Comparative Chi-Square Analysis of Interactions User Guide (CPχ^2 Version 2013-10-04)

The Song Lab
Department of Computer Science
New Mexico State University
Las Cruces, New Mexico

Comparative chi-square analysis (CPχ^2) identifies conserved and differential interactions in two networks sharing the same set of nodes, from either steady-state or dynamic observations of the networks. The analysis is accomplished through computing chi-square statistics of heterogeneity and homogeneity of interactions across conditions. The input is two trajectory collection files observed for each network. Continuous values must be quantized first into discrete ones to use this analysis. The output includes two networks in which interactions towards each node are declared conserved, absolutely differential, relatively differential or null, across the two conditions.

Required parameters:

- -M comparison
- -1 {an input file name. No default value}
  This is an input trajectory collection file (TCF) under condition 1. The TCF file format is defined in `TrajColFile.pdf`.
- -2 {an input file name. No default value}
  This is an input trajectory collection file under condition 2.

Optional parameters:

- -A {a number. Default: 0.05}
  This number is the α level for the maximum false positive rate.
- -k {an integer value of 1 or 2. Default: 1}
  This integer can be only 1 or 2.
  1: do not allow different parents for the same child across conditions.
  2: allow different parents for the same child across conditions.
- -D {an output file. No default}
  This file contains a graph of the networks in Graphviz dot format, and can be rendered in various ways to produce visualization of the networks.
  Under the same network topology (-k 1), an edge is labelled differential if and only if it is incident upon a child node that is differential and has a changed working zone; an edge is labelled conserved if and only if it is incident upon a child node that is conserved with both the child and its parent set having a changed working zone.
  Under different network topologies (-k 2), an edge is labelled differential/conserved if and only if it is incident upon a child node that is differential/conserved. No working zone information is indicated for the child or its parents.
- -G {an output file name. No default}
  This is a tab-delimited file giving an 11-column table of statistics on differential and conserved interactions in the two networks.
- -K {a non-positive integer. Default: -1}
  This number specifies the maximum time delay in an interaction, also known as the Markovian order. For example, -5 implies the child value at time t can be influenced by the value of a parent at up to $t-5$, but not $t-6$ or older values. 0 considers association of parent and child values without a delay.

- -J {a non-positive integer. Default: -1}

  This number specifies the minimum time delay in an interaction. For example, -2 implies the child value at time t can be influenced by the value of a parent at at least $t-2$, but not $t-1$ or more recent values.

- -P {a positive integer. Default: 3}

  This parameter specifies the statistical test to compute the p-value of a contingency table.

  3: Pearson's chi-square test of independence

  Other options are not fully tested.

- -p {an integer. Default: network size}

  This is the maximum number of unique parents, considering time delay, allowed for each child combined for all conditions.

- -g {an integer. Default: 0}

  This is the minimum number of unique parents, considering time delay, allowed for each child combined for all conditions.

- -N {an positive integer. Default: 1}

  The node index to start the analysis.

- -w {a string}

  This argument specifies how to compute the null marginal distributions of each contingency table for each condition.

  INDEP_SUP: The *superset* of all conditions $\Pi = \Pi_1 \cup \cdots \cup \Pi_K$ is used to computed the heterogeneity chisq:

  χ^2_1 (with parent superset Π) + ... + χ^2_K (with parent superset Π) - χ^2_{pooled} (with parent superset Π)

  This option is not suitable for parent selection, but is theoretically proved for approaching a null chi-square distribution.

- -Y {a string. Default: `BY_TOT`}

  This parameter specifies how to compare different candidate parent sets of a child to select the best one covering all conditions:

  `BY_TOT`: by total strength of interactions across conditions.

  `BY_EACH_COND`: by individual interaction strength under each condition.

  `BY_INTX_TYPE`: by interaction type. Priority is given to differential interactions.

  `BY_HOM`: by homogeneity p-values.

  `BY_HET:`by heterogeneity p-values.

- -Z

  When specified, self-cycle edges are allowed; otherwise, self-cycle edges are not allowed. No values are needed after -Z.

The most influential parameter options include -k, -A, -K, -J and -Y.

