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 $Y = f(A \times X) + \epsilon \in \mathbb{R}$, where $f : \mathbb{R} \to \mathbb{R}$ is a function of X, A is a $1 \times d$ transformation, and ϵ is random noise. 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. Note that we may also use a random decay rather than the fixed decay, and the numerical phenomenon is similar.

2. Independence tests

Given two sample data $\mathcal{X}$ and $\mathcal{Y}$ of size $d \times n$ and size $1 \times n$ (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 our Ranking Distance Correlation improves over distance correlation, the experiment is done as follows: we first obtain $\mathcal{X}$ and $\mathcal{Y}$ by generating (X, Y) for $n = 100$ pairs, and calculate the distance correlation and Ranking Distance Correlation between them. Repeat it for 1000 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 1000 Monte-Carlo replicates, we obtain the empirical distribution of DC and RDC under the alternative.

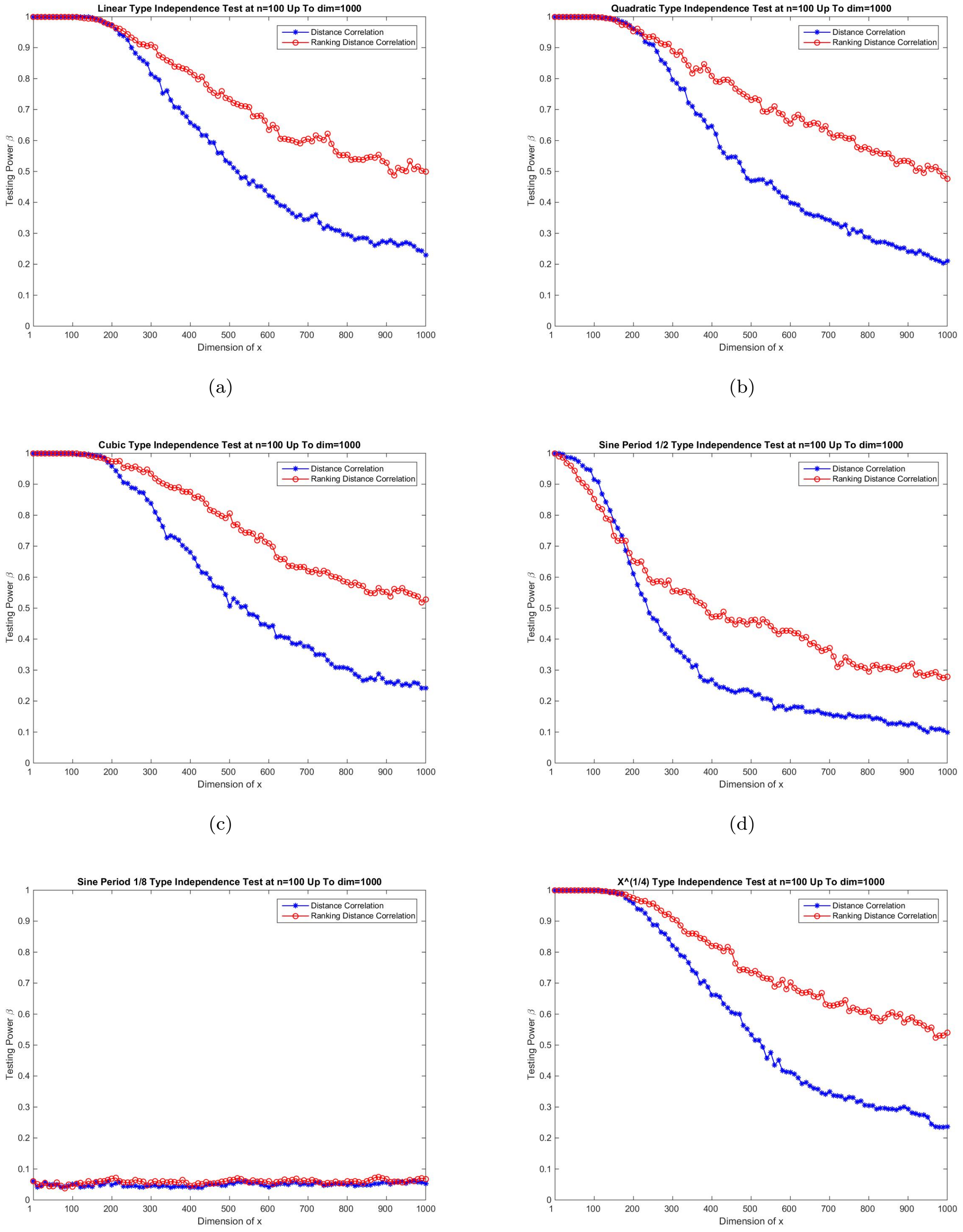
Given DC and RDC under the null and the alternative, we calculate the testing power at type 1 error level 0.95. To assess the effect for high-dimensional data, we do the experiment for $d = 1, 10, 20, \ldots, 1000$ at $n = 100$.

