Ranking Distance Correlation

1. Set-up

Let $X \sim U(0,1)^d \in \mathbb{R}^d$, where U is the uniform distribution, d is the dimension size.

Let $Y = f(X) * A \in \mathbb{R}$, where $f : \mathbb{R}^d \to \mathbb{R}^d$ is certain function of X, and A is a $d \times 1$ vector. So Y is a one-dimensional variable that is related to X.

For the choice of $f()$, we use linear, quadratic, cubic, sine period 1/2, sine period 1/8, $X^{0.25}$, circle, step function, exponential, and log function. Please see the code for detail.

For the choice of A, we use $A(i) = 1/i, \forall i = 1, \ldots, d$, so that the entries of A decays as the dimension increases.

For another choice of A, we use $A(i) = 1/\lceil d * C(i) \rceil$, where $C_i \overset{i.i.d.}{\sim} U(0,1)$. So A randomly decays as the dimension increases, rather than fixed decay in the previous case. In experiment, A is pre-generated and never changed in each Monte-Carlo run.

2. Independence tests

Given two sample data $\mathcal{X}$ and $\mathcal{Y}$ of size $n \times d$ and size $n \times 1$ (n is the sample size), the null hypothesis is that they are not independent. The alternative hypothesis is they are independent.

Fix f and A. To test whether Ranking Distance Correlation improves over distance correlation, the experiment is done as follows: we first obtain $\mathcal{X}$ and $\mathcal{Y}$ by generate (X, Y) for $n = 100$ pairs, and calculate the distance correlation and Ranking Distance Correlation between them. Repeat it for 500 Monte-Carlo replicates, we obtain the empirical distribution of DC and RDC for that type of f and A, which are the test statistics under the null.

Now repeat the same procedure, but $\mathcal{X}$ and $\mathcal{Y}$ are generated from (Z, Y) for $n = 100$ pairs. $Z \sim U(0,1)^d \in \mathbb{R}^d$ but is independent from X, so Z is also independent from Y. After 500 Monte-Carlo replicates, we obtain the empirical distribution of DC and RDC under the alternative.

Given DC and RDC under the null and the alternative, we calculate the testing power at type 1 error level 0.95. To assess the effect for high-dimensional data, we do the experiment for $d = 1, \ldots, 1000$ at $n = 100$.

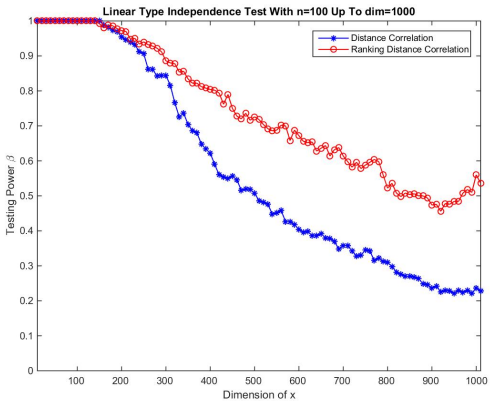
The next two sections show the results. It is clear that no matter in fixed decay or random decay, RDC performs at least similarly as DC; and for more than half the function types, RDC is significantly better than DC as the dimension increases. This shows that RDC has a very obvious advantage in identifying correlation among high-dimensional data.

The last section considers $Y = f(X * A)$ with fixed decay, which also shows the advantage of RDC.

