Evaluating The Repeatability of Two Studies with a Large Number of Objects: Modified Kendall Rank-Order Association Test

Tian Zheng

Department of Statistics
Columbia University

May 6th, 2010

Acknowledgement

- Collaborator: Shaw-Hwa Lo, Statistics, Columbia.
- The paper can be downloaded at http://www.stat.columbia.edu/∼tzheng

Motivational example

Data from van 't Veer et al. (2002):

- 78 breast cancer patients;
- 44 remained disease-free for more than 5 years;
- 34 developed metastases within 5 years;
- Gene expression levels of 24,479 oligonucleotides were measured on each individual;
- Goal: identify important genes that are associated with the metastases risk.
- Our concern: different genes are identified using different data sets due to sampling variation.

Motivational example

Data from van 't Veer et al. (2002):

- 78 breast cancer patients;
- 44 remained disease-free for more than 5 years;
- 34 developed metastases within 5 years;
- Gene expression levels of 24,479 oligonucleotides were measured on each individual;
- Goal: identify important genes that are associated with the metastases risk.
- Our concern: different genes are identified using different data sets due to sampling variation.

Motivational example

Data from van 't Veer et al. (2002):

- 78 breast cancer patients;
- 44 remained disease-free for more than 5 years;
- 34 developed metastases within 5 years;
- Gene expression levels of 24,479 oligonucleotides were measured on each individual;
- Goal: identify important genes that are associated with the metastases risk.
- Our concern: different genes are identified using different data sets due to sampling variation.

Motivational example

Data from van 't Veer et al. (2002):

- 78 breast cancer patients;
- 44 remained disease-free for more than 5 years;
- 34 developed metastases within 5 years;
- Gene expression levels of 24,479 oligonucleotides were measured on each individual;
- Goal: identify important genes that are associated with the metastases risk.
- Our concern: different genes are identified using different data sets due to sampling variation.

Motivational example

Data from van 't Veer et al. (2002):

- 78 breast cancer patients;
- 44 remained disease-free for more than 5 years;
- 34 developed metastases within 5 years;
- Gene expression levels of 24,479 oligonucleotides were measured on each individual;
- Goal: identify important genes that are associated with the metastases risk.
- Our concern: different genes are identified using different data sets due to sampling variation.

Motivational example

Data from van 't Veer et al. (2002):

- 78 breast cancer patients;
- 44 remained disease-free for more than 5 years;
- 34 developed metastases within 5 years;
- Gene expression levels of 24,479 oligonucleotides were measured on each individual;
- Goal: identify important genes that are associated with the metastases risk.
- Our concern: different genes are identified using different data sets due to sampling variation.

Motivational example

Simulation experiment:

- Randomly split the data into two halves. Each consists of 22 patients with good prognosis and 17 with poor prognosis;
- Called sample 1 and sample 2;
- For each sample and each gene, compute the correlation coefficient between the expression levels and the prognosis outcome;
- Important genes are those with high absolute correlation values.

Motivational example

Simulation experiment:

- Randomly split the data into two halves. Each consists of 22 patients with good prognosis and 17 with poor prognosis;
- Called sample 1 and sample 2;
- For each sample and each gene, compute the correlation coefficient between the expression levels and the prognosis outcome;
- Important genes are those with high absolute correlation values.

Motivational example

Simulation experiment:

- Randomly split the data into two halves. Each consists of 22 patients with good prognosis and 17 with poor prognosis;
- Called sample 1 and sample 2;
- For each sample and each gene, compute the correlation coefficient between the expression levels and the prognosis outcome;
- Important genes are those with high absolute correlation values.

Motivational example

Simulation experiment:

- Randomly split the data into two halves. Each consists of 22 patients with good prognosis and 17 with poor prognosis;
- Called sample 1 and sample 2;
- For each sample and each gene, compute the correlation coefficient between the expression levels and the prognosis outcome;
- Important genes are those with high absolute correlation values.

Motivational example

- The original Kendall association correlation coefficient: 0.00522 (p-value=0.112);
- Say, top 5000 genes are regarded as important;
- Take the top 5000 genes using ALL 78 patients, only 17% are regarded as important in both sample 1 and sample 2 (the black dots);
- Sample 1 and sample 2 have 1,131 top genes in common;
- The modified test obtains a p-value = 0.0000578

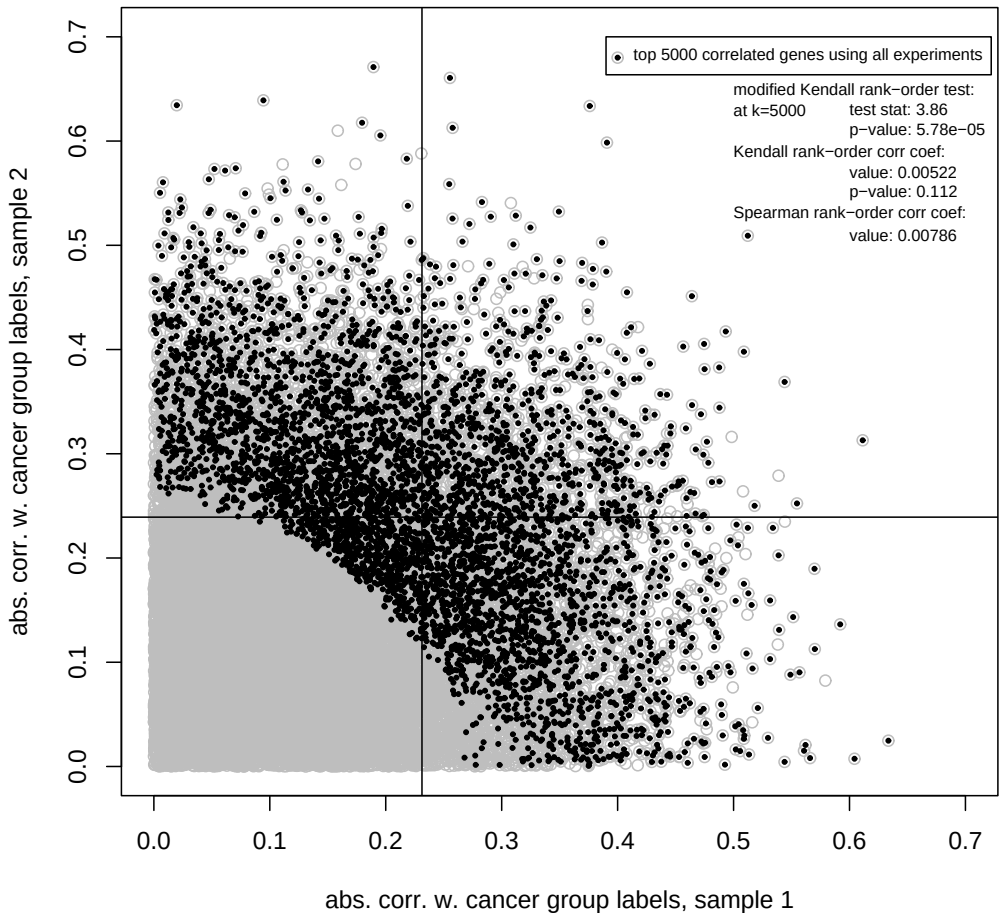


Fig. 1. An example of two microarray experiment samples showing weak association.

Association between two rankings on the same objects

- X_i, $i = 1, \ldots, n$ and Y_i, $i = 1, \ldots, n$ be two sets of independent rankings of n objects.
- (Throughout, we discuss rankings in decreasing order.)
- Denote α_i as the *importance* of object i.
- X_i's and Y_i's are random representations of the true ranking, $\text{Rank}(\alpha_i)$.
- We assume that

$$\begin{aligned} X_i &= \text{Rank}(\alpha_i + \varepsilon_i), \\ Y_i &= \text{Rank}(\alpha_i + \delta_i), \end{aligned}$$

  where $\varepsilon_i \overset{iid}{\sim} F$ and $\delta_i \overset{iid}{\sim} G$.
- α's, F and G are introduced for the convenience of discussion.

Association between two rankings on the same objects

- X_i, $i = 1, \ldots, n$ and Y_i, $i = 1, \ldots, n$ be two sets of independent rankings of n objects.
- (Throughout, we discuss rankings in decreasing order.)
- Denote α_i as the *importance* of object i.
- X_i's and Y_i's are random representations of the true ranking, $\text{Rank}(\alpha_i)$.
- We assume that

$$\begin{aligned} X_i &= \text{Rank}(\alpha_i + \varepsilon_i), \\ Y_i &= \text{Rank}(\alpha_i + \delta_i), \end{aligned}$$

  where $\varepsilon_i \overset{iid}{\sim} F$ and $\delta_i \overset{iid}{\sim} G$.
- α's, F and G are introduced for the convenience of discussion.