For the first subsection, we present the testing powers with $\epsilon = 0$; but the performance is similar for small error as well. For the second subsection, we present the testing powers with increasing noise level (which is generated by Gaussian) at fixed $d = 500$; the interpretation is similar for other dimensions as well.

It is clear that RDC performs very similarly as DC for low-dimensional data; but for high-dimensional data we observe a clear advantage of RDC in testing independence. Such advantage is also robust against noise.

3. Powers with respect to Dimension

4. Powers with respect to Noise



(a)

(b)

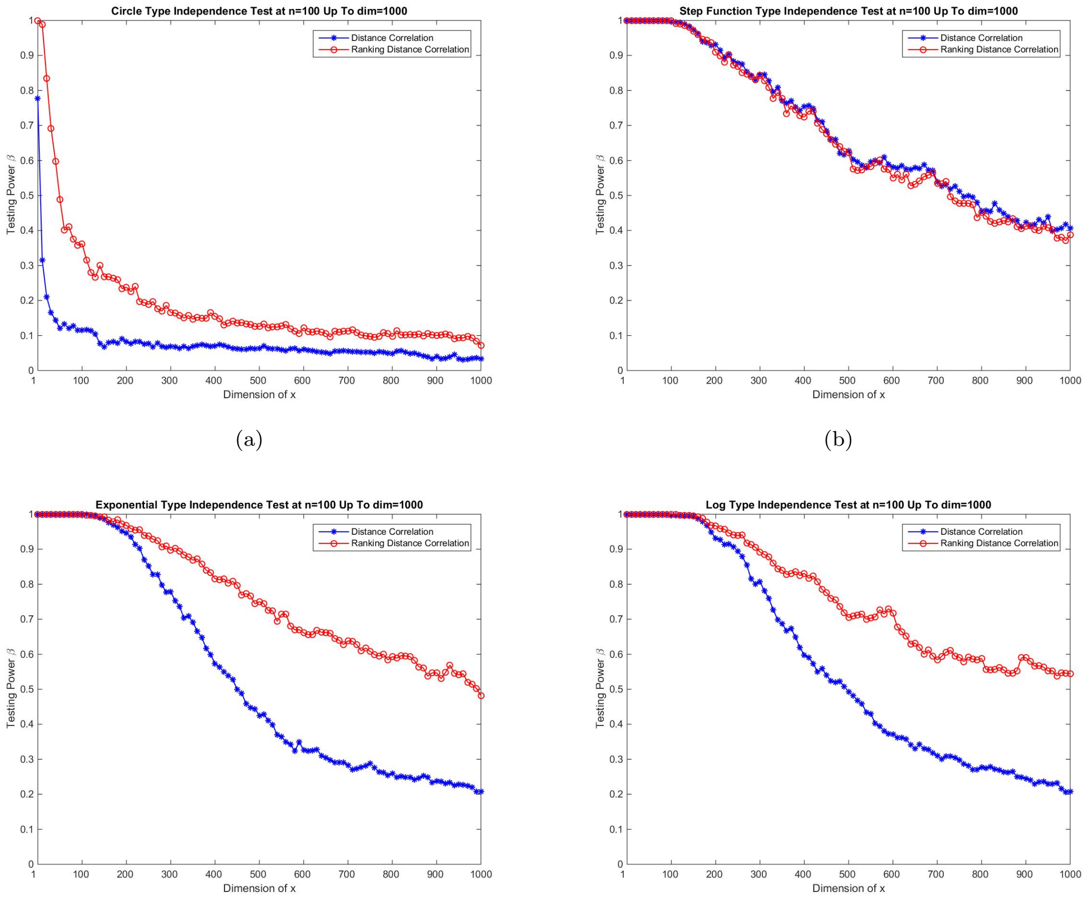
(c)