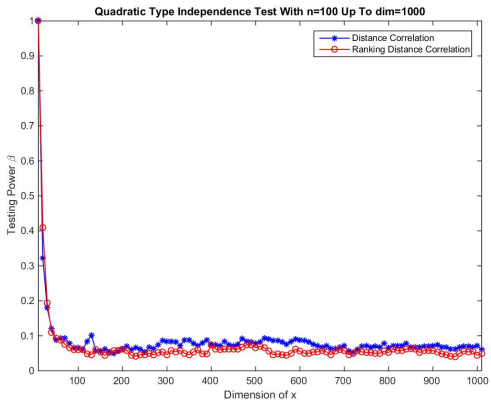
3. $y = f(x) * A$ with fixed decay

4. $y = f(x) * A$ with random decay

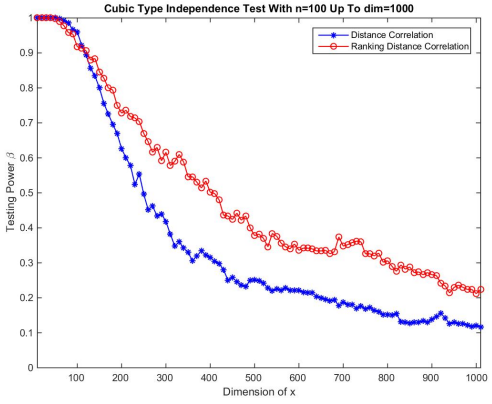
5. $Y = f(X * A)$ with fixed decay



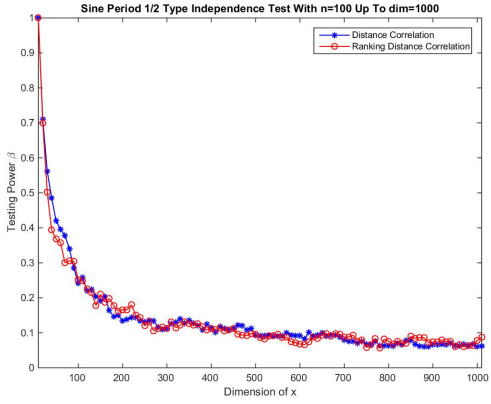
(a)



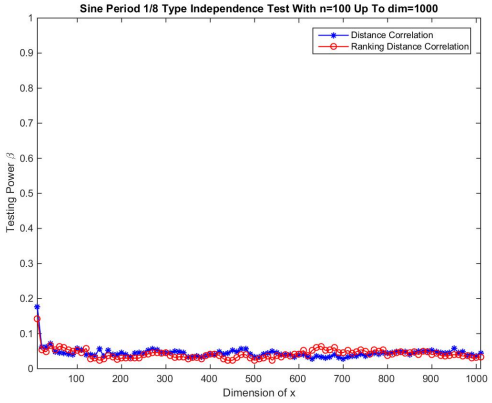
(b)



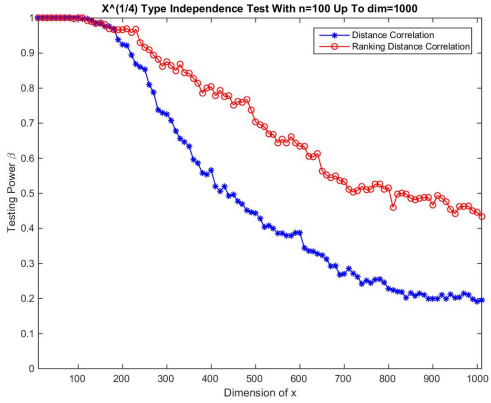
(c)



(d)

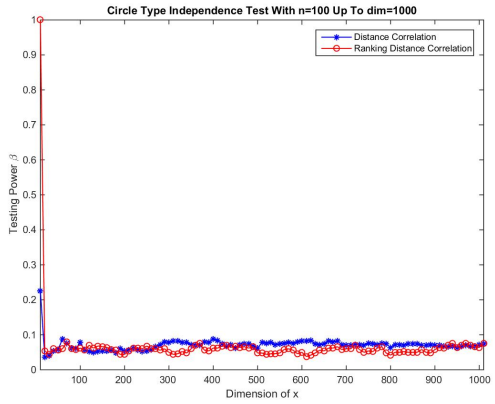


(e)

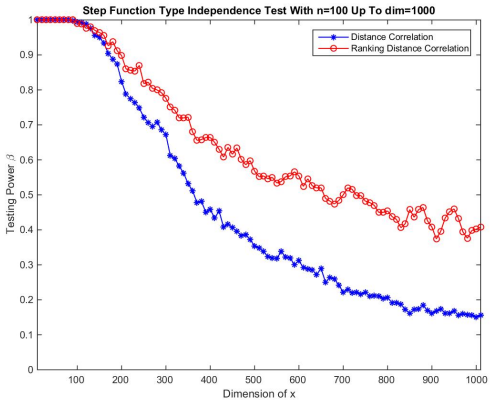


(f)