Association between two rankings on the same objects

- X_i, $i = 1, \ldots, n$ and Y_i, $i = 1, \ldots, n$ be two sets of independent rankings of n objects.
- (Throughout, we discuss rankings in decreasing order.)
- Denote α_i as the *importance* of object i.
- X_i's and Y_i's are random representations of the true ranking, $\text{Rank}(\alpha_i)$.
- We assume that

$$\begin{aligned} X_i &= \text{Rank}(\alpha_i + \varepsilon_i), \\ Y_i &= \text{Rank}(\alpha_i + \delta_i), \end{aligned}$$

  where $\varepsilon_i \overset{iid}{\sim} F$ and $\delta_i \overset{iid}{\sim} G$.
- α's, F and G are introduced for the convenience of discussion.

Association between two rankings on the same objects

- X_i, $i = 1, \ldots, n$ and Y_i, $i = 1, \ldots, n$ be two sets of independent rankings of n objects.
- (Throughout, we discuss rankings in decreasing order.)
- Denote α_i as the *importance* of object i.
- X_i's and Y_i's are random representations of the true ranking, Rank (α_i).
- We assume that

$$\begin{aligned} X_i &= \text{Rank}(\alpha_i + \varepsilon_i), \\ Y_i &= \text{Rank}(\alpha_i + \delta_i), \end{aligned}$$

  where $\varepsilon_i \overset{iid}{\sim} F$ and $\delta_i \overset{iid}{\sim} G$.
- α's, F and G are introduced for the convenience of discussion.

Association between two rankings on the same objects

- X_i, $i = 1, \ldots, n$ and Y_i, $i = 1, \ldots, n$ be two sets of independent rankings of n objects.
- (Throughout, we discuss rankings in decreasing order.)
- Denote α_i as the *importance* of object i.
- X_i's and Y_i's are random representations of the true ranking, Rank (α_i).
- We assume that

$$\begin{aligned} X_i &= \text{Rank}(\alpha_i + \varepsilon_i), \\ Y_i &= \text{Rank}(\alpha_i + \delta_i), \end{aligned}$$

  where $\varepsilon_i \overset{iid}{\sim} F$ and $\delta_i \overset{iid}{\sim} G$.
- α's, F and G are introduced for the convenience of discussion.

Association between two rankings on the same objects

- X_i, $i = 1, \ldots, n$ and Y_i, $i = 1, \ldots, n$ be two sets of independent rankings of n objects.
- (Throughout, we discuss rankings in decreasing order.)
- Denote α_i as the *importance* of object i.
- X_i's and Y_i's are random representations of the true ranking, Rank (α_i).
- We assume that

$$\begin{aligned} X_i &= \text{Rank}(\alpha_i + \varepsilon_i), \\ Y_i &= \text{Rank}(\alpha_i + \delta_i), \end{aligned}$$

  where $\varepsilon_i \overset{iid}{\sim} F$ and $\delta_i \overset{iid}{\sim} G$.
- α's, F and G are introduced for the convenience of discussion.

Association between two rankings on the same objects

- Without loss of generality, assume that the objects are arranged in the order of their importance, that is

$$\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n.$$

- If $\alpha_1 = \alpha_2 = \cdots = \alpha_n$, X is reduced to Rank(ε), would be independent of ranking Y=Rank(δ).
- If $\alpha_1 > \alpha_2 > \cdots > \alpha_n$, X and Y will be positively correlated and the degree of correlation depends on the random variation of ε's and δ's.
- The correlation between rankings X and Y can be used to measure the variation among the objects' importance (signal), relative to the amount of noises.

Association between two rankings on the same objects

- Without loss of generality, assume that the objects are arranged in the order of their importance, that is

$$\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n.$$

- If $\alpha_1 = \alpha_2 = \cdots = \alpha_n$, X is reduced to Rank(ε), would be independent of ranking Y=Rank(δ).
- If $\alpha_1 > \alpha_2 > \cdots > \alpha_n$, X and Y will be positively correlated and the degree of correlation depends on the random variation of ε's and δ's.
- The correlation between rankings X and Y can be used to measure the variation among the objects' importance (signal), relative to the amount of noises.

Association between two rankings on the same objects

- Without loss of generality, assume that the objects are arranged in the order of their importance, that is

$$\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n.$$

- If $\alpha_1 = \alpha_2 = \cdots = \alpha_n$, X is reduced to Rank(ε), would be independent of ranking Y=Rank(δ).
- If $\alpha_1 > \alpha_2 > \cdots > \alpha_n$, X and Y will be positively correlated and the degree of correlation depends on the random variation of ε's and δ's.
- The correlation between rankings X and Y can be used to measure the variation among the objects' importance (signal), relative to the amount of noises.

Association between two rankings on the same objects

- Without loss of generality, assume that the objects are arranged in the order of their importance, that is

$$\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n.$$

- If $\alpha_1 = \alpha_2 = \cdots = \alpha_n$, X is reduced to Rank(ε), would be independent of ranking Y=Rank(δ).
- If $\alpha_1 > \alpha_2 > \cdots > \alpha_n$, X and Y will be positively correlated and the degree of correlation depends on the random variation of ε's and δ's.
- The correlation between rankings X and Y can be used to measure the variation among the objects' importance (signal), relative to the amount of noises.

Kendall rank-order correlation coefficient

The Kendall rank-order correlation coefficient (Kendall, 1955) is formulated as

$$T = \frac{\#\text{ agreements} - \#\text{ disagreements}}{\text{total number of pairs}}$$

- Consider all possible *pairs* of (X_i, X_j) in which X_i is lower than X_j, if
  - if Y_i is lower than Y_j, it is then an agreement,
  - if Y_i is higher than Y_j, it is then a disagreement.
-

$$\#\text{ agreements} = \sum_{i=1}^{n} \sum_{i \neq j} \mathbf{1}_{(X_i < X_j)} \mathbf{1}_{(Y_i < Y_j)},$$

$$\#\text{ disagreements} = \sum_{i=1}^{n} \sum_{i \neq j} \mathbf{1}_{(X_i < X_j)} \mathbf{1}_{(Y_i > Y_j)}.$$

Kendall rank-order correlation coefficient

The Kendall rank-order correlation coefficient (Kendall, 1955) is formulated as

$$T = \frac{\#\text{ agreements} - \#\text{ disagreements}}{\text{total number of pairs}}$$

- Consider all possible *pairs* of (X_i, X_j) in which X_i is lower than X_j, if
  - if Y_i is lower than Y_j, it is then an *agreement*;
  - if Y_i is higher than Y_j, it is then an *disagreement*.
-

$$\#\text{ agreements} = \sum_{i=1}^{n} \sum_{i \neq j} \mathbf{1}_{(X_i < X_j)} \mathbf{1}_{(Y_i < Y_j)},$$

$$\#\text{ disagreements} = \sum_{i=1}^{n} \sum_{i \neq j} \mathbf{1}_{(X_i < X_j)} \mathbf{1}_{(Y_i > Y_j)}.$$

Kendall rank-order correlation coefficient

The Kendall rank-order correlation coefficient (Kendall, 1955) is formulated as

$$T = \frac{\#\text{ agreements} - \#\text{ disagreements}}{\text{total number of pairs}}$$

- Consider all possible *pairs* of (X_i, X_j) in which X_i is lower than X_j, if
  - if Y_i is lower than Y_j, it is then an *agreement*;
  - if Y_i is higher than Y_j, it is then an *disagreement*.
-

$$\#\text{ agreements} = \sum_{i=1}^{n} \sum_{i \neq j} \mathbf{1}_{(X_i < X_j)} \mathbf{1}_{(Y_i < Y_j)},$$

$$\#\text{ disagreements} = \sum_{i=1}^{n} \sum_{i \neq j} \mathbf{1}_{(X_i < X_j)} \mathbf{1}_{(Y_i > Y_j)}.$$

Kendall rank-order correlation coefficient

The Kendall rank-order correlation coefficient (Kendall, 1955) is formulated as

$$T = \frac{\#\text{ agreements} - \#\text{ disagreements}}{\text{total number of pairs}}$$

- Consider all possible *pairs* of (X_i, X_j) in which X_i is lower than X_j, if
  - if Y_i is lower than Y_j, it is then an *agreement*;
  - if Y_i is higher than Y_j, it is then an *disagreement*.
- $$\#\text{ agreements} = \sum_{i=1}^{n} \sum_{i \neq j} \mathbf{1}_{(X_i < X_j)} \mathbf{1}_{(Y_i < Y_j)},$$
- $$\#\text{ disagreements} = \sum_{i=1}^{n} \sum_{i \neq j} \mathbf{1}_{(X_i < X_j)} \mathbf{1}_{(Y_i > Y_j)}.$$

Kendall rank-order correlation coefficient

The Kendall rank-order correlation coefficient (Kendall, 1955) is formulated as

$$T = \frac{\#\text{ agreements} - \#\text{ disagreements}}{\text{total number of pairs}}$$

