{P}(\widetilde{Z} = k)$, and $\rho^2_{kU} = \mathrm{cov}^2(X, Y \mid \widetilde{Z} = k)/\{\mathrm{var}(X \mid \widetilde{Z} = k)\,\mathrm{var}(Y \mid \widetilde{Z} = k)\}$.

Remark 2.5. Relations and distinctions between the specified and unspecified scenarios.

1. $\rho^2_{GU} \geq \rho^2_{GS}$. The proof is in the supplementary material.
2. R^2_{GU} is not an estimator of ρ^2_{GS}; rather, it is an estimator of ρ^2_{GU}. Hence, the consistency of R^2_{GU} does not rely on a specific distributional assumption, for example, bivariate Gaussian mixture model, just like the R^2. If the goal were to use R^2_{GU} as an estimator of ρ^2_{GS}, a specific mixture model must be assumed. Then the K-lines algorithm should be replaced by the Expectation-Maximization (EM) algorithm to decide the sample surrogate indices $\widehat{Z}_1, \ldots, \widehat{Z}_n$. When the EM algorithm converges to the global optimum and returns the maximum-likelihood estimates of mixture model parameters, the corresponding R^2_{GU} will be an asymptotically unbiased estimator of ρ^2_{GS}.

3. Asymptotic Distributions of Sample-Level Generalized Pearson Correlation Squares

To enable statistical inference of the population-level measures ρ^2_{GS} (2.1) and ρ^2_{GU} (2.13), we derive the first-order asymptotics of the sample-level measures R^2_{GS} (2.2) and R^2_{GU} (2.9). In the asymptotic results below, we consider all parameters as fixed and only allow the sample size n to go to infinity.

Theorem 3.1. Under the line-membership specified scenario, we define

$$\mu_{X^cY^d,kS} = \mathbb{E}\left\{\left(\frac{X - \mathbb{E}(X \mid Z = k)}{\mathrm{var}(X \mid Z = k)^{1/2}}\right)^c \left(\frac{Y - \mathbb{E}(Y \mid Z = k)}{\mathrm{var}(Y \mid Z = k)^{1/2}}\right)^d \,\middle|\, Z = k\right\}, \quad c, d \in \mathbb{N}.$$

Assume $\mu_{X^4,kS} < \infty$ and $\mu_{Y^4,kS} < \infty$ for all $k = 1, \ldots, K$. Then

$$\sqrt{n}\left(R^2_{GS} - \rho^2_{GS}\right) \xrightarrow{d} N\left(0, \sum_{k=1}^{K}(A_{kS} + B_{kS}) + 2\sum_{1 \leq k < r \leq K}\sum C_{krS}\right), \qquad (3.1)$$

where

$$A_{kS} = p_{kS}\left[\rho^4_{kS}\left(\mu_{X^4,kS} + 2\mu_{X^2Y^2,kS} + \mu_{Y^4,kS}\right) - 4\rho^3_{kS}\left(\mu_{X^3Y,kS} + \mu_{XY^3,kS}\right) + 4\rho^2_{kS}\mu_{X^2Y^2,kS}\right],$$

$$B_{kS} = p_{kS}\left(1 - p_{kS}\right)\rho^4_{kS}, \quad \text{and} \quad C_{krS} = -p_{kS}\,p_{rS}\,\rho^2_{kS}\,\rho^2_{rS}.$$

Note that Theorem 3.1 does not rely on any distributional assumptions. When it is applied to the special case where $(X, Y) \mid Z$ follows a bivariate Gaussian distribution, we obtain a much simpler form of the first-order asymptotic distribution of R^2_{GS}.

Corollary 3.1. Under the special case where $(X, Y) \mid (Z = k)$ follows a bivariate Gaussian distribution for all $k = 1, \ldots, K$, the asymptotic variance of $\sqrt{n}(R^2_{GS} - \rho^2_{GS})$ in Theorem 3.1 is simplified and becomes

$$\sum_{k=1}^{K}\left[4\,p_{kS}\,\rho_{kS}^2\left(1-\rho_{kS}^2\right)^2+p_{kS}\left(1-p_{kS}\right)\rho_{kS}^4\right] \tag{3.2}$$

$$-2\sum\sum_{1\le k<r\le K} p_{kS}\,p_{rS}\,\rho_{kS}^2\,\rho_{rS}^2\,, \tag{3.3}$$

which only depends on p_{kS} and ρ_{kS}^2, $k=1,\ldots,K$.

To derive an analog of Theorem 3.1 and Corollary 3.1 for the unspecified scenario, we need to show that each sample surrogate index $\widehat{Z}_i$, $i=1,\ldots,n$, converges in distribution to the random surrogate index $\widetilde{Z}$. A sufficient condition is the strong consistency of the K sample-level unspecified line centers $\widehat{B}_{KU}=\{\widehat{\beta}_{1U},\ldots,\widehat{\beta}_{KU}\}$ to the K population-level unspecified line centers $B_{KU}=\{\beta_{1U},\ldots,\beta_{KU}\}$.

Theorem 3.2. Suppose that $\int \left\|(x,y)^{\mathsf{T}}\right\|^2 \mathbb{P}\left((dx,dy)^{\mathsf{T}}\right) < \infty$ and that for each $k=1,\ldots,K$ there is a unique multiset $B_{kU}=\arg\min_{B_k} W(B_k,P)$. Also assume that the globally optimal sample-level unspecified line centers $\widehat{B}_{KU}=\arg\min_{B_K} W(B_K,P_n)$ is attained and unique. Then as $n\to\infty$, $\widehat{B}_{KU}\to B_{KU}$ and $W(\widehat{B}_{KU},P_n)\to W(B_{KU},P)$ almost surely.

The first statement of Theorem 3.2 means that there exists an ordering of the elements in $\widehat{B}_{KU}=\{\widehat{\beta}_{1U},\ldots,\widehat{\beta}_{KU}\}$ and $B_{KU}=\{\beta_{1U},\ldots,\beta_{KU}\}$ such that as the sample size $n\to\infty$,

$$\widehat{\beta}_{kU}\to\beta_{kU}\ \text{ almost surely},\ k=1,\ldots,K\,.$$

Based on Theorems 3.1 and 3.2, we derive the asymptotic distribution of R_{GU}^2.