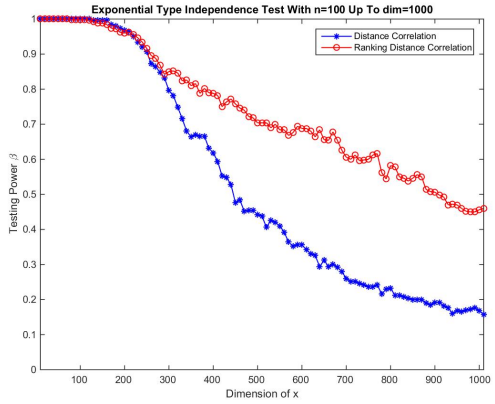
Figure 1: Fixed Decay 1



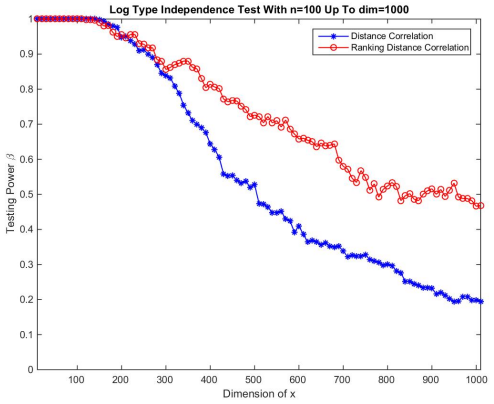
(a)



(b)

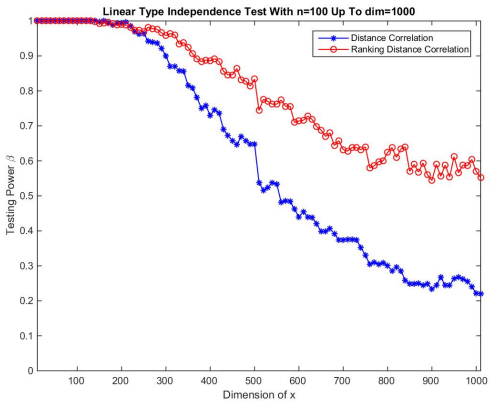


(c)

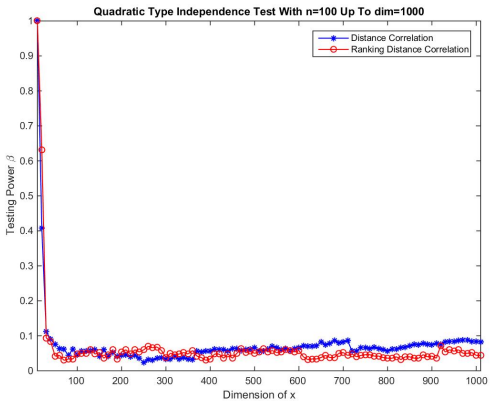


(d)

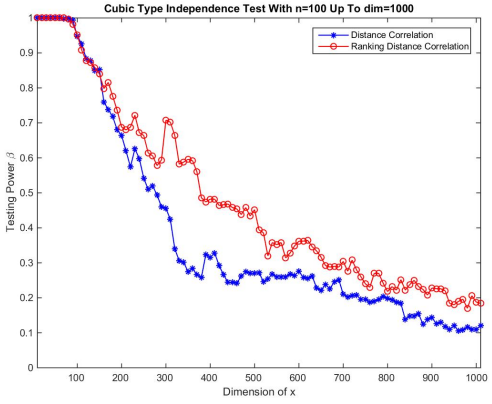
Figure 2: Fixed Decay 2



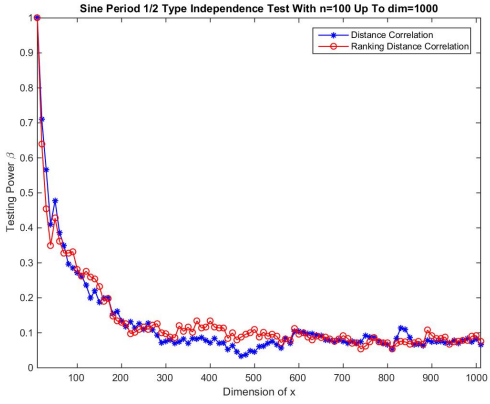
(a)



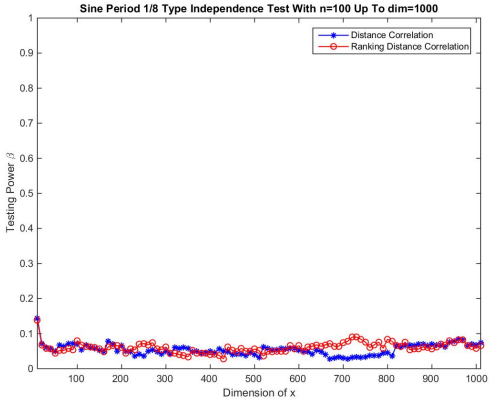
(b)