- Consider all possible *pairs* of (X_i, X_j) in which X_i is lower than X_j, if
  - if Y_i is lower than Y_j, it is then an *agreement*;
  - if Y_i is higher than Y_j, it is then an *disagreement*.
-

$$\#\text{ agreements} = \sum_{i=1}^{n} \sum_{i \neq j} \mathbf{1}_{(X_i < X_j)} \mathbf{1}_{(Y_i < Y_j)},$$

$$\#\text{ disagreements} = \sum_{i=1}^{n} \sum_{i \neq j} \mathbf{1}_{(X_i < X_j)} \mathbf{1}_{(Y_i > Y_j)}.$$

Kendall rank-order correlation coefficient

- If there are no ties,
  # agreements + # disagreements $= n(n-1)/2$.
- Under the null hypothesis,

$$\begin{aligned} \mathrm{E}(\#\text{ agreements}) &= \mathrm{E}(\#\text{ disagreements}) = \frac{1}{4}n(n-1), \\ \mathrm{var}(\#\text{agreements}) &= \frac{1}{16}\left(\frac{4n}{9}+\frac{10}{9}\right)n(n-1). \end{aligned}$$

$$\mathrm{E}(T) = 0 \text{ and } \mathrm{var}(T) = \frac{2(2n+5)}{9n(n-1)}.$$

Kendall rank-order correlation coefficient

- If there are no ties,
  $\#$ agreements $+ \#$ disagreements $= n(n-1)/2$.
- Under the null hypothesis,

$$\begin{aligned} \mathrm{E}(\#\text{ agreements}) &= \mathrm{E}(\#\text{ disagreements}) = \frac{1}{4}n(n-1), \\ \mathrm{var}(\#\text{agreements}) &= \frac{1}{16}\left(\frac{4n}{9}+\frac{10}{9}\right)n(n-1). \end{aligned}$$

$$\mathrm{E}(T) = 0 \text{ and } \mathrm{var}(T) = \frac{2(2n+5)}{9n(n-1)}.$$

When the number of the objects is large

- In most current studies, the number of objects is large, while the number of objects with higher importance is small.
  - $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_n \equiv \alpha$
  - versus a *local* alternative
    $H_a : \exists 1 \leq k_0 \ll n$, s.t.,
    $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_{k_0} > \alpha_{k_0+1} = \alpha_{k_0+2} = \cdots = \alpha_{n-1} = \alpha_n$.
- The strength of association measured by the original statistic is weakened by the large number of objects with undifferentiated importance.

When the number of the objects is large

- In most current studies, the number of objects is large, while the number of objects with higher importance is small.
  - $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_n \equiv \alpha$
  - versus a *local* alternative
    $H_a : \exists 1 \leq k_0 \ll n$, s.t.,
    $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_{k_0} > \alpha_{k_0+1} = \alpha_{k_0+2} = \cdots = \alpha_{n-1} = \alpha_n$.
- The strength of association measured by the original statistic is weakened by the large number of objects with undifferentiated importance.

When the number of the objects is large

- In most current studies, the number of objects is large, while the number of objects with higher importance is small.
  - $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_n \equiv \alpha$
  - versus a *local* alternative
    $H_a : \exists 1 \leq k_0 \ll n$, s.t.,
    $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_{k_0} > \alpha_{k_0+1} = \alpha_{k_0+2} = \cdots = \alpha_{n-1} = \alpha_n$.
- The strength of association measured by the original statistic is weakened by the large number of objects with undifferentiated importance.

When the number of the objects is large

- In most current studies, the number of objects is large, while the number of objects with higher importance is small.
  - $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_n \equiv \alpha$
  - versus a *local* alternative
    $H_a : \exists 1 \leq k_0 \ll n$, s.t.,
    $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_{k_0} > \alpha_{k_0+1} = \alpha_{k_0+2} = \cdots = \alpha_{n-1} = \alpha_n$.
- The strength of association measured by the original statistic is weakened by the large number of objects with undifferentiated importance.

Modified Kendall association test

- Consider the truncated rankings $X_i^c = \min(X_i, k)$.
- The number of agreements can then be computed and tested on the truncated X and Y
- Using the truncated rankings, the noises from the objects with no signals are reduced.

$$\#\text{ agreements} = \sum_{i=1}^{n} \sum_{i \neq j} \mathbf{1}_{(X_i^c < X_j^c)} \mathbf{1}_{(Y_i^c < Y_j^c)}$$

and under the null hypothesis

$$\mathrm{E}(\#\text{ agreements}) = \frac{1}{4} n(n-1) \left(1 - \frac{\binom{n-k+1}{2}}{\binom{n}{2}} \right)^2$$

Modified Kendall association test

- Consider the truncated rankings $X_i^c = \min(X_i, k)$.
- The number of agreements can then be computed and tested on the truncated X and Y
- Using the truncated rankings, the noises from the objects with no signals are reduced.

$$\#\text{ agreements} = \sum_{i=1}^{n} \sum_{i \neq j} \mathbf{1}_{(X_i^c < X_j^c)} \mathbf{1}_{(Y_i^c < Y_j^c)}$$

and under the null hypothesis

$$\mathrm{E}(\#\text{ agreements}) = \frac{1}{4} n(n-1) \left(1 - \frac{\binom{n-k+1}{2}}{\binom{n}{2}} \right)^2$$

Modified Kendall association test

- Consider the truncated rankings $X_i^c = \min(X_i, k)$.
- The number of agreements can then be computed and tested on the truncated X and Y
- Using the truncated rankings, the noises from the objects with no signals are reduced.

$$\#\text{ agreements} = \sum_{i=1}^{n} \sum_{i \neq j} \mathbf{1}_{(X_i^c < X_j^c)} \mathbf{1}_{(Y_i^c < Y_j^c)}$$

and under the null hypothesis

$$\mathrm{E}(\#\text{ agreements}) = \frac{1}{4} n(n-1) \left(1 - \frac{\binom{n-k+1}{2}}{\binom{n}{2}} \right)^2$$

Modified Kendall association test

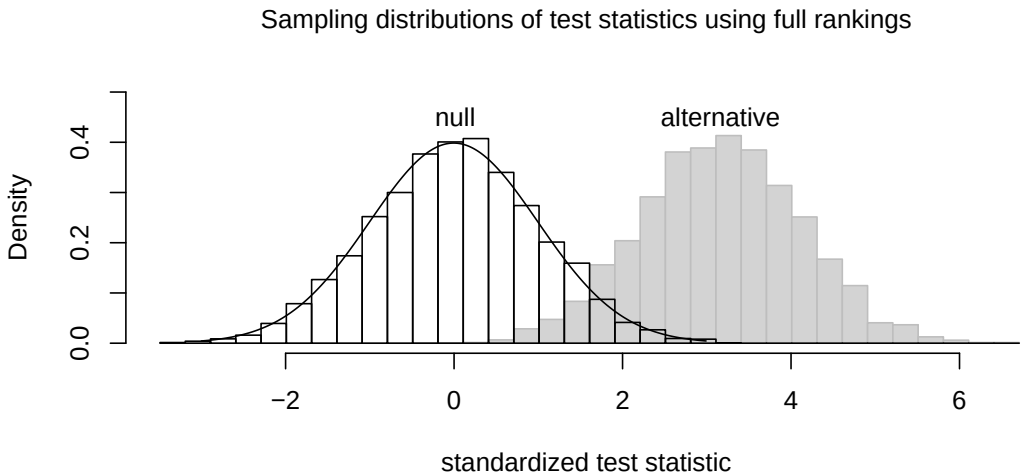
$$
\begin{aligned}
& \operatorname{var}\left(\sum_{i=1}^{n}\sum_{j\neq i}\mathbf{1}_{(X_i^c<X_j^c)}\mathbf{1}_{(Y_i^c<Y_j^c)}\right) \\
= \quad & n(n-1)\left\{\frac{1}{4}n(n-1)\left(1-\frac{(n-k+1)(n-k)}{n(n-1)}\right)^2\right. \\
& +(n-2)(n-3)\left(\frac{1}{4}\frac{\binom{k-1}{4}}{\binom{n}{4}}+\frac{1}{4}\frac{\binom{k-1}{3}\binom{n-k+1}{1}}{\binom{n}{4}}+\frac{1}{6}\frac{\binom{k-1}{2}\binom{n-k+1}{2}}{\binom{n}{4}}\right)^2 \\
& +(n-2)\frac{1}{6}\left(\frac{\binom{k}{3}}{\binom{n}{3}}+\frac{\binom{k-1}{2}\binom{n-k}{1}}{\binom{n}{3}}\right)^2 \\
& +(n-2)\frac{1}{9}\left(1-\frac{\binom{n-k+1}{3}}{\binom{n}{3}}\right)^2 \\
& \left.-\frac{1}{16}(n^2-n)\left(1-\frac{(n-k+1)(n-k)}{n(n-1)}\right)^4\right\}.
\end{aligned}
$$