Example

The input includes two trajectory collection files `1.trj` and `2.trj` shown in Fig. 1, corresponding to two experimental conditions. A TCF file can contain multiple trajectories, representing either steady-state sample or dynamic time-series.

We can compare the two input trajectory collection files to identify conserved and differential interactions using the following command line:

```
> CPX2 -M comparison -k 2 -1 1.trj -2 2.trj -G summary.txt -D summary.dot
-w INDEP_SUP -Y BY_EACH_COND
```

This command generates three types of output, including `summary.txt`, `summary.dot` and the screen dump. Figure 2, produced from `summary.dot`, visualizes the networks under both conditions, including differential and conserved interactions. Table 1 shows the content of `summary.txt` and explains the rows and columns.

Table 1: The content of `summary.txt`. Each row gives the statistics of the comparative analysis of each child node.

Type	Same Parents	p_d	Parents1	Parents2	Common Parents	p_c	p_t	p_z (child)	p_z (parents)	Child Name
C	-	0.172037	2	2	2	3.81241e-16	5.30611e-15	0.998994	1	u
C	-	0.670541	1	3	1,3	5.3449e-08	7.90397e-06	1	0	v
D	-	1.82687e-17	2	2	2	5.87642e-05	5.34337e-20	0.990119	1	w
C	-	0.619336	1	3	1,3	2.27864e-15	4.16875e-12	0.998882	0	y

In the 11 columns, the first column (Type) is the comparative interaction type: D denotes differential while C denotes conserved. The 2nd column (SameParents) is reserved for other usage. The 3rd column (p_d) is the *p*-value for interaction heterogeneity. The 4th column (Parents1) is parent set in condition 1 and 5th column (Parents2) is parent set in condition 2. The 6th column (CommonParents) is shared parent across conditions, as we have used $-w\ \ INDEP\_SUP$, this denotes the parents supersets. The 7th-10th columns denotes statistical significance for interaction homogeneity (p_c), total interaction strength (p_t), child (p_z(child)) and parent (p_z(parents)) working zone change respectively. The last column (ChildName) is the child node name.

The screen dump, shown as follows, contains the best contingency tables for each child node, in addition to comparative chi-square statistics.

screen dump

```
N1-0,N1-1,N1-2,N1-3,N1-4,N2-0,N2-1,N2-2,N2-3,N2-4,
N3-0,N3-1,N3-2,N3-3,N3-4,N4-0,N4-1,N4-2,N4-3,N4-4,

id name type num.parents parents offsets p.value chisq df
1 u i 1 2, -1, 5.56766e-05 24.781 4
2 v i 1 1, -1, 5.56766e-05 24.781 4
3 w i 1 2, -1, 2.90086e-12 60 4
4 y i 1 1, -1, 2.90086e-12 60 4
Overall p-value of the reconstructed generalized logical network = 1.02962e-27
```

Data from the first condition

"1.trj"

```
TRAJECTORY_VER2
1 4 0
3         3         3         3
u         v         w         y
31
1         0         1         0
1         0         1         0
1         0         1         0
2         1         1         0
2         1         2         1
2         1         2         1
2         1         2         1
2         1         2         1
0         2         2         1
0         2         0         2
0         2         0         2
0         2         0         2
1         0         0         2
1         0         1         0
1         0         1         0
2         1         1         0
2         1         2         1
2         1         2         1
0         2         2         1
0         2         0         2
0         2         0         2
0         2         0         2
1         0         0         2
1         0         1         0
1         0         1         0
2         1         1         0
2         1         2         1
2         1         2         1
0         2         2         1
0         2         0         2
0         2         0         2
```

Data from the second condition

"2.trj"

```
TRAJECTORY_VER2
1 4 0
3         3         3         3
u         v         w         y
31
2         0         2         1
1         0         2         0
1         0         2         0
1         0         2         0
1         0         2         0
1         0         2         0
1         0         2         0
1         0         2         0
1         0         2         0
1         1         0         0
2         1         0         1
2         1         0         1
2         1         0         1
2         1         0         1
2         1         0         1
2         1         0         1
2         1         0         1
2         1         0         1
2         1         0         1
2         1         0         1
2         2         1         1
0         2         1         2
0         2         1         2
0         2         1         2
0         2         1         2
0         2         1         2
0         2         1         2
0         2         1         2
0         2         1         2
0         2         1         2
```

Figure 1: The input trajectory collections files to the comparative interaction analysis.