Theorem 3.3. Under the line-membership unspecified scenario, we define

$$\mu_{X^cY^d,kU}=\mathbb{E}\left\{\left(\frac{X-\mathbb{E}[X\mid\widetilde{Z}=k]}{\operatorname{var}(X\mid\widetilde{Z}=k)^{1/2}}\right)^c \left(\frac{Y-\mathbb{E}[Y\mid\widetilde{Z}=k]}{\operatorname{var}(Y\mid\widetilde{Z}=k)^{1/2}}\right)^d\middle|\,\widetilde{Z}=k\right\},\ c,d\in\mathbb{N},$$

where $\widetilde{Z}$ is the random surrogate index defined in (2.12). Assume $\mu_{X^4,kU}<\infty$ and $\mu_{Y^4,kU}<\infty$ for all $k=1,\ldots,K$. Then

$$\sqrt{n}\left(R_{GU}^2-\rho_{GU}^2\right) \xrightarrow{d} N\left(0,\sum_{k=1}^{K}(A_{kU}+B_{kU})+2\sum\sum_{1\le k<r\le K}C_{krU}\right), \tag{3.4}$$

where

$$A_{kU}=p_{kU}\left[\rho_{kU}^4\left(\mu_{X^4,kU}+2\mu_{X^2Y^2,kU}+\mu_{Y^4,kU}\right)\right.$$
$$\left.-4\rho_{kU}^3\left(\mu_{X^3Y,kU}+\mu_{XY^3,kU}\right)+4\rho_{kU}^2\mu_{X^2Y^2,kU}\right],$$
$$B_{kU}=p_{kU}\left(1-p_{kU}\right)\rho_{kU}^4\,,\quad\text{and}\quad C_{krU}=-\,p_{kU}\,p_{rU}\,\rho_{kU}^2\,\rho_{rU}^2\,.$$

Corollary A.1 presents a simpler form of Theorem 3.3 under an unrealistic assumption that $(X,Y)\mid\widetilde{Z}$ follows a bivariate Gaussian distribution. Despite being unrealistic, this simpler form empirically works well in numerical simulations (Section 4.1).

Remark 3.1. When $K=1$, the asymptotic distributions of R_{GS}^2 and R_{GU}^2 in Theorems 3.1 and 3.3 both reduce to the asymptotic distribution of R^2. In the special case that $K=1$ and (X,Y) follows bivariate Gaussian distribution with correlation ρ, the asymptotic distribution of R_{GS}^2 in Corollary 3.1 reduces to $\sqrt{n}(R^2-\rho^2)\xrightarrow{d}N\left(0,4\rho^2\left(1-\rho^2\right)^2\right)$.

Proofs are in the supplementary material. In Section 4.1, we will numerically show that the asymptotic distribution in Theorem 3.3 works well when K is chosen by the AIC.

4. Numerical Simulations

In this section, we perform simulation studies to numerically verify the theoretical results in Section 3 and to compare our generalized Pearson correlation squares with multiple existing association measures in terms of statistical power. We also demonstrate the effectiveness of our proposed approaches for choosing K, the number of line components in the line-membership unspecified scenario.

4.1. Numerical Verification of Theoretical Results

We first compare the asymptotic distributions in Section 3 with numerically simulated finite-sample distributions under 8 settings (Table 1), where $(X,Y)\mid Z$ follows a bivariate Gaussian distribution under the first 4 settings and a bivariate t distribution under the latter 4 settings. Under each setting, we generate $B=1000$ samples with sizes $n=50$ or 100, calculate R_{GS}^2 and R_{GU}^2 on each sample, and compare the simulated finite-sample distributions of R_{GS}^2 and R_{GU}^2 to the corresponding asymptotic distributions. In the first 4 settings, the asymptotic distributions are from Corollaries 3.1 and A.1 (the bivariate Gaussian results); in the latter 4 settings, the asymptotic distributions are from Theorems 3.1 and 3.3 (the general results). The comparison results (Figure 3) show that the finite-sample distributions and the asymptotic results have good agreement, justifying the use of the asymptotic distributions for statistical inference of ρ_{GS}^2 or ρ_{GU}^2 on a finite sample.

In practice, K often needs to be found in a data-driven way under the line-membership unspecified scenario. To verify the behavior of R_{GU}^2 when K is chosen by the AIC in (2.7), we conduct another simulation study to compare R_{GU}^2's finite-sample distributions with the asymptotic distributions. The results (Fig. S1 in the supplementary material) show that when $n=100$, finite-sample and asymptotic distributions still agree well.

However, the asymptotic distributions in Section 3 involve unobservable parameters in the asymptotic variance terms. A classical solution is to plug-in estimates of these parameters. Another common inferential approach is to use the bootstrap, which is computationally more intensive, instead of the closed-form asymptotic distributions. Here we numerically verify whether the plug-in approach works reasonably well for statistical inference of ρ_{GS}^2 and ρ_{GU}^2. Under each of the eight settings, we simulate two samples with sizes $n=50$ and 100, respectively. We then use each sample to construct a 95% confidence interval (CI) of ρ_{GS}^2 and ρ_{GU}^2 as $R_{GS}^2\pm1.96\text{se}(R_{GS}^2)$