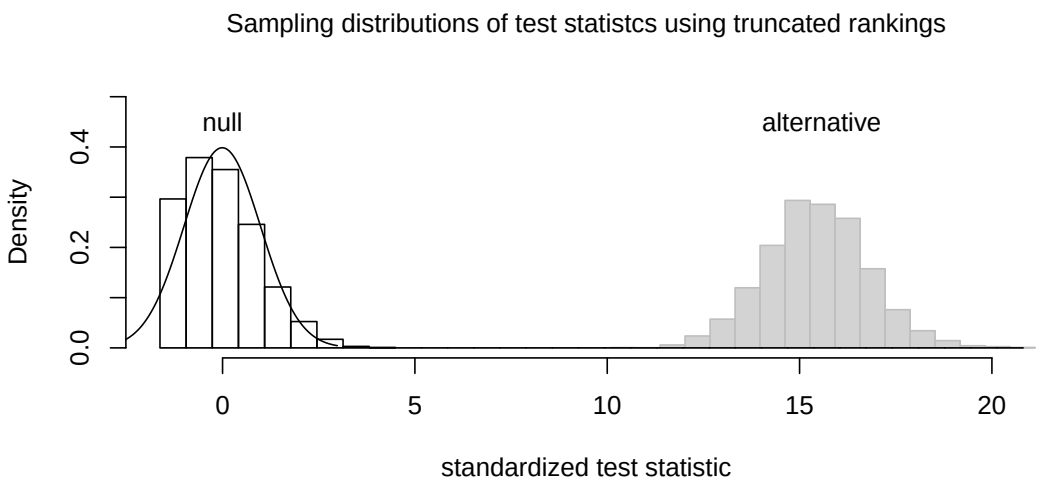
Modified Kendall association test

- The modified Kendall rank-order test statistic is defined as

$$T^c = \frac{\#\text{ agreements} - \mathrm{E}(\#\text{ agreements})}{\sqrt{\mathrm{Var}\,(\#\text{ agreements})}}.$$

Modified Kendall association test





Simulation setup

- n objects, among which the top k_0 objects are important and with linearly increasing importance.
- Signal: δ is the highest importance value.
- Noises with standard deviation σ is added to the observed importance.

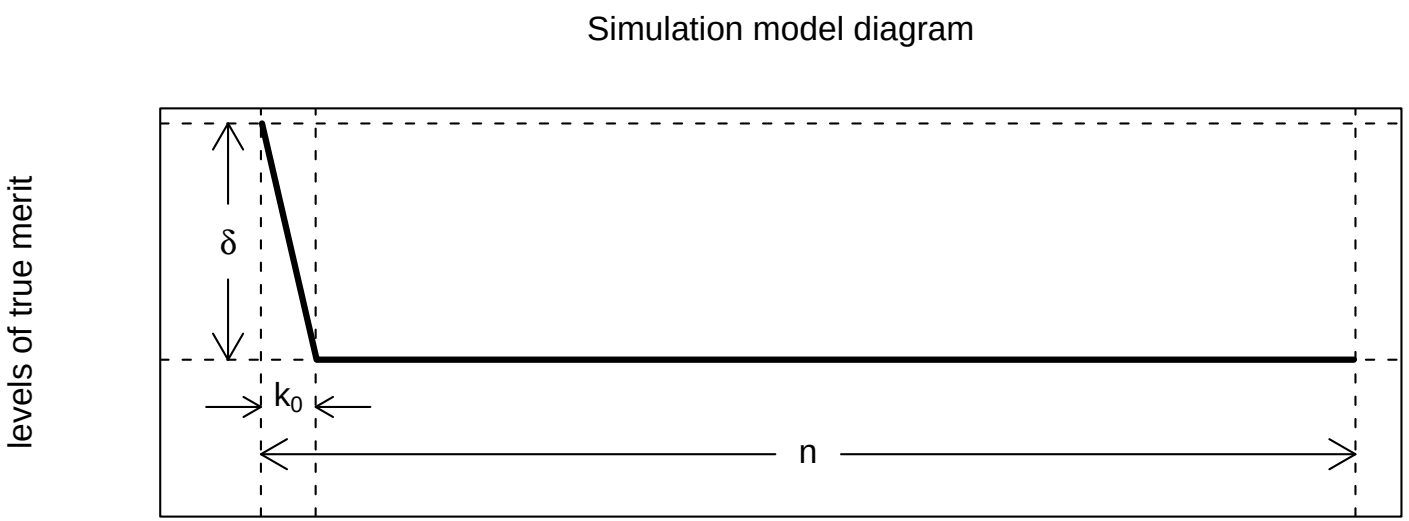


Fig. 3. Alternative model used for simulations

Simulation setup

- n objects, among which the top k_0 objects are important and with linearly increasing importance.
- Signal: δ is the highest importance value.
- Noises with standard deviation σ is added to the observed importance.

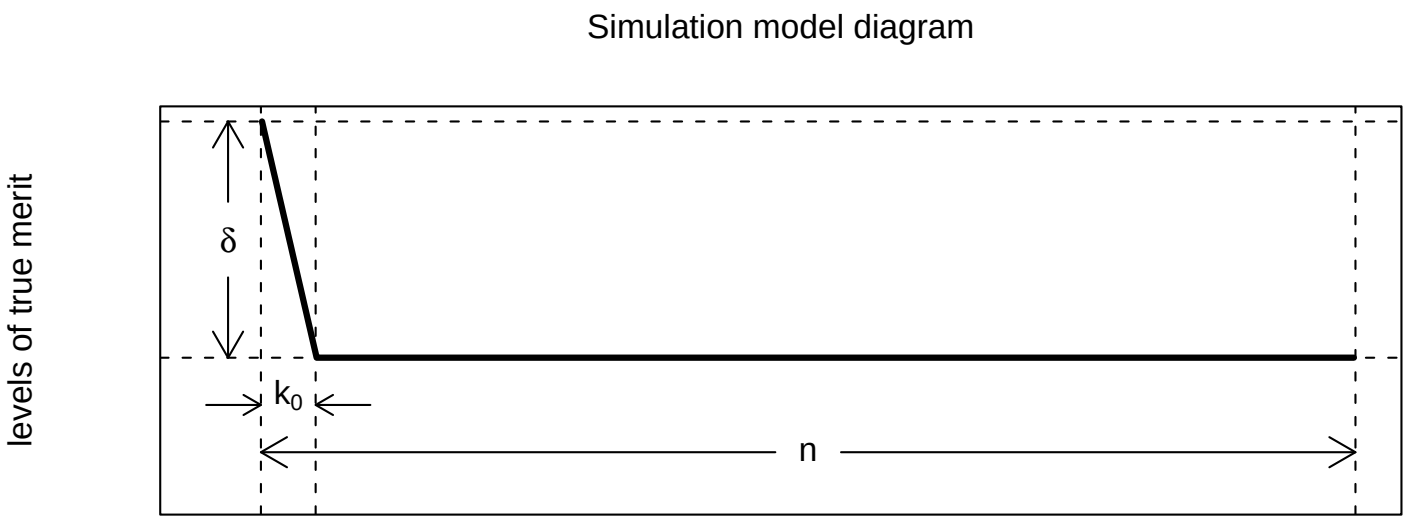


Fig. 3. Alternative model used for simulations

Simulation setup

- n objects, among which the top k_0 objects are important and with linearly increasing importance.
- Signal: δ is the highest importance value.
- Noises with standard deviation σ is added to the observed importance.