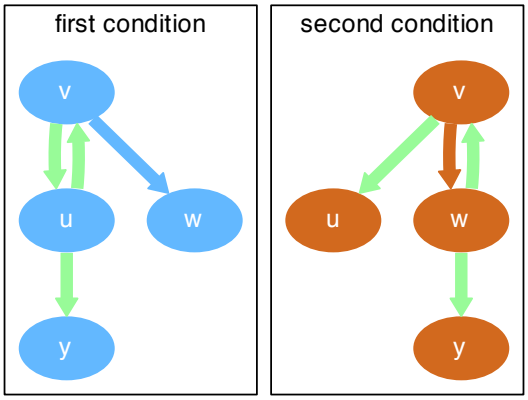

Figure 2: The two networks generated as an output of the comparative interaction analysis. Blue nodes are from the network under condition 1 and chocolate ones from the network under condition 2. Green edges indicate conserved interaction across the two conditions. Blue edges are interactions in network 1, differential from network 2. Chocolate edges are interactions in network 2, differential from network 1. Both networks include the edge $y \rightarrow v$, but they are differential in the interaction strength and thus not conserved. The width of an edge is proportional to the statistical significance of the differential/conserved edge. The graphs are rendered from `summary.dot` using Graphviz.

```
Adjusted p-values:
node.id        parents        adjusted.p.value
1         2,          0.000668119
2         1,          0.000668119
3         2,          3.48104e-11
4         1,          3.48104e-11
Significant transition tables:
Child=u[1], with 1 parents = {v[2]-1} (name[ID]delay)
0         6         3
3         0         8
8         2         0

Child=v[2], with 1 parents = {u[1]-1} (name[ID]delay)
2         0         8
6         3         0
0         8         3

Child=w[3], with 1 parents = {v[2]-1} (name[ID]delay)
0         9         0
0         0         11
10         0         0

Child=y[4], with 1 parents = {u[1]-1} (name[ID]delay)
0         0         10
9         0         0
```

```
0        11        0

Significant generalized truth tables:
1(u): 1, 2, 0,
2(v): 2, 0, 1,
3(w): 1, 2, 0,
4(y): 2, 0, 1,
N1-0,N1-1,N1-2,N1-3,N1-4,N2-0,N2-1,N2-2,N2-3,N2-4,
N3-0,N3-1,N3-2,N3-3,N3-4,N4-0,N4-1,N4-2,N4-3,N4-4,

id name type num.parents parents offsets p.value chisq df
1 u i 1 2, -1, 2.90086e-12 60 4
2 v i 1 3, -1, 5.09472e-10 49.2837 4
3 w i 1 2, -1, 5.09472e-10 49.2837 4
4 y i 1 3, -1, 2.90086e-12 60 4
Overall p-value of the reconstructed generalized logical network = 1.36331e-37

Adjusted p-values:
node.id        parents        adjusted.p.value
1        2,         3.48104e-11
2        3,         6.11366e-09
3        2,         6.11366e-09
4        3,         3.48104e-11
Significant transition tables:
Child=u[1], with 1 parents = {v[2]-1} (name[ID]delay)
0        9        0
0        0        11
10        0        0

Child=v[2], with 1 parents = {w[3]-1} (name[ID]delay)
0        10        1
0        0        10
8        1        0

Child=w[3], with 1 parents = {v[2]-1} (name[ID]delay)
1        0        8
10        1        0
0        10        0

Child=y[4], with 1 parents = {w[3]-1} (name[ID]delay)
0        11        0
0        0        10
9        0        0

Significant generalized truth tables:
1(u): 1, 2, 0,
2(v): 1, 2, 0,
```

```
3(w): 2, 0, 1,
4(y): 1, 2, 0,


************************************************************
Significant comparative interactions identified:
_____________________________________________________________________________


_____________________________________________________________
ABSOLUTE DIFFERENTIAL, with significant homogeneous component:
3        w
chi2d=84.6194        vd=4        pd=1.82687e-17        (heterogeneity)
chi2c=24.6643        vc=4        pc=5.87642e-05        (homogeneity)
chi2t=109.284        vt=8        pt=5.34337e-20        (total strength)
chi2z=0.295238        vz=4        pz=0.990119        (child working zone change)
chi2z=0.00000        vz=4        pz=1.00000        (parent working zone change)
... Contingency table under condition 1:
Child=w[3], with 1 parents = {v[2]-1} (name[ID]delay)
