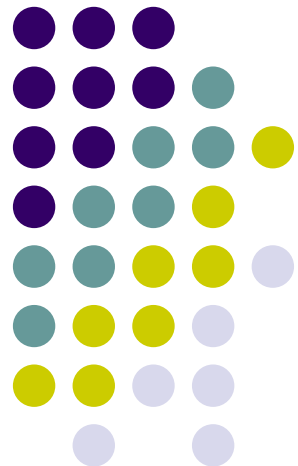
Knowledge Digraph Contribution (KDC) Analysis Software Tutorial

Richard Golden

School of Behavioral and Brain Sciences

University of Texas at Dallas

golden@utdallas.edu

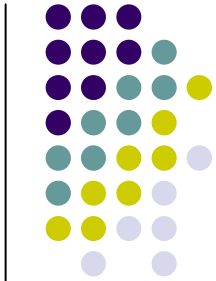


Revision Date: January 28, 2006

Supported in part by the NSF ITR Award Initiative through the Research on Learning and Education Program Award 0113369

Any opinions findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

What Problems is KDC Analysis Designed to Solve?



- Deciding which of several models of sequential order “best-fits” a particular sequence of response data items for a group of participants.

- Using the sequential order of response data items to identify which sequential regularities (i.e.,“digraphs”) are invariant (or non-invariant) across experimental conditions (e.g., an independent-groups design).

- Moreover, estimate digraph parameters for individual participants for the purpose of investigating individual differences.

What is KDC (Knowledge Digraph Contribution) Analysis?

- Assume the data for individual subjects can be represented as a categorical time-series. (e.g., Subject's recall of a story is given by the sequence: "2, 4, 1, 3, 1, 5" which means that first proposition #2 was mentioned, then proposition #4 was mentioned, then proposition #1 was mentioned, and so on…)

- Assume the theorist has certain preconceived ideas regarding the likelihood of particular patterns of sequential information. These can be highly constrained (e.g., "subjects will recall the ideas in their original order of presentation in the text" or "subjects will recall the superordinate ideas in the text in a particular order"). These possible patterns of sequential information are mathematically represented as *directed-graphs* (i.e., "*digraphs")*

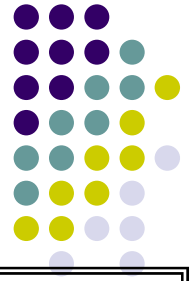
- Then KDC analysis can compute "weighting coefficients" analogous to "beta weights" in linear regression. Each *contribution weight* in KDC analysis indicates the degree a particular digraph is effective at "explaining" statistical regularities in the data generated by the individual subjects.

How does KDC analysis compare to standard methods of data analysis?

- Most standard methods of data analysis focus on *"what"* items are mentioned without explicitly taking into account the "order" in which the items are mentioned.

- The few existing sequential methods of data analysis tend to be *exploratory* in nature and focus on estimating evidence for "local sequential patterns" (e.g., the percentage of times that one item follows another) instead of seeking *confirmatory* evidence for "global sequential patterns" (e.g., the degree to which the order in which participants recalled a sequence of items is consistent with the order in which the items were originally presented).

Example *Categorical Data Analysis*
Application: Free Response Data Analysis



RAW DATA

(participant free response data):

"*A dragon kidnapped the three daughters. As they were being dragged off they called for help. Some knights rescued them at the end of the story."*

DICTIONARY

1. START-NODE
2. KIDNAP(DRAGON, DAUGHTERS)
3. TRANSFER(DRAGON, DAUGHTERS)
4. SCREAM(DAUGHTERS)
5. HEAR(KNIGHTS,SCREAM(DAUGHTERS))
6. RESCUE(KNIGHTS, DAUGHTERS)
7. REWARD(CZAR, KNIGHTS)
8. MARRIED(KNIGHTS, DAUGHTERS)
9. END-NODE