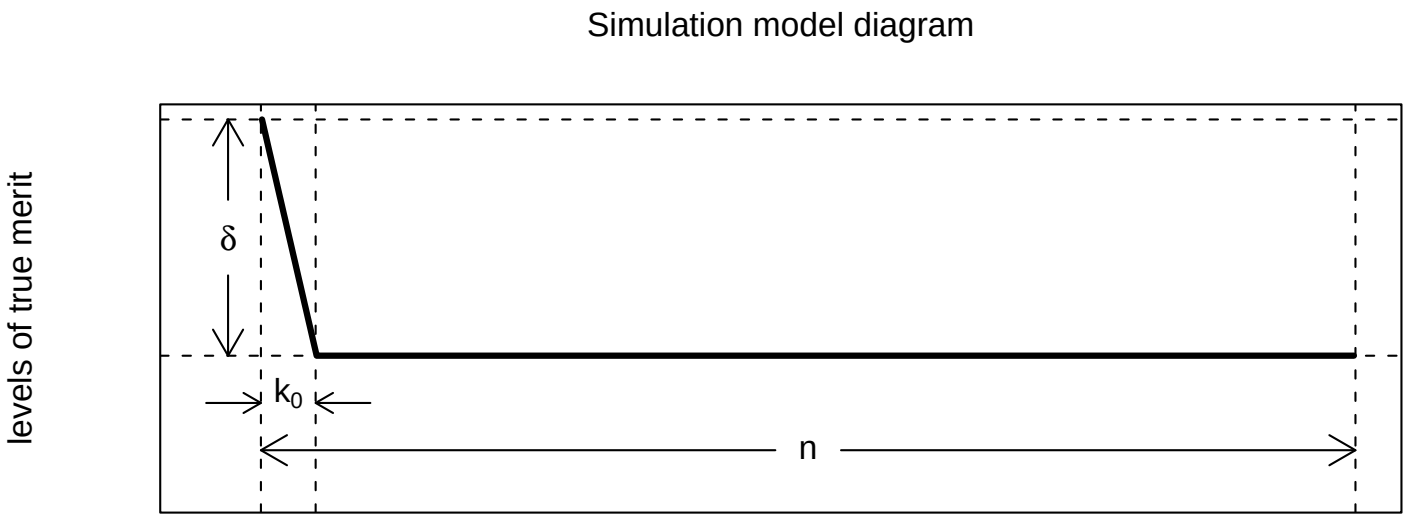
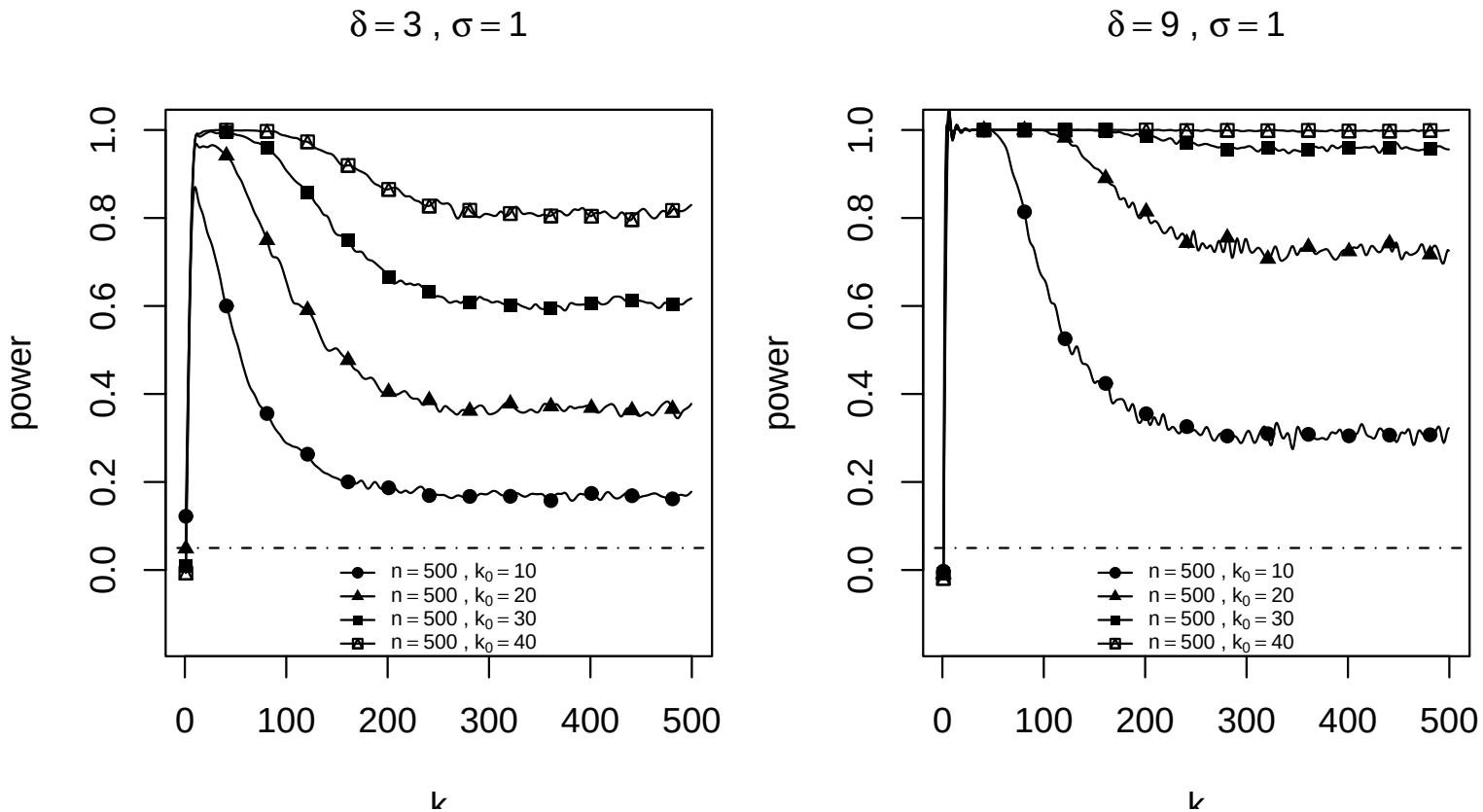


Fig. 3. Alternative model used for simulations

Performance of the modified test under different signal-to-noise ratios



Computational notes

- The computational complexity of a single rank-order association test statistic is of $O(n^2)$, comparing with the computation of the correlation coefficient at $O(n)$. However, $T^c(k+1)$ can be updated from $T^c(k)$.
- From $\mathbf{1}_{(\min(X_i,k+1)<\min(X_j,k+1))}\mathbf{1}_{((\min(Y_i,k+1)<\min(Y_j,k+1))}$ to $\mathbf{1}_{(\min(X_i,k)<\min(X_j,k))}\mathbf{1}_{((\min(Y_i,k)<\min(Y_j,k))}$, only a small number of elements change values.
- Using such a sequential update, all $T^c(k)$, $k = 1, \ldots, n$ can be computed in an operation of $O(n^2)$ complexity.

Computational notes

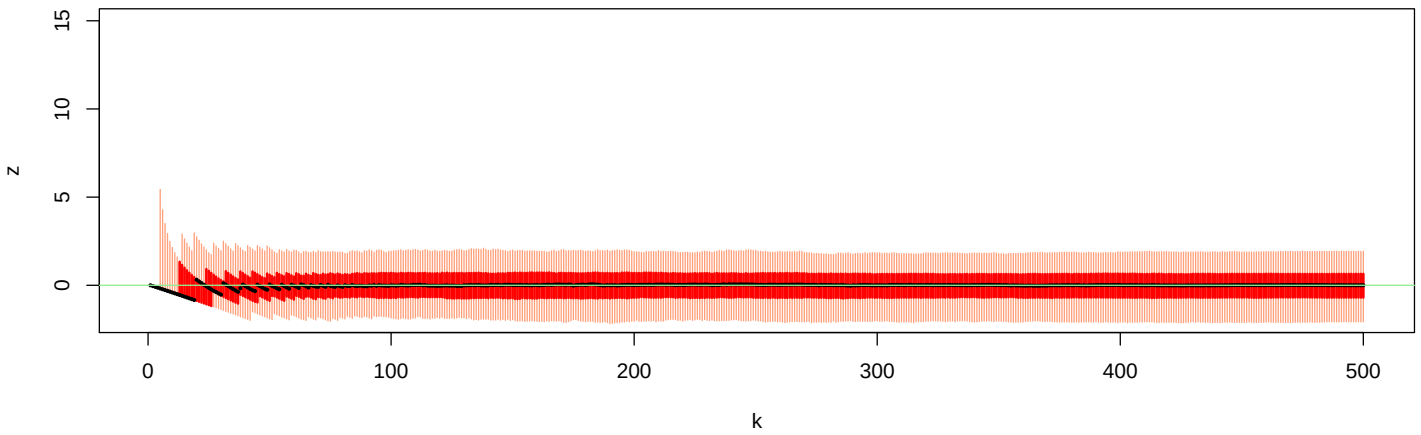
- The computational complexity of a single rank-order association test statistic is of $O(n^2)$, comparing with the computation of the correlation coefficient at $O(n)$. However, $T^c(k+1)$ can be updated from $T^c(k)$.
- From $\mathbf{1}_{(\min(X_i,k+1)<\min(X_j,k+1))}\mathbf{1}_{((\min(Y_i,k+1)<\min(Y_j,k+1))}$ to $\mathbf{1}_{(\min(X_i,k)<\min(X_j,k))}\mathbf{1}_{((\min(Y_i,k)<\min(Y_j,k))}$, only a small number of elements change values.
- Using such a sequential update, all $T^c(k)$, $k = 1, \ldots, n$ can be computed in an operation of $O(n^2)$ complexity.