0        9        0
0        0        11
10        0        0

... Contingency table under condition 2:
Child=w[3], with 1 parents = {v[2]-1} (name[ID]delay)
1        0        8
10        1        0
0        10        0

... Pooled contingency table of all conditions:
Child=w[3], with 1 parents = {v[2]-1} (name[ID]delay)
1        9        8
10        1        11
10        10        0

~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~


__________
CONSERVED:
1        u
chi2d=6.38721        vd=4        pd=0.172037        (heterogeneity)
chi2c=78.3938        vc=4        pc=3.81241e-16        (homogeneity)
chi2t=84.7810        vt=8        pt=5.30611e-15        (total strength)
chi2z=0.0910973        vz=4        pz=0.998994        (child working zone change)
chi2z=0.00000        vz=4        pz=1.00000        (parent working zone change)
... Contingency table under condition 1:
Child=u[1], with 1 parents = {v[2]-1} (name[ID]delay)
0        6        3
```

```
3        0        8
8        2        0

... Contingency table under condition 2:
Child=u[1], with 1 parents = {v[2]-1} (name[ID]delay)
0        9        0
0        0        11
10        0        0

... Pooled contingency table of all conditions:
Child=u[1], with 1 parents = {v[2]-1} (name[ID]delay)
0        15        3
3        0        19
18        2        0

~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~
2        v
chi2d=13.0303        vd=16        pd=0.670541        (heterogeneity)
chi2c=65.8086        vc=16        pc=5.34490e-08        (homogeneity)
chi2t=78.8389        vt=32        pt=7.90397e-06        (total strength)
chi2z=0.00000        vz=4        pz=1.00000        (child working zone change)
chi2z=0.00000        vz=0        pz=0.00000        (parent working zone change)
... Contingency table under condition 1:
Child=v[2], with 1 parents = {u[1]-1} (name[ID]delay)
2        0        8
6        3        0
0        8        3

... Contingency table under condition 2:
Child=v[2], with 1 parents = {w[3]-1} (name[ID]delay)
0        10        1
0        0        10
8        1        0

... Pooled contingency table of all conditions:
Child=v[2], with 2 parents = {u[1]-1;w[3]-1} (name[ID]delay)
2        0        5
0        0        9
0        0        3
2        1        0
4        3        0
7        1        0
0        9        1
0        3        1
1        5        3

~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~
```

```
4        y
chi2d=13.7231        vd=16        pd=0.619336        (heterogeneity)
chi2c=106.277        vc=16        pc=2.27864e-15        (homogeneity)
chi2t=120.000        vt=32        pt=4.16875e-12        (total strength)
chi2z=0.0961098        vz=4        pz=0.998882        (child working zone change)
chi2z=0.00000        vz=0        pz=0.00000        (parent working zone change)
... Contingency table under condition 1:
Child=y[4], with 1 parents = {u[1]-1} (name[ID]delay)
0        0        10
9        0        0
0        11        0

... Contingency table under condition 2:
Child=y[4], with 1 parents = {w[3]-1} (name[ID]delay)
0        11        0
0        0        10
9        0        0

... Pooled contingency table of all conditions:
Child=y[4], with 2 parents = {u[1]-1;w[3]-1} (name[ID]delay)
0        0        7
0        0        9
0        0        3
2        1        0
7        0        0
8        0        0
0        10        0
0        3        1
1        8        0

~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~


done
```