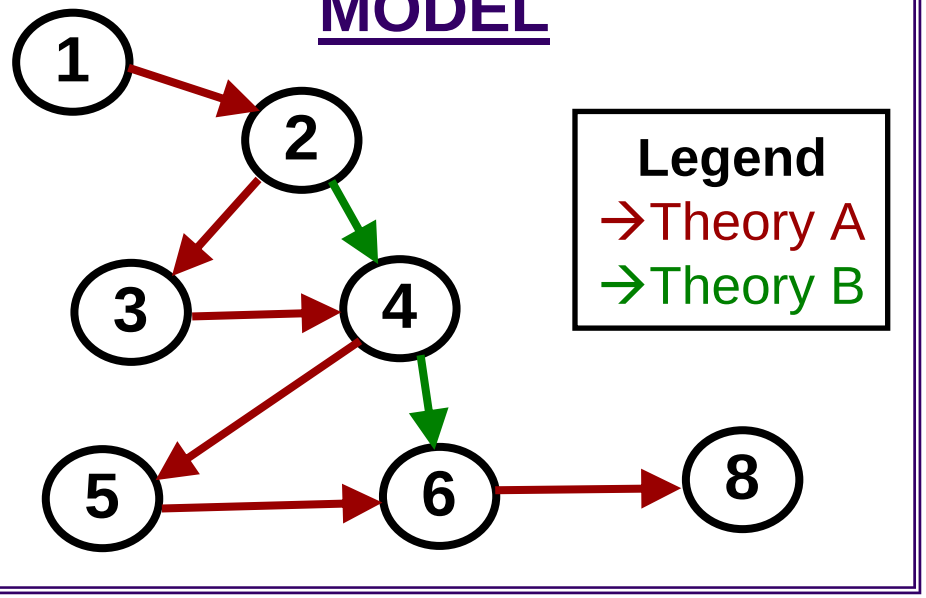
CODED DATA: 1, 2, 3, 4, 6, 9

Analyses of Categorical Time-Series Data

OBSERVED DATA

Participant 1: 1, 2, 3, 4, 6, 9
Participant 2: 1, 2, 4, 6, 9
Participant 3: 1, 2, 3, 4, 6, 9
Participant 4: 1, 3, 4, 6, 2, 9
Participant 5: 1, 4, 6, 2, 9

MODEL



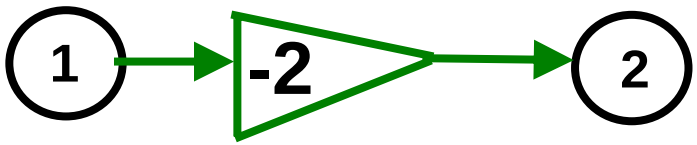
KEY QUESTIONS

- Does Theory A account for data as effectively as Theory B?
- How does predictive relevance of Theory A versus Theory B vary as a function of experimental manipulations?

Temporal Digraph Notation

Theory B predicts observation 2 immediately follows observation 1 in the observed data ("lag 1 link")

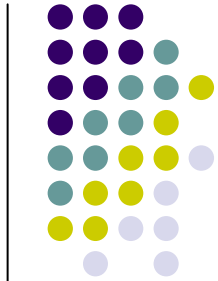
Theory A predicts observation 2 immediately follows observation 1 in the observed data ("lag 1 link")



Theory B predicts observation 2 follows some other observation X, and observation X follows observation 1 in the observed data ("lag 2 link")

Theory A predicts observation 2 follows some other observation X, and observation X follows another observation Y, and observation Y follows observation 1 in the observed data ("lag 3 link")

Format of the “Data” File



“test.data”

```
% Participant 1:
1 2 3 4 5 7
%
% Participant 2:
1 2 4 6 7
%
% Participant 3:
1 2 3 4 6 7
```

- Comment Lines have a % sign at beginning of line
- Each list of integers corresponds to a sequence of observations from a participant in the study
- Data files are created using a standard text editor and must have a filename with the suffix “.data”