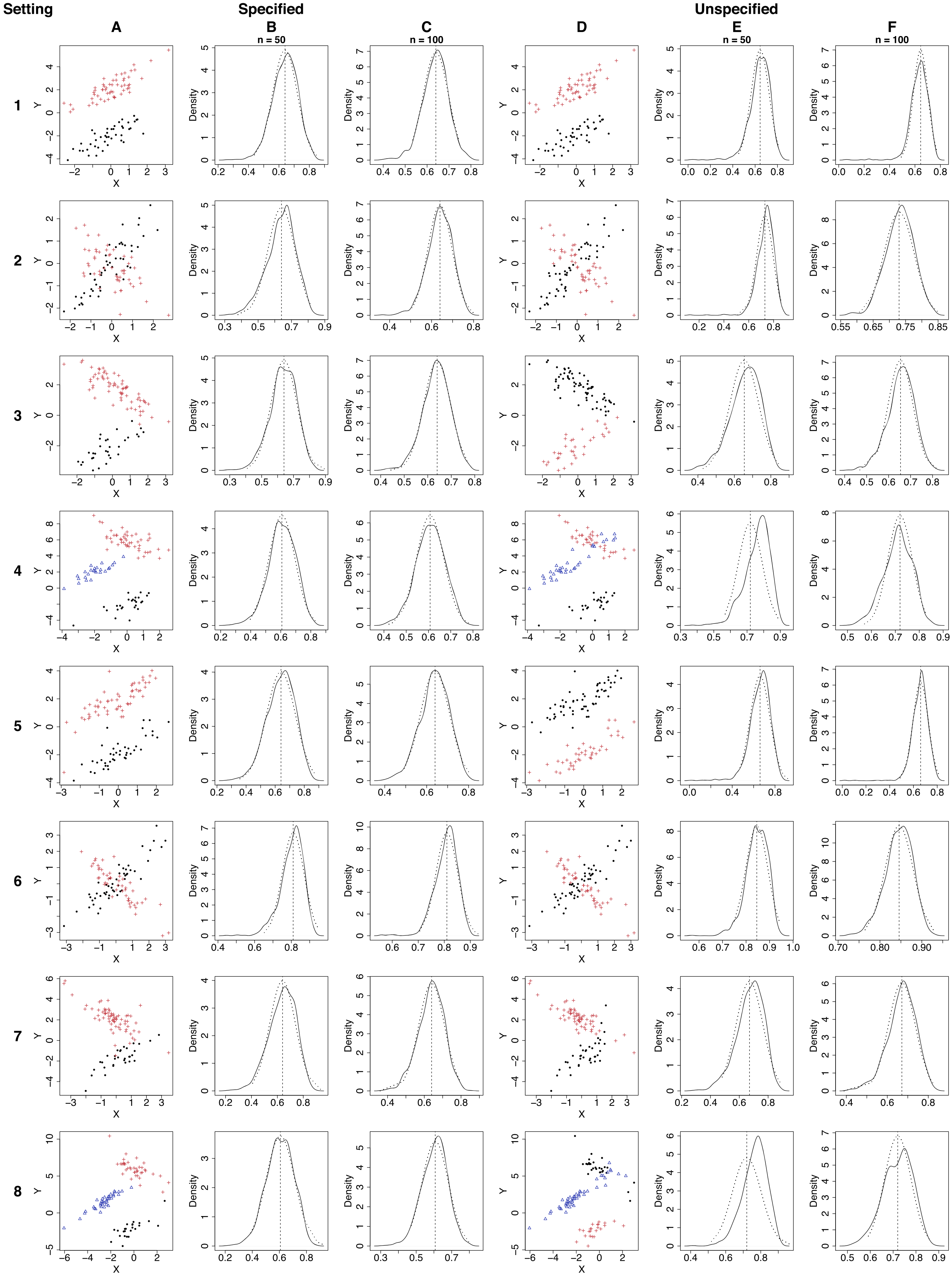

Figure 3. Comparison of the asymptotic distributions and the finite-sample distributions of R^2_{GS} and R^2_{GU}. **A**: Example samples with $n = 100$; colors and symbols represent values of Z. **B–C**: Finite-sample distributions $n = 50$ or 100 (black solid curves) versus the asymptotic distribution (black dotted curves) of R^2_{GS}; the vertical dashed lines mark the values of ρ^2_{GS}. **D**: Example samples with $n = 100$; colors and symbols represent values of $\widetilde{Z}$ inferred by the K-lines algorithm. **E–F**: Finite-sample distributions of $n = 50$ or 100 (black solid curves) versus the asymptotic distribution (black dotted curves) of R^2_{GU}; the vertical dashed lines mark the values of ρ^2_{GU}.

Table 1. Eight settings in simulation studies (Section 4), with each setting indicating a mixture of linear dependences.

Setting	K	Population	Parameters
1	$K=2$	Specified: $\mathbf{P}(Z=k)=p_k$ $(X,Y)\mid(Z=k)\sim N(\mu_k,\Sigma_k)$ $k=1,\dots,K$ Unspecified: $\sum_{k=1}^{K} p_k N(\mu_k,\Sigma_k)$	$p_1=p_2=0.5$ $\mu_1=(0,-2)^{\mathrm{T}},\mu_2=(0,2)^{\mathrm{T}}$ $\Sigma_1=\Sigma_2=\begin{bmatrix}1 & 0.8\\ 0.8 & 1\end{bmatrix}$
2	$K=2$		$p_1=p_2=0.5$ $\mu_1=\mu_2=(0,0)^{\mathrm{T}}$ $\Sigma_1=\begin{bmatrix}1 & 0.8\\ 0.8 & 1\end{bmatrix},\Sigma_2=\begin{bmatrix}1 & -0.8\\ -0.8 & 1\end{bmatrix}$
3	$K=2$		$p_1=0.3,p_2=0.7$ $\mu_1=(0,-2)^{\mathrm{T}},\mu_2=(0,2)^{\mathrm{T}}$ $\Sigma_1=\begin{bmatrix}1 & 0.8\\ 0.8 & 1\end{bmatrix},\Sigma_2=\begin{bmatrix}1 & -0.8\\ -0.8 & 1\end{bmatrix}$
4	$K=3$		$p_1=0.25,p_2=0.5,p_3=0.25$ $\mu_1=(0,-2)^{\mathrm{T}},\mu_2=(0,6)^{\mathrm{T}},\mu_3=(-2,2)^{\mathrm{T}}$ $\Sigma_1=\begin{bmatrix}1 & 0.8\\ 0.8 & 1\end{bmatrix},\Sigma_2=\begin{bmatrix}1 & -0.7\\ -0.7 & 1\end{bmatrix},\Sigma_3=\begin{bmatrix}1 & 0.9\\ 0.9 & 1\end{bmatrix}$
5	$K=2$	Specified: $\mathbf{P}(Z=k)=p_k$ $(X,Y)\mid(Z=k)\sim t_{\nu_k}(\mu_k,\Sigma_k)$ $k=1,\dots,K$ Unspecified: $\sum_{k=1}^{K} p_k t_{\nu_k}(\mu_k,\Sigma_k)$	$p_1=p_2=0.5,\nu_1=\nu_2=8$ $\mu_1=(0,-2)^{\mathrm{T}},\mu_2=(0,2)^{\mathrm{T}}$ $\Sigma_1=\Sigma_2=\begin{bmatrix}1 & 0.8\\ 0.8 & 1\end{bmatrix}$
6	$K=2$		$p_1=p_2=0.5,\nu_1=\nu_2=8$ $\mu_1=\mu_2=(0,0)^{\mathrm{T}}$ $\Sigma_1=\begin{bmatrix}1 & 0.8\\ 0.8 & 1\end{bmatrix},\Sigma_2=\begin{bmatrix}1 & -0.8\\ -0.8 & 1\end{bmatrix}$
7	$K=2$		$p_1=0.3,p_2=0.7,\nu_1=\nu_2=8$ $\mu_1=(0,-2)^{\mathrm{T}},\mu_2=(0,2)^{\mathrm{T}}$ $\Sigma_1=\begin{bmatrix}1 & 0.8\\ 0.8 & 1\end{bmatrix},\Sigma_2=\begin{bmatrix}1 & -0.8\\ -0.8 & 1\end{bmatrix}$
8	$K=3$		$p_1=0.25,p_2=0.5,p_3=0.25$ $\nu_1=\nu_2=\nu_3=8$ $\mu_1=(0,-2)^{\mathrm{T}},\mu_2=(0,6)^{\mathrm{T}},\mu_3=(-2,2)^{\mathrm{T}}$ $\Sigma_1=\begin{bmatrix}1 & 0.8\\ 0.8 & 1\end{bmatrix},\Sigma_2=\begin{bmatrix}1 & -0.7\\ -0.7 & 1\end{bmatrix},\Sigma_3=\begin{bmatrix}1 & 0.9\\ 0.9 & 1\end{bmatrix}$