Computational notes

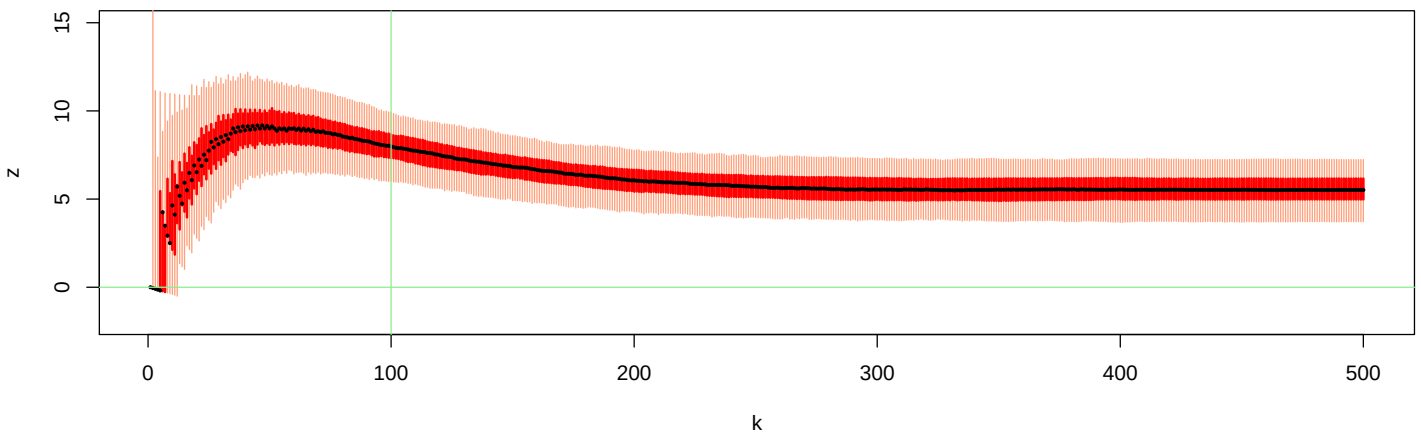
- The computational complexity of a single rank-order association test statistic is of $O(n^2)$, comparing with the computation of the correlation coefficient at $O(n)$. However, $T^c(k+1)$ can be updated from $T^c(k)$.
- From $\mathbf{1}_{(\min(X_i,k+1)<\min(X_j,k+1))}\mathbf{1}_{((\min(Y_i,k+1)<\min(Y_j,k+1))}$ to $\mathbf{1}_{(\min(X_i,k)<\min(X_j,k))}\mathbf{1}_{((\min(Y_i,k)<\min(Y_j,k))}$, only a small number of elements change values.
- Using such a sequential update, all $T^c(k)$, $k = 1, \ldots, n$ can be computed in an operation of $O(n^2)$ complexity.