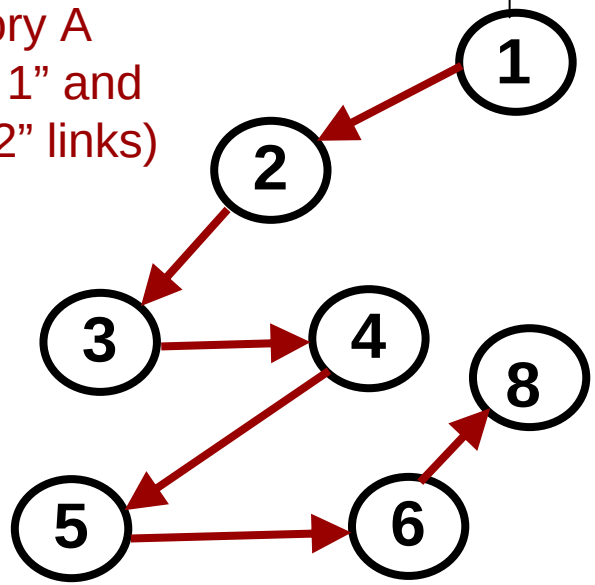
Format of the “Model” File

```
COMMENT: Example Model File for illustrating KDC Analysis
NODE 1: START-NODE
NODE 2: KIDNAP(DRAGON, DAUGHTERS)
NODE 3: TRANSFER(DRAGON, DAUGHTERS)
NODE 4: SCREAM(DAUGHTERS)
NODE 5: HEAR(KNIGHTS,SCREAM(DAUGHTERS))
NODE 6: RESCUE(KNIGHTS, DAUGHTERS)
NODE 7: REWARD(CZAR, KNIGHTS)
NODE 8: MARRIED(KNIGHTS, DAUGHTERS)
NODE 9: END-NODE
```

Node Interpretation List

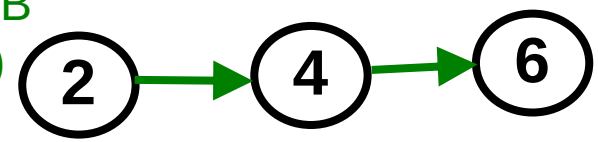
```
----------     Digraph # 1 (“THEORYA[1]")    ----------
LAG: 1
TO-NODE    FROM-NODE       WEIGHT-VALUE
8             8               0
2             1               1
3             2               1
4             3               1
5             4               1
6             5               1
8             6               1

----------     Digraph # 2 (“THEORYA[2]")    ----------
LAG: 2
TO-NODE    FROM-NODE       WEIGHT-VALUE
8             8               0
2             1               1
3             2               1
4             3               1
5             4               1
6             5               1
8             6               1
```

Theory A
(“lag 1” and
“lag 2” links)




```
----------     Digraph # 3 (“THEORYB[1]")    ----------
LAG: 1
TO-NODE    FROM-NODE       WEIGHT-VALUE
8             8               0
4             2               1
6             4               1
```

Theory B
(“lag 1”)

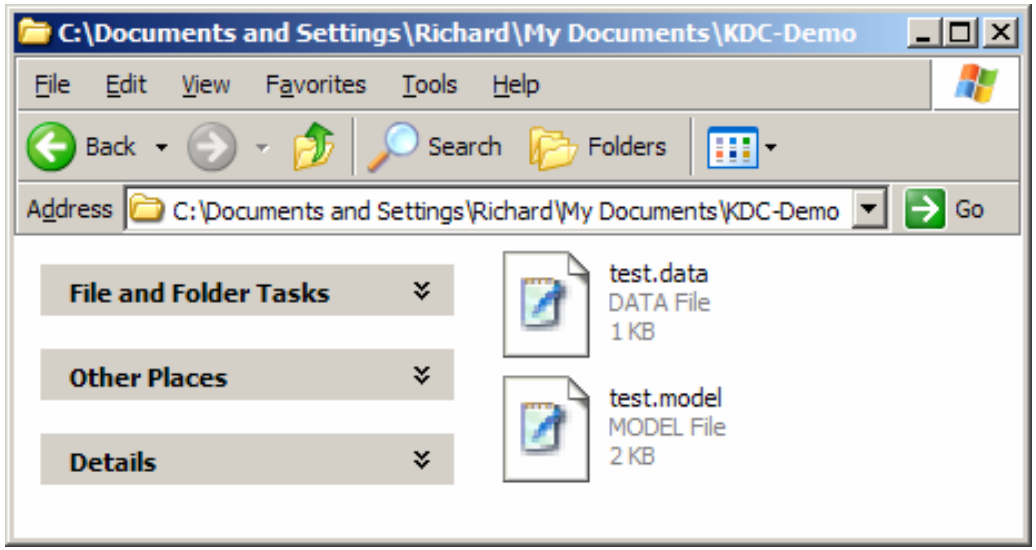



```
NODE 1 ATOMS: 1;
NODE 2 ATOMS: 2;
NODE 3 ATOMS: 3;
NODE 4 ATOMS: 4;
NODE 5 ATOMS: 5;
NODE 6 ATOMS: 6;
NODE 7 ATOMS: 7;
NODE 8 ATOMS: 8;
NODE 9 ATOMS: 9;
```