NOTE: In the settings 1–4, $N(\mu_k,\Sigma_k)$ represents a bivariate Gaussian distribution with the mean vector μ_k and the covariance matrix Σ_k. In the settings 5–8, $t_{\nu_k}(\mu_k,\Sigma_k)$ represents a bivariate t distribution with the degrees of freedom ν_k, the location vector μ_k and the shape matrix Σ_k.

and $R^2_{GU}\pm 1.96\mathrm{se}(R^2_{GU})$, respectively. We construct the standard errors $\mathrm{se}(R^2_{GS})$ and $\mathrm{se}(R^2_{GU})$ in two ways: square roots of (a) the plug-in estimates of the asymptotic variances of R^2_{GS} and R^2_{GU}, or (b) the bootstrap estimates of $\mathrm{var}(R^2_{GS})$ and $\mathrm{var}(R^2_{GU})$. We also calculate the true asymptotic variances of R^2_{GS} and R^2_{GU} based on true parameter values and use them to construct the theoretical CIs. The results (Figure S2 in the supplementary material) show that the plug-in and bootstrap approaches construct similar CIs on the same sample. When n increases from 50 to 100, the CIs constructed by both approaches agree better with the theoretical CIs.

We also evaluate the coverage probabilities of the 95% CIs constructed by the plug-in approach and compare them with those of the theoretical CIs. Table S3 in the supplementary material summarizes the results. The theoretical CIs have coverage probabilities close to 95% under all the eight settings, providing additional verification of the asymptotic distributions. Overall, the plug-in confidence intervals have good coverage probabilities, which are increasingly closer to 95% as n increases; their coverage probabilities are in general closer to 95% under the first four bivariate Gaussians settings than under the last four bivariate t settings. The reason is that mixtures of bivariate Gaussians are more concentrated on K lines and better allow the K-lines algorithm to find the sample-level unspecified line centers, thus, reducing the unwanted variance due to failed algorithm convergence and making the plug-in variance estimate of R^2_{GU} more accurate. Comparing the line-membership specified and unspecified scenarios, the plug-in confidence intervals, as expected, have better coverage probabilities under the specified scenario that has less uncertainty. Table S3 also shows that the two plug-in options do not have obvious differences, suggesting that the first plug-in option ("P1"), which uses the asymptotic variances in the special bivariate Gaussian forms (Corollaries 3.1 and A.1), is robust and can be used in practice for its simplicity.

4.2. *Use of Scree Plot and AIC to Choose K*

Following Section 2.2, here we demonstrate the performance of the scree plot and the AIC in choosing K under the eight simulation settings. For each setting, we simulate a sample of size $n = 100$ and evaluate $W(B_K, P_n)$ in (2.4) and AIC(K) in (2.7) on this sample for K ranging from 1 to 10 (Figure S3 in the supplementary material). For all the eight settings, the scree plots and the AIC both suggest the correct K values. Even though Settings 5–8 violate the bivariate Gaussian assumption required by the AIC, the AIC results are still reasonable. In practice, users may use the scree plot together with the AIC to choose K.

4.3. Power Analysis

To confirm that R^2_{GU} is a powerful measure for capturing mixed linear dependences, we conduct a simulation study to compare $R^2_{GU}(K = 2)$ with six popular association measures or tests: the squared Pearson correlation (R^2), the maximal correlation (maxCor) estimated by the alternating conditional expectation algorithm (Breiman and Friedman 1985), the distance correlation (dCor) (Székely, Rizzo, and Bakirov 2007; Székely and Rizzo 2009), the maximal information coefficient (MIC) (Reshef et al. 2011), Chatterjee's rank correlation ξ (xiCor), and the Heller-Heller-Gorfine (HHG) test for independence.[3] All these measures[4] have values in $[0, 1]$.[5] Our simulation procedure follows Simon and Tibshirani (2014), where each relationship between two real-valued random variables X and Y is composed of a marginal distribution of $X \sim N(0, 5^2)$, a noiseless pattern (i.e., relationship) between X and Y, and a random error from $N(0, \sigma^2)$ added to Y. The null hypothesis is that X and Y are independent, while the alternative hypothesis is specified by the noiseless pattern and σ. Given a sample size $n = 30$, 50 or 200, we simulate $B = 1000$ samples from the alternative hypothesis. On each of these alternative samples, we randomly permute the Y observations to create a null sample. Then for each n we calculate the association measures on the B null samples and decide a rejection threshold for each measure as the $(1 - \alpha)$ quantile of its B null values, where $\alpha = 0.05$ is the significance level. Next, we calculate the association measures on the B alternative samples, compare each measure's B alternative values to its rejection threshold, and estimate the measure's power as the proportion of alternative values above the rejection threshold. Figure S4 in the supplementary material illustrates each measure's empirical distribution across alternative samples at each n and σ; all measures' variances decrease as n increases.[6]

Figure 4 shows that R^2_{GU} is the most powerful measure when the pattern is a mixture of positive and negative linear dependences. When the pattern is a mixture of nonlinear relationships that can be approximated by a mixture of linear dependences, R^2_{GU} is still the most powerful. When the pattern is linear, R^2 is expectedly the most powerful, and the other measures including R^2_{GU} also have perfect power up to $\sigma = 3$ at $n = 30$. Under a parabola pattern, which can be approximated by two intersecting lines (i.e., the "V" pattern), R^2_{GU} still has good power and is comparable to xiCor, maxCor, dCor, HHG, and MIC. These results confirm the application potential of R^2_{GU} in capturing complex relationships that can be approximated by mixtures of linear dependences.