(d)

(e)

(f)

Figure 1: Testing Powers w.r.t. Increasing Dimension



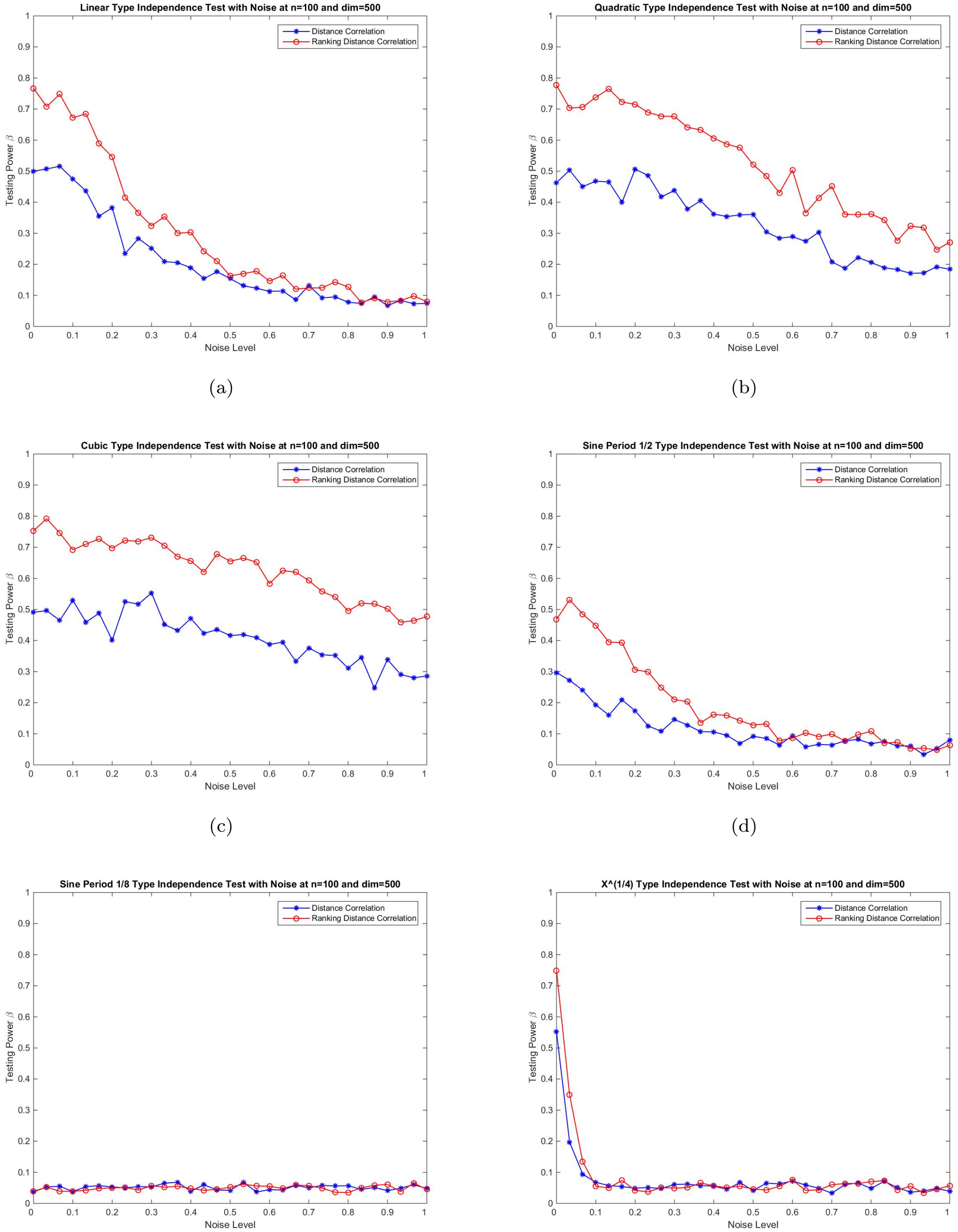
(a)