Node Translation List

“NODE 1 ATOMS: 35 23;”
indicates observations 35 and 23
should be relabeled as Node 1

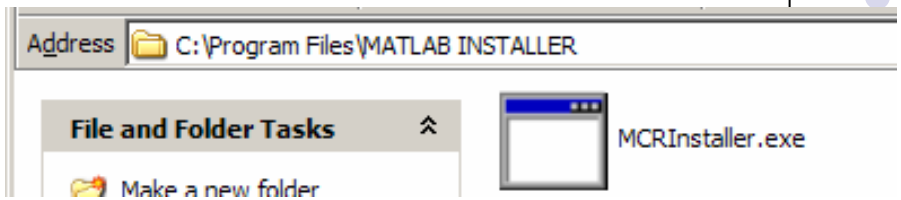
Setting up the Project Folder



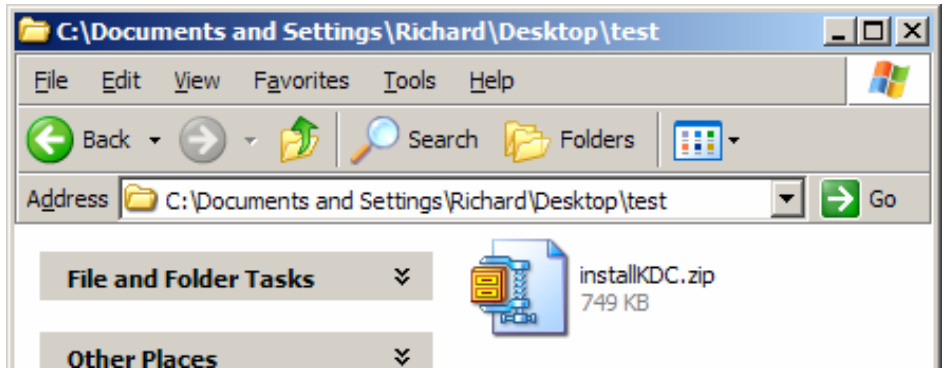
- Place the "`.data`" file and the "`.model`" file into a folder (e.g., `KDC-Demo`)
- All data analyses will be done within the project folder

Installing KDC on a Windows Operating System

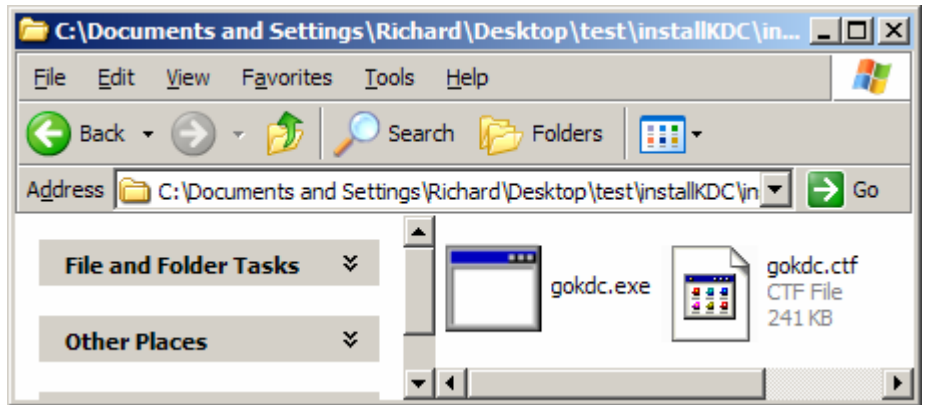
- **STEP 1:**
Move the file *MCRInstaller.exe* into the folder *MATLAB INSTALLER* in the *Program Files* folder. Then install the MATLAB Run-Time Component Library by clicking on: *MCRInstaller.exe* and following the directions. Note that this step may be omitted if the MATLAB Run-Time Library has been previously installed.