5. Real Data Applications

5.1. Analysis of the *A. thaliana* Gene Expression Dataset

Back to our motivating example in *A. thaliana*, here we use this gene expression dataset (Li et al. 2008) to demonstrate the use of our generalized Pearson correlation squares to capture biologically meaningful gene–gene relationships. The glucosinolate (GSL) biosynthesis pathway has been well studied in *A. thaliana*, and 31 genes in this pathway have been experimentally identified (Kim et al. 2012). Since genes in the same pathway are functionally related, their relationships should be distinct from their relationships with the other genes outside of the pathway. Hence, a powerful association measure should distinguish the pairwise gene–gene relationships within the GSL pathway from the relationships of randomly paired GSL and non-GSL genes.

The dataset (Table S4 in the supplementary material) contains $n = 232$ samples, 26 GSL genes, and four index variables: `condition` (oxidation, wounding, UV-B light, and drought), `treatment` (yes and no), `replicate` (1 and 2), and `tissue` (root and shoot). We observe that only the `tissue` variable is a good indicator of linear dependences, as illustrated in Figures 1 and S5–S8 in the supplementary material.

Figure 5(A) shows the values of R^2, maxCor, dCor, MIC, xiCor, HHG,[7] and $R^2_{GU}(K = 2)$, all of which do not use index variables, as well as R^2_{GS}, which uses the index variable as `condition`, `treatment`, `replicate`, or `tissue`. All these measures are computed for the $\binom{26}{2} = 325$ GSL gene pairs and 2600 random gene pairs, each of which consists of a GSL gene (out of 26 GSL genes) and a randomly selected non-GSL gene (out of 100 randomly selected non-GSL genes). Among these measures, only R^2_{GS}(tissue) and $R^2_{GU}(K = 2)$ show significantly stronger relationships (at the significance level of 0.01) within the GSL pathway than in random gene pairs (with respective p-values 1.29×10^{-28} and 7.65×10^{-20} from one-sided Wilcoxon rank-sum test). Hence, R^2_{GS} is a useful and powerful measure when a good index variable is available; otherwise, R^2_{GU} is advantageous in capturing complex but interpretable gene–gene relationships without knowledge of index variables[8].

To verify the agreement between the K-lines clusters and the `tissue` index variable, for every GSL gene pair, we compare the $K = 2$ sample clusters identified by the K-lines algorithm with the sample groups defined by each of the four index variables (`condition`, `treatment`, `replicate`, and `tissue`) using Fisher's exact test[9] and the adjusted mutual

[3] For the implementation of these measures, see Table S2 in the supplementary material.

[4] The HHG test statistic is not an association measure, so (1− HHG test p-value) is used as a measure.

[5] Note that xiCor may take negative values at the sample level.

[6] The HHG test is not included because (1− HHG test p-value) is too close to 1 most of the time.

[7] The HHG test statistic is not an association measure, so (1− HHG test p-value) is used as a measure.

[8] In this application, the two variables X and Y refer to two genes with comparable expression levels, so no normalization is performed before K-lines clustering.

[9] A smaller p-value more strongly rejects the null hypothesis that the K-lines clusters are independent of an index variable.