Sequences of $T^c(k)$

(a) under the null hypothesis

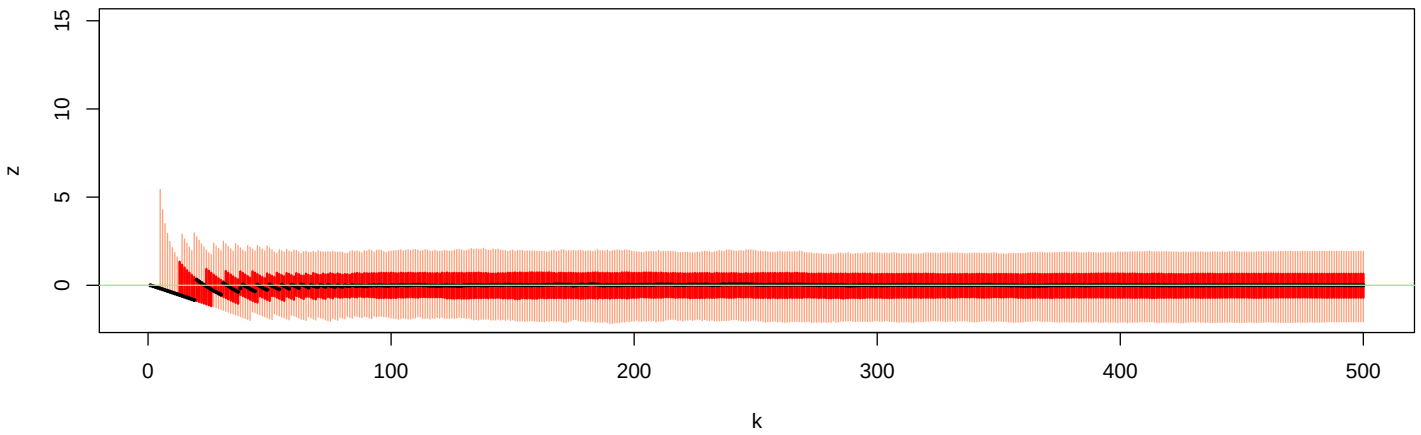


(b) $\delta = 3$, $\sigma = 1$

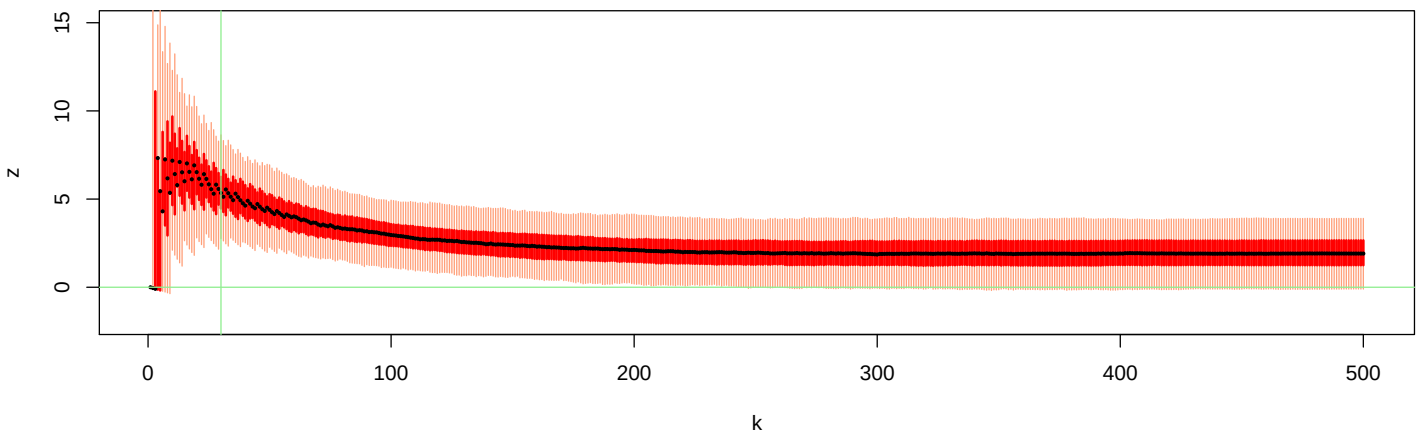


Sequences of $T^c(k)$

(a) under the null hypothesis

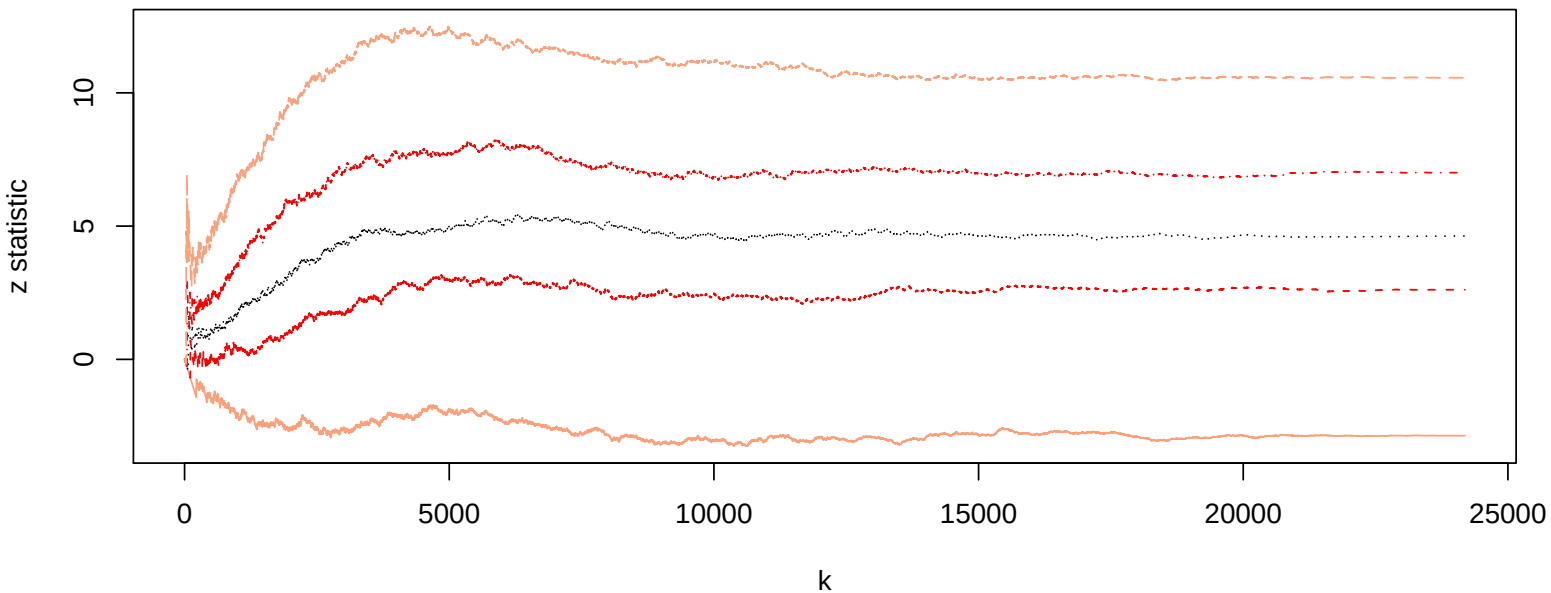


(b) $\delta = 3$, $\sigma = 1$



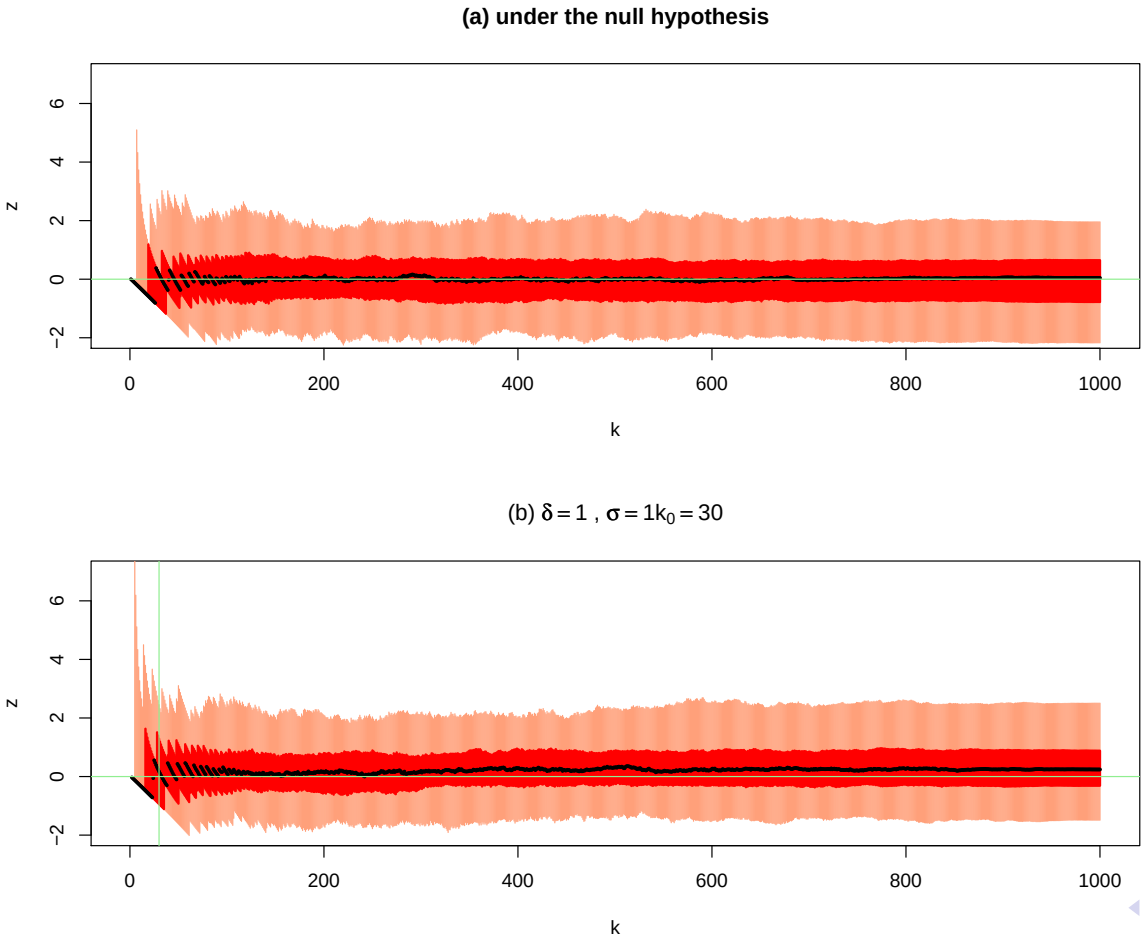
Back to the breast cancer example

100 simulation experiments using randomly partitioned samples.



Pattern due to weak signals?

Simulation with small number of important objects and low signal-to-noise ratio.



Application to an eQTL example

- Affymetrix Human Focus Arrays, with 8500 transcripts were measured on 194 individuals in 14 CEPH families (Morley et al., 2004).
- Genotypes of these CEPH individuals on 2882 SNPs across the genome were obtained from The SNP Consortium (http://snp.cshl.org/linkage_maps/).
- We examined 18 transcripts that are related to several candidate genes of breast cancer.

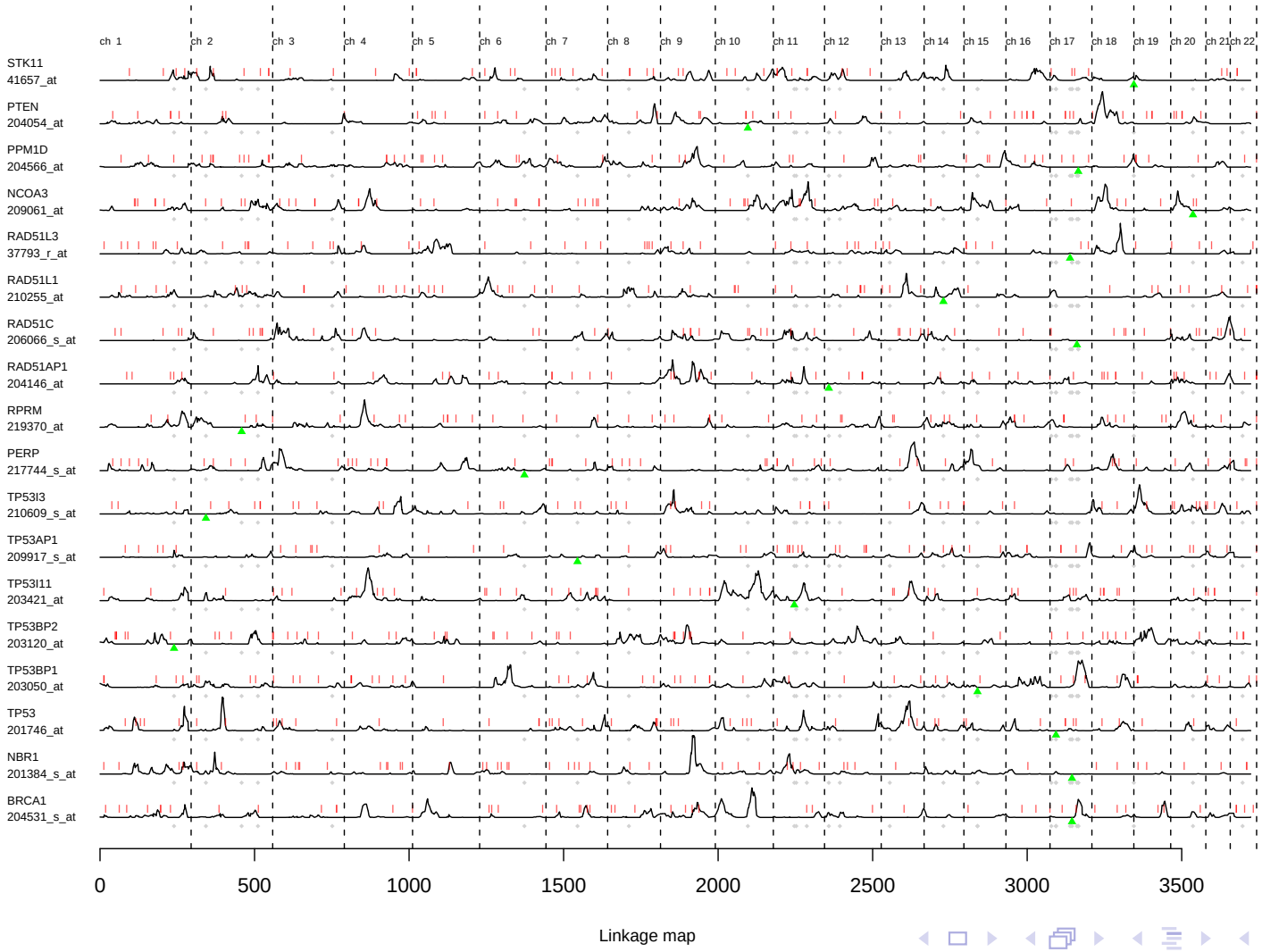
Application to an eQTL example

- Affymetrix Human Focus Arrays, with 8500 transcripts were measured on 194 individuals in 14 CEPH families (Morley et al., 2004).
- Genotypes of these CEPH individuals on 2882 SNPs across the genome were obtained from The SNP Consortium (http://snp.cshl.org/linkage_maps/).
- We examined 18 transcripts that are related to several candidate genes of breast cancer.