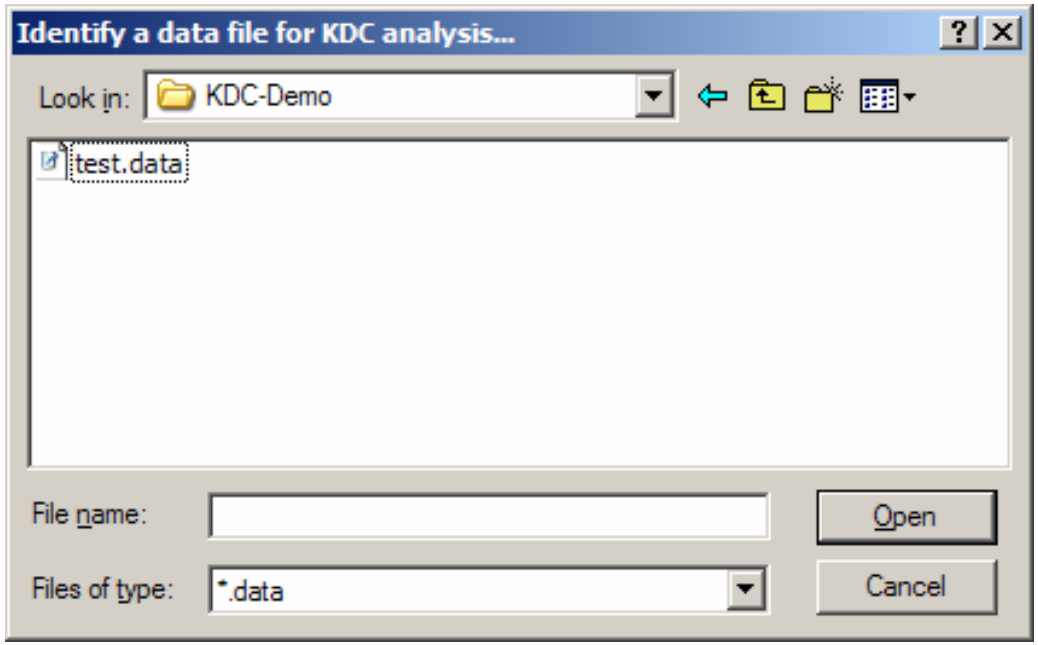
- **STEP 2:**
Unzip the file: *KDC.zip,* obtain the files *gokdc.exe* and *gokdc.ctf,* put both of these files in a folder called KDC located in your *Program Folder* with the *help folder. No other files or folders should be located in the folder KDC at this point in the installation process.*



- **STEP 3:**
Create a short-cut to *gokdc.exe* by highlighting *gokdc.exe* and right-clicking "`Create Shortcut`". You can copy and paste this short cut anywhere you wish to invoke the software. Alternatively you can click on "`Pin to Start Menu`" to access *gokdc* from the start menu of your system



Selecting the Project Folder…

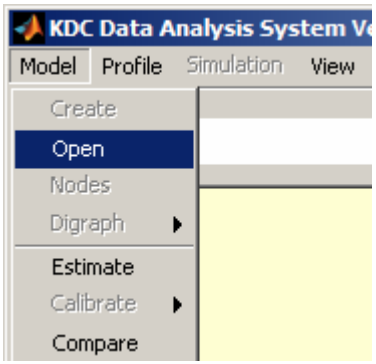


- The software will ask you to select a data file (must have ".data" suffix and be a KDC data file) for KDC data analysis for the purposes of identifying the project folder.
- All subsequent data analyses in this session must take place in the project folder which contains the data file have selected.
- If you wish to analyze data from another analysis, then you will need to abort the software and restart KDC analysis.