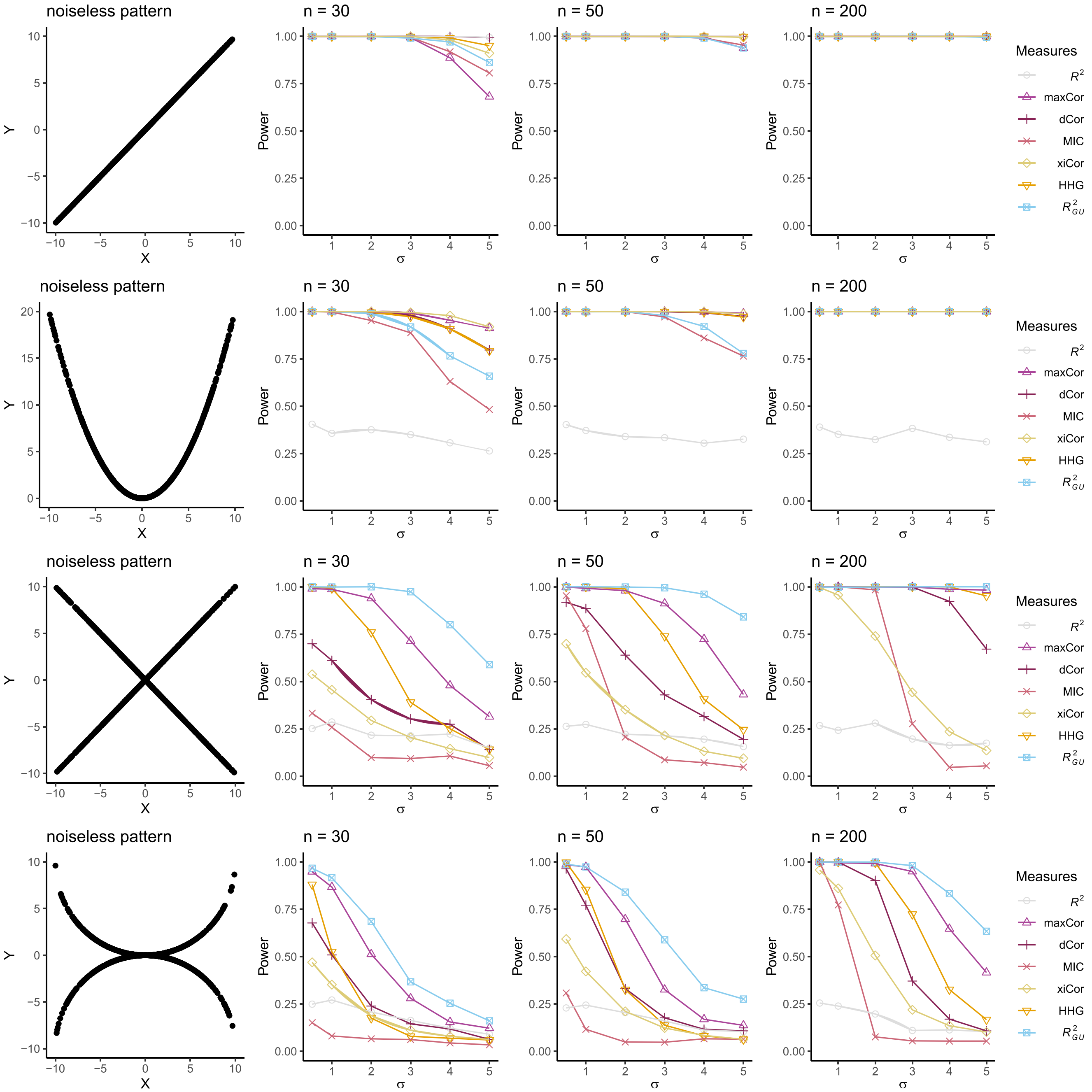

Figure 4. Power analysis. Simulation studies that compare the statistical power of seven measures/tests: the squared Pearson correlation (R^2), the maximal correlation (maxCor), the distance correlation (dCor), the maximal information coefficient (MIC), Chatterjee's rank correlation (xiCor), the Heller-Heller-Gorfine (HHG) test, and R^2_{GU} with $K = 2$. In each row, the noiseless pattern illustrates a relationship between two real-valued random variables X and Y when no noise is added ($\sigma = 0$). Under the null hypothesis, X and Y are independent. Varying alternative hypotheses are formed by the noiseless pattern with noise $\sim N(0, \sigma^2)$ at varying σ added to Y. Under each alternative hypothesis corresponding to one σ, we estimate the power of the seven measures/tests given each sample size n (columns 2–4).

information[10]. The results in Table 2 confirm that the K-lines clusters exhibit higher consistency with the `tissue` variable compared to the other three index variables across the 325 GSL gene pairs, suggesting that the K-lines algorithm separates the samples in good accordance with their tissue types.

5.2. Identification of Beta Cell Subtypes by K-Lines Clustering

The second example is from a single-cell gene expression dataset of mouse pancreas (Baron et al. 2016). A previous study found that using projective nonnegative matrix factorization (PNMF) to project cells from a high-dimensional (7838) gene expression space to a low-dimensional space of PNMF factors, some PNMF factors exhibited heterogeneous linear relationships with the cell library size (i.e., the total number of sequencing reads mapped to each cell), and each relationship corresponded to a known

[10] A larger adjusted mutual information indicates a better agreement between the K-lines clusters and an index variable.

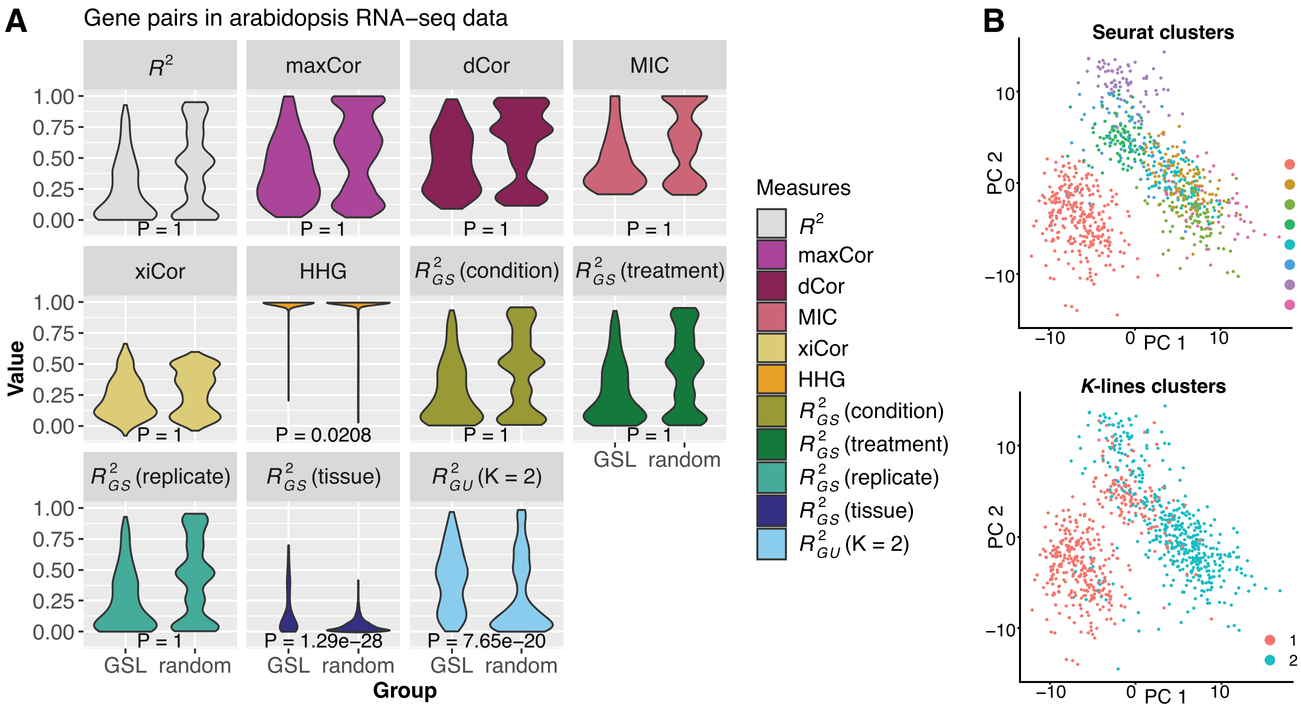

Figure 5. Real data applications of R^2_{GS}, R^2_{GU}, and the K-lines clustering algorithm. **A**: Analysis of the *Arabidopsis* gene expression dataset. We compare 11 measures, including seven unspecified measures (R^2, maxCor, dCor, MIC, xiCor, HHG, and R^2_{GU} with $K = 2$) and four R^2_{GS} measures with different index variables, in terms of measuring pairwise gene relationships within the GSL pathway ("GSL") versus relationships between a GSL gene and a randomly paired non-GSL gene ("random"). For each measure, the one-sided Wilcoxon rank-sum test is used to compare the measure's values in the two groups ("GSL" vs. "random"), and the resulting p-value is marked in each panel. At the significance level of 0.01, R^2_{GS}(tissue) and $R^2_{GU}(K = 2)$ are the only two measures indicating that the gene pairs within the GSL pathway have significantly stronger relationships than the random GSL-nonGSL gene pairs do. **B**: Beta cell clusters found by Seurat (left) or the K-lines clustering algorithm (right).

cell type (Song et al. 2021). Motivated by this finding, here we apply the K-lines clustering algorithm to $n = 894$ beta cells' PNMF factors and library sizes. For each of 10 PNMF factors, we pair it up with the cell library size and apply the K-lines clustering algorithm,[11] which finds a notable $K = 2$ pattern (by AIC) for PNMF factor 4 versus cell library size (Figure S9A). Comparing the resulting two beta cell clusters to the default clustering results by the most popular pipeline Seurat using PCA and UMAP visualization, we find that the default Seurat leads to eight clusters including many hardly separable ones, while the two K-lines clusters are distinct in both PCA and UMAP visualization (Figures 5(B) and S9B).