(b)

(c)

(d)

Figure 2: Testing Powers w.r.t. Increasing Dimension



(a)

(b)

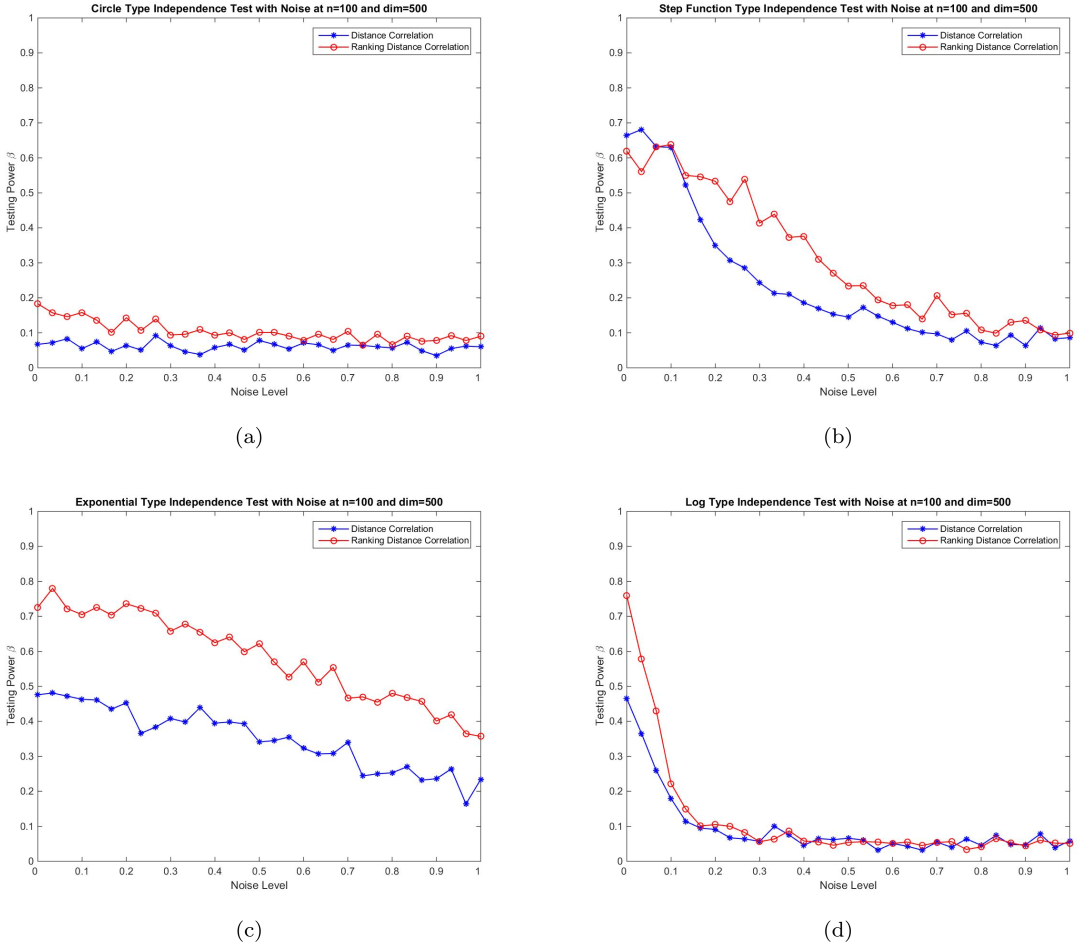
(c)

(d)

(e)

(f)

Figure 3: Testing Powers w.r.t. Increasing Noise



(a)

(b)

(c)

(d)

Figure 4: Testing Powers w.r.t. Increasing Noise