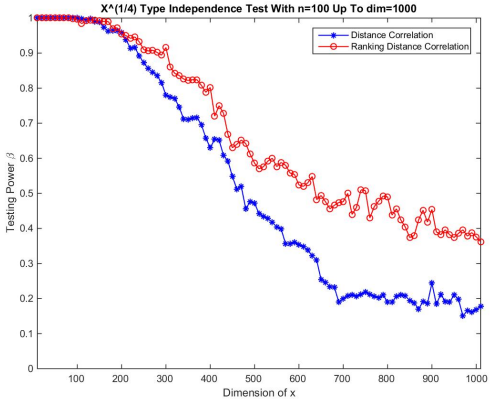
(c)



(d)

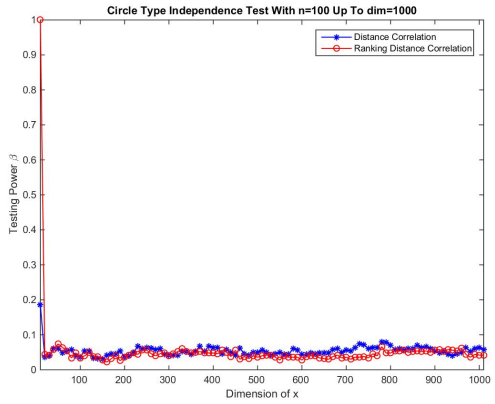


(e)

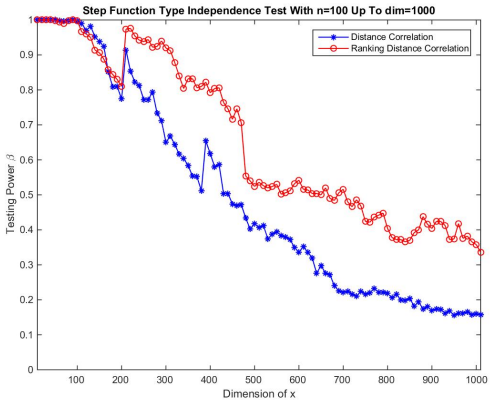


(f)

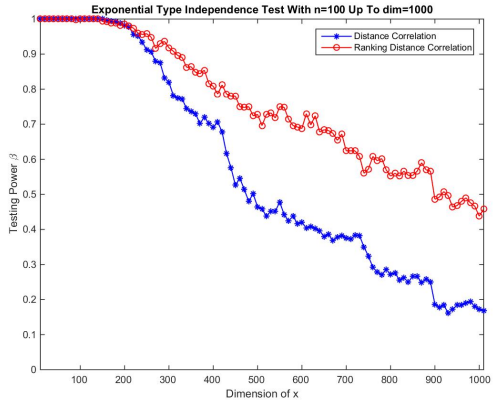
Figure 3: Random Decay 1



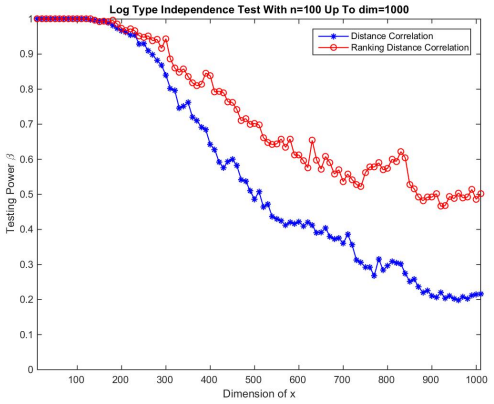
(a)



(b)

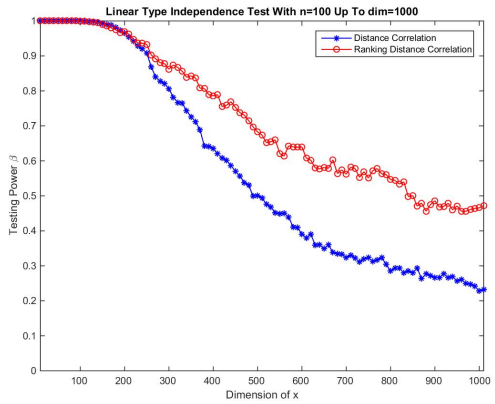


(c)

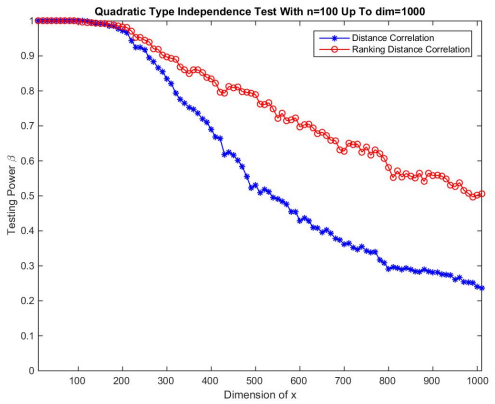


(d)