We verify the two K-lines clusters by conducting literature review and performing KEGG pathway enrichment analysis. First, based on the review by Miranda, Macias-Velasco, and Lawson (2021), we map the clusters 1 and 2 to the immature and mature beta subtypes respectively by verifying that mature cell marker genes *Gck*, *Ins1*, and *Iapp* are significantly more highly expressed in cluster 2 than cluster 1 (with one-sided Wilcoxon rank-sum test p-values 0.001001, $< 2.2 \times 10^{-16}$, and $< 2.2 \times 10^{-16}$, respectively). Second, we confirm this mapping result by using (a) the `FindMarkers()` function in Seurat to find clusters 1 and 2's respective marker genes and (b) the R package clusterProfiler 4.0 to find the enriched KEGG pathways within each cluster's marker genes (Tables S5–S6 in the supplementary material). The results suggest that beta cell cluster 1 is related to the insulin resistance, while cluster 2 is related to the insulin signaling pathway, consistent with previous findings that immature beta cells are over-represented in insulin resistant patients and mature beta cells are responsible for regular insulin secretion, respectively (Miranda, Macias-Velasco, and Lawson 2021).

Table 2. Comparison of the K-lines clusters and the four index variables (`condition`, `treatment`, `replicate`, and `tissue`) in the *A. thaliana* gene expression dataset.

	Fisher's exact test p-value			Adjusted mutual information		
	<0.05	<0.005	<0.0005	Mean	3rd quartile	Max
`condition`	29.54%	19.69%	11.08%	0.007	0.010	0.090
`treatment`	34.15%	22.46%	14.77%	0.013	0.019	0.152
`replicate`	1.85%	0.00%	0.00%	−0.001	0.000	0.019
`tissue`	72.00%	63.38%	58.77%	0.293	0.578	1.000

NOTE: In columns 1–3, each percentage represents the percentage of the 325 GSL gene pairs whose K-lines clusters have a significant p-value (given a threshold) when compared with an index variable by Fisher's exact test. In columns 4–6, the mean, 3rd quartile, and maximum of the adjusted mutual information values are reported for the 325 GSL gene pairs whose K-lines clusters are compared with each index variable.

6. Discussion

The generalized Pearson correlation squares extend the classic and popular Pearson correlation to capturing heterogeneous linear relationships. This new suite of measures has broad potential use in scientific applications. In addition to gene expression analysis, statistical genetics is a potential application domain because genetic variants could exhibit heterogeneous effects on

[11] In this application, the two variables X and Y refer to the cell library size and a PNMF factor, whose values are on the same scale, so no normalization is performed before K-lines clustering.

a phenotype. When known subpopulations, for example, race, gender, and geography, cannot explain heterogenous associations between a genetic variant and a phenotype, R^2_{GU} and the K-lines algorithm could be useful.

A future direction is to extend the generalized Pearson correlation squares to be rank-based. This extension will make the measures robust to outliers and capable of capturing a mixture of monotone relationships.

Another future direction is to generalize the K-lines clustering algorithm to the K-hyperplanes clustering algorithm, following the subspace clustering literature (Vidal 2010). Specifically, for $p \geq 2$ variables, the kth cluster center becomes the $(p-1)$-dimensional hyperplane defined by the top $(p-1)$ principal components of the data points (p-dimensional vectors) assigned to the kth cluster, $k = 1, \ldots, K$. In this generalization, we only need to generalize the recentering step in the K-lines clustering algorithm (Algorithm 1) while keeping the assignment step as assigning every data point to the closest hyperplane based on the perpendicular distance. Then, we can generalize a multivariate dependence measure by calculating the weighted sum of the measure's values across the K clusters.

Supplementary Materials

The online supplementary materials (Supplements.zip) contain 1. the Author Contributions Checklist form (acc_form.pdf); 2. the Supplementary Material file (supplementary_material.pdf) containing proofs of theorems, convergence properties of the K-lines algorithm, more simulation results, another real data application about the dependence of glioma patient survival on CD44 expression, real data description, more tables and figures, and additional references; 3. the reproducibility_materials.zip file including (1) a README.md file, (2) an Rmd file to reproduce the results and the html file knitted from the Rmd file, (3) a data folder containing the datasets and data_description.txt, (4) a results folder containing the intermediate results, (5) a code folder containing R code used in the Rmd file, and (6) the pdf files of Figure S9A and Figure S9B.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

The authors gratefully acknowledge NSF DBI-1846216 and DMS-2113754, NIH/NIGMS R01GM120507 and R35GM140888, Johnson and Johnson WiSTEM2D Award, Sloan Research Fellowship, and UCLA David Geffen School of Medicine W.M. Keck Foundation Junior Faculty Award.