Application to an eQTL example

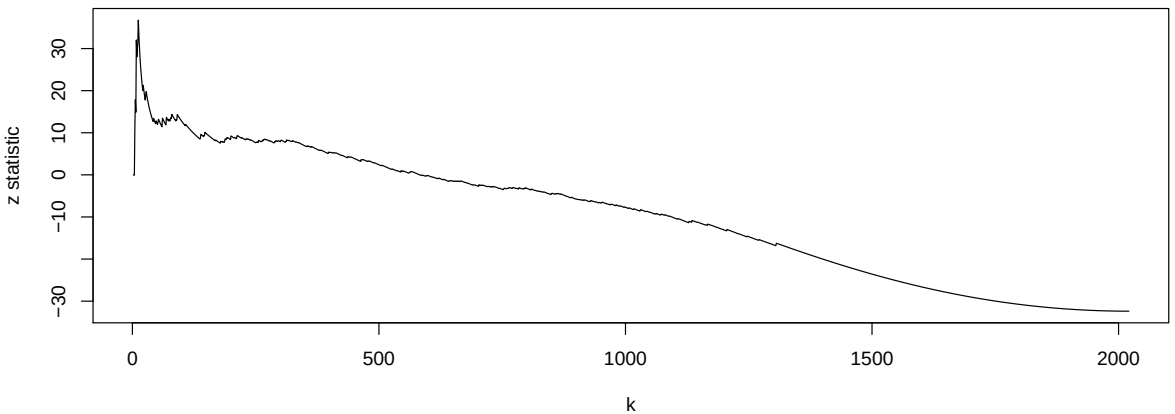
- Affymetrix Human Focus Arrays, with 8500 transcripts were measured on 194 individuals in 14 CEPH families (Morley et al., 2004).
- Genotypes of these CEPH individuals on 2882 SNPs across the genome were obtained from The SNP Consortium (http://snp.cshl.org/linkage_maps/).
- We examined 18 transcripts that are related to several candidate genes of breast cancer.

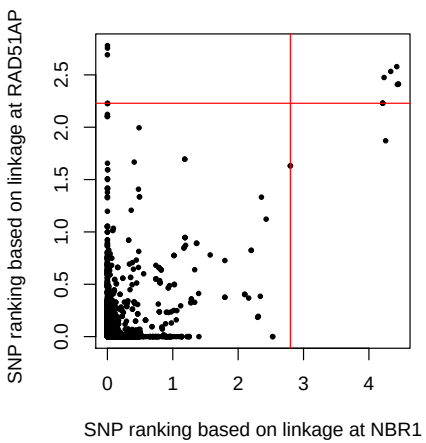
Application to an eQTL example



Application to an eQTL example

Compare linkage signals for two gene expression traits (NBR1 and RAD51AP). (Overlapped linkage signals indicate evidence for co-regulation of these two transcripts.)

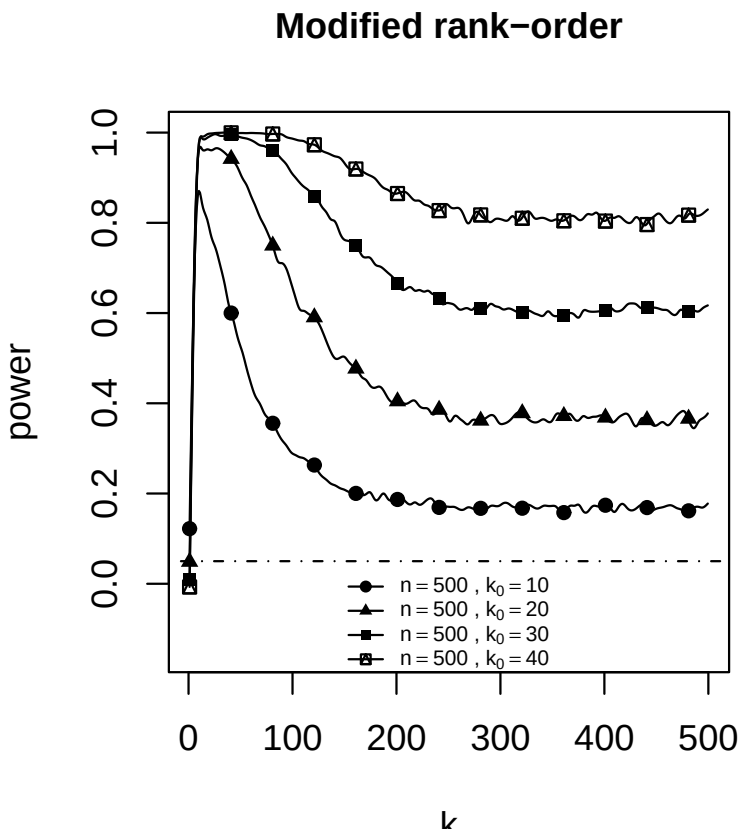


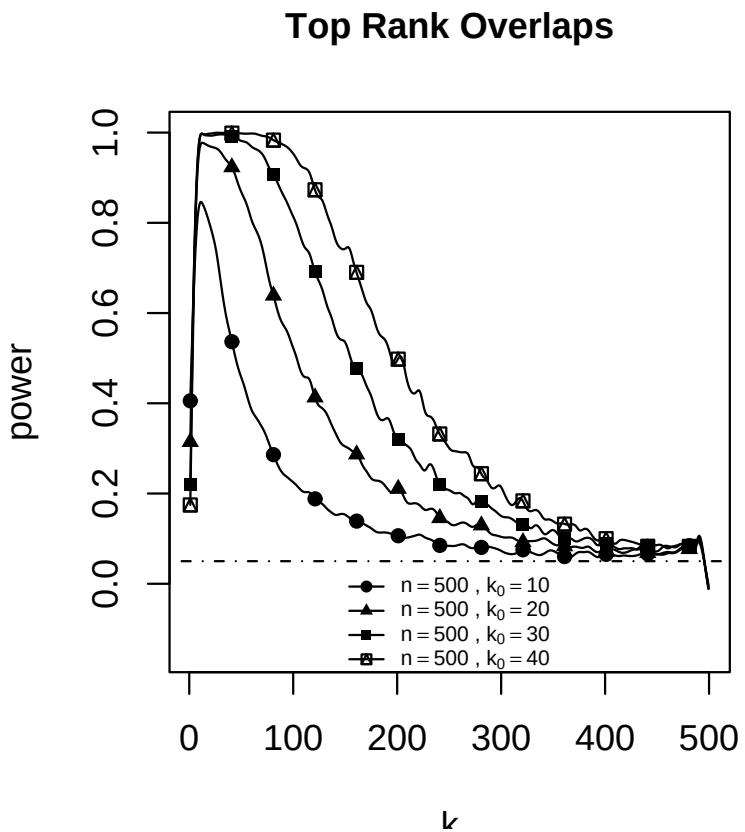


Application to an eQTL example

From OMIM (Online Mendelian Inheritance in Man):
"Dong et al. (2003) isolated a holoenzyme complex containing BRCA1 (113705), BRCA2, BARD1 (610593), and RAD51, which they called the BRCA1- and BRCA2-containing complex (BRCC). . . . . concluded that the BRCC is a ubiquitin E3 ligase that enhances cellular survival following DNA damage."

An alternative method: examine the extent of overlap





Conclusion and future work

- This modified association test removes noises from uninformative rank values and thus is more powerful in detecting the true signal.
- Due to the use of ranks, this test can be used to compare information extracted differently (such as linkage and association in gene mapping efforts) or on different scales (differently normalized gene expression experiments).
- If used on random partitions of a data set, the sequence of the test statistic contains information on the number of truly important objects.
- Future work on this project including more theoretical and computation evaluation of this test statistic, especially on the $T^c(k)$ sequences and its application to the selection of important objects and incorporation of multiple evaluations.

Conclusion and future work

- This modified association test removes noises from uninformative rank values and thus is more powerful in detecting the true signal.
- Due to the use of ranks, this test can be used to compare information extracted differently (such as linkage and association in gene mapping efforts) or on different scales (differently normalized gene expression experiments).
- If used on random partitions of a data set, the sequence of the test statistic contains information on the number of truly important objects.
- Future work on this project including more theoretical and computation evaluation of this test statistic, especially on the $T^c(k)$ sequences and its application to the selection of important objects and incorporation of multiple evaluations.

Conclusion and future work

- This modified association test removes noises from uninformative rank values and thus is more powerful in detecting the true signal.
- Due to the use of ranks, this test can be used to compare information extracted differently (such as linkage and association in gene mapping efforts) or on different scales (differently normalized gene expression experiments).
- If used on random partitions of a data set, the sequence of the test statistic contains information on the number of truly important objects.
- Future work on this project including more theoretical and computation evaluation of this test statistic, especially on the $T^c(k)$ sequences and its application to the selection of important objects and incorporation of multiple evaluations.

Conclusion and future work

- This modified association test removes noises from uninformative rank values and thus is more powerful in detecting the true signal.
- Due to the use of ranks, this test can be used to compare information extracted differently (such as linkage and association in gene mapping efforts) or on different scales (differently normalized gene expression experiments).
- If used on random partitions of a data set, the sequence of the test statistic contains information on the number of truly important objects.
- Future work on this project including more theoretical and computation evaluation of this test statistic, especially on the $T^c(k)$ sequences and its application to the selection of important objects and incorporation of multiple evaluations.

Thank you!