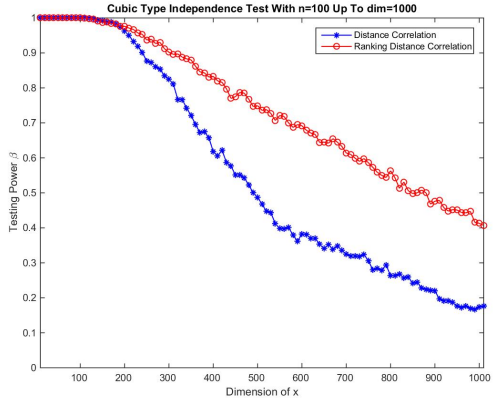
Figure 4: Random Decay 2



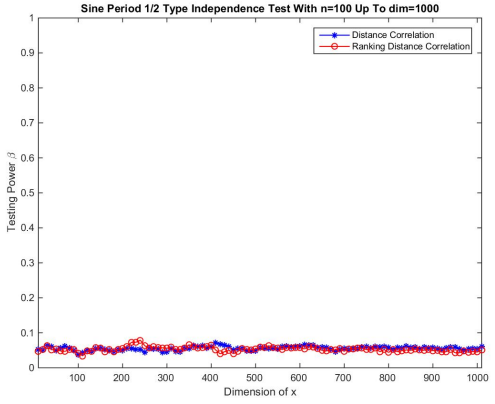
(a)



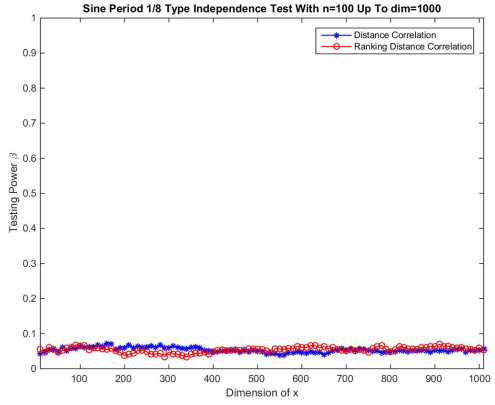
(b)



(c)



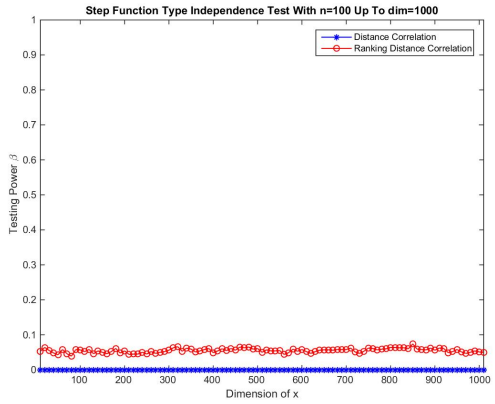
(d)



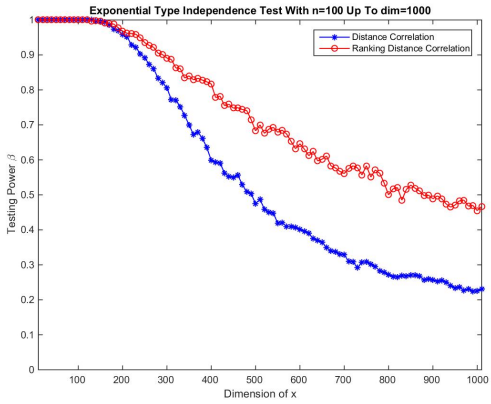
(e)

(f)

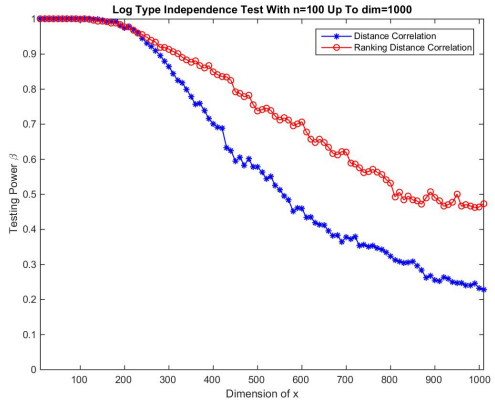
Figure 5: Fix Decay 3



(a)



(b)



(c)

Figure 6: Fix Decay 4