References

Baron, M., Veres, A., Wolock, S. L., Faust, A. L., Gaujoux, R., Vetere, A., Ryu, J. H., Wagner, B. K., Shen-Orr, S. S., Klein, A. M. et al. (2016), "A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter-and Intra-cell Population Structure," *Cell Systems*, 3, 346–360. [2460]

Benaglia, T., Chauveau, D., Hunter, D., and Young, D. (2009), "mixtools: An r Package for Analyzing Finite Mixture Models," *Journal of Statistical Software*, 32, 1–29. [2451]

Bjerve, S., and Doksum, K. (1993), "Correlation Curves: Measures of Association as Functions of Covariate Values," *The Annals of Statistics*, 21, 890–902. [2450]

Breiman, L., and Friedman, J. H. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation," *Journal of the American Statistical Association*, 80, 580–598. [2459]

Chatterjee, S. (2021), "A New Coefficient of Correlation," *Journal of the American Statistical Association*, 116, 2009–2022. [2451]

De Veaux, R. D. (1989), "Mixtures of Linear Regressions," *Computational Statistics & Data Analysis*, 8, 227–245. [2451]

Delicado, P., and Smrekar, M. (2009), "Measuring Non-linear Dependence for Two Random Variables Distributed Along a Curve," *Statistics and Computing*, 19, 255–269. [2450]

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005), "Measuring Statistical Dependence with Hilbert-Schmidt Norms," in *International Conference on Algorithmic Learning Theory*, pp. 63–77, Springer. [2451]

Hawkins, D. S., Allen, D. M., and Stromberg, A. J. (2001), "Determining the Number of Components in Mixtures of Linear Models," *Computational Statistics & Data Analysis*, 38, 15–48. [2451]

Heller, R., Heller, Y., and Gorfine, M. (2012), "A Consistent Multivariate Test of Association based on Ranks of Distances," *Biometrika*, 100, 503–510. [2451]

Hurn, M., Justel, A., and Robert, C. P. (2003), "Estimating Mixtures of Regressions," *Journal of Computational and Graphical Statistics*, 12, 55–79. [2451]

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991), "Adaptive Mixtures of Local Experts," *Neural Computation*, 3, 79–87. [2451]

Jolliffe, I. T. (1982), "A Note on the Use of Principal Components in Regression," *Applied Statistics*, 31, 300–303. [2453]

Jones, P., and McLachlan, G. (1992), "Fitting Finite Mixture Models in a Regression Context," *Australian & New Zealand Journal of Statistics*, 34, 233–240. [2451]

Kim, K., Jiang, K., Teng, S. L., Feldman, L. J., and Huang, H. (2012), "Using Biologically Interrelated Experiments to Identify Pathway Genes in Arabidopsis," *Bioinformatics*, 28, 815–822. [2450,2459]

Lee, D., and Zhu, B. (2021), "A Semiparametric Kernel Independence Test with Application to Mutational Signatures," *Journal of the American Statistical Association*, 116, 1648–1661. [2451]

Leisch, F. (2008), "Modelling Background Noise in Finite Mixtures of Generalized Linear Regression Models," in *COMPSTAT 2008*, pp. 385–396, Springer. [2451]

Li, J., Hansen, B. G., Ober, J. A., Kliebenstein, D. J., and Halkier, B. A. (2008), "Subclade of Flavin-Monooxygenases Involved in Aliphatic Glucosinolate Biosynthesis," *Plant Physiology*, 148, 1721–1733. [2450,2459]

Li, K.-C. (2002), "Genome-Wide Coexpression Dynamics: Theory and Application," *Proceedings of the National Academy of Sciences*, 99, 16875–16880. [2450]

Lloyd, S. (1982), "Least Squares Quantization in PCM," *IEEE Transactions on Information Theory*, 28, 129–137. [2453]

Miranda, M. A., Macias-Velasco, J. F., and Lawson, H. A. (2021), "Pancreatic β-cell Heterogeneity in Health and Diabetes: Classes, Sources, and Subtypes," *American Journal of Physiology-Endocrinology and Metabolism*, 320, E716–E731. [2461]

Murtaph, F., and Raftery, A. (1984), "Fitting Straight Lines to Point Patterns," *Pattern Recognition*, 17, 479–483. [2451]

Quandt, R. E., and Ramsey, J. B. (1978), "Estimating Mixtures of Normal Distributions and Switching Regressions," *Journal of the American Statistical Association*, 73, 730–738. [2451]

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., and Sabeti, P. C. (2011), "Detecting Novel Associations in Large Data Sets," *Science*, 334, 1518–1524. [2451,2459]

Scharl, T., Grün, B., and Leisch, F. (2009), "Mixtures of Regression Models for Time Course Gene Expression Data: Evaluation of Initialization and Random Effects," *Bioinformatics*, 26, 370–377. [2451]

Shen, C., Priebe, C. E., and Vogelstein, J. T. (2019), "From Distance Correlation to Multiscale Graph Correlation," *Journal of the American Statistical Association*, 115, 280–291. [2451]

Simon, N., and Tibshirani, R. (2014), Comment on "Detecting Novel Associations in Large Data Sets" by Reshef et al, Science Dec 16, 2011," arXiv preprint arXiv:1401.7645 . [2459]

Smith, R. J. (2009), "Use and Misuse of the Reduced Major Axis for Line-Fitting," *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, 140, 476–486. [2453]

Song, D., Li, K., Hemminger, Z., Wollman, R., and Li, J. J. (2021), "scPNMF: Sparse Gene Encoding of Single Cells to Facilitate Gene Selection for Targeted Gene Profiling," *Bioinformatics*, 37, i358–i366. [2461]

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007), "Measuring and Testing Dependence by Correlation of Distances," *The Annals of Statistics*, 35, 2769–2794. [2451,2459]

Székely, G. J., and Rizzo, M. L. (2009), "Brownian Distance Covariance," *The Annals of Applied Statistics*, 3, 1236–1265. [2451,2459]

Turner, T. R. (2000), "Estimating the Propagation Rate of a Viral Infection of Potato Plants via Mixtures of Regressions," *Journal of the Royal Statistical Society*, Series C, 49, 371–384. [2451]

Vidal, R. (2010), "A Tutorial on Subspace Clustering," *https://api.semanticscholar.org/CorpusID:7099537* [2452,2453,2462]

Wang, X., Jiang, B., and Liu, J. S. (2017), "Generalized R-squared for Detecting Dependence," *Biometrika*, 104, 129–139. [2451]

Wang, Y. R., Waterman, M. S., and Huang, H. (2014), "Gene Coexpression Measures in Large Heterogeneous Samples Using Count Statistics," *Proceedings of the National Academy of Sciences*, 111, 16371–16376. [2451]

Wedel, M., and DeSarbo, W. S. (1994), "A Review of Recent Developments in Latent Class Regression Models," in *Advanced Methods of Marketing Research*, ed. R.P. Bagozzi, pp. 352–388, Cambridge: Blackwell Publishers. [2451]

Zheng, S., Shi, N.-Z., and Zhang, Z. (2012), "Generalized Measures of Correlation for Asymmetry, Nonlinearity, and Beyond," *Journal of the American Statistical Association*, 107, 1239–1252. [2451]