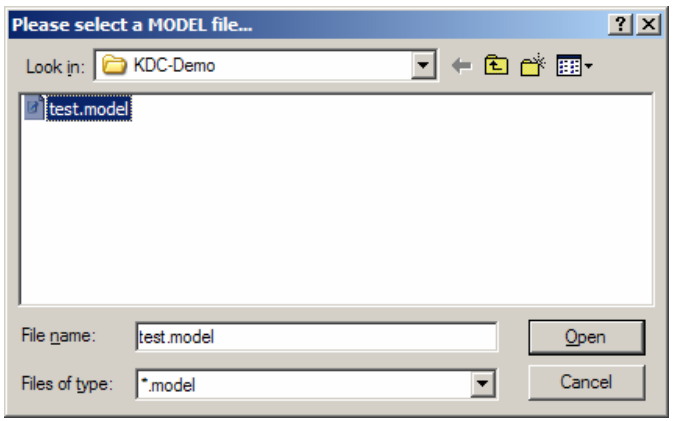
Loading a Model…

- The following sequence of steps is used to load a model into KDC's "workspace"…

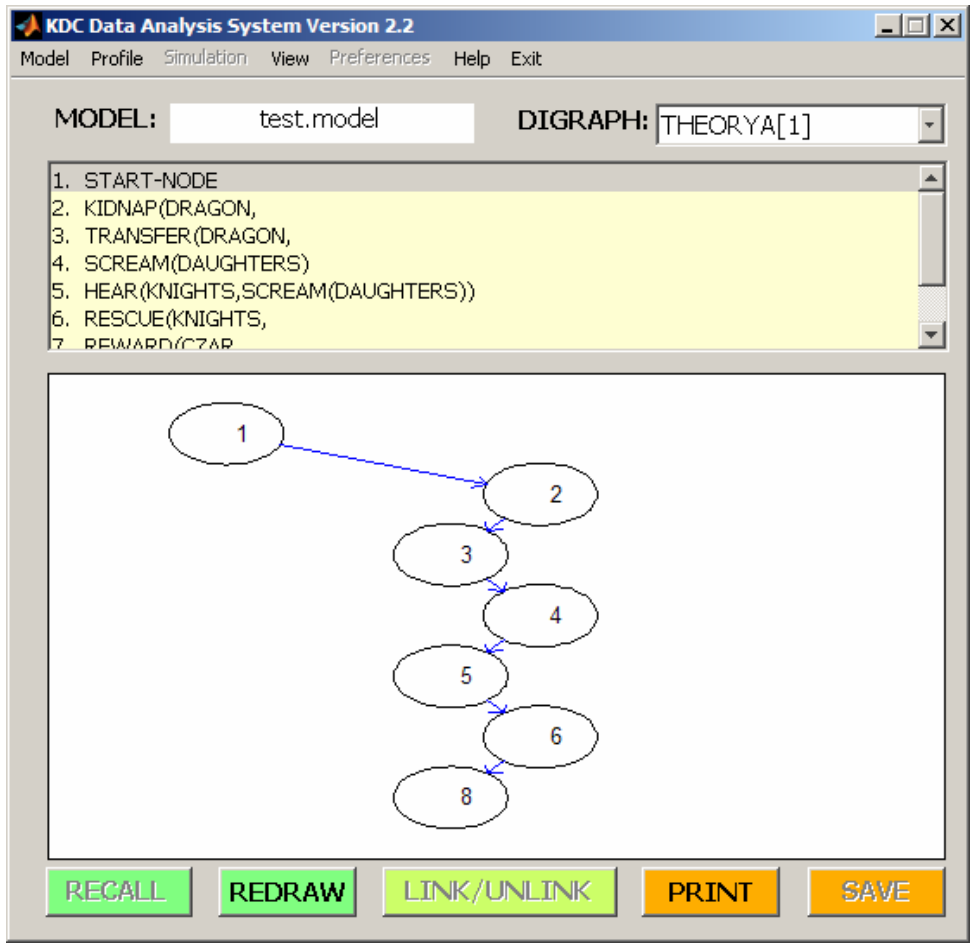
Step 1: Select `Model→Open`



Step 2: Identify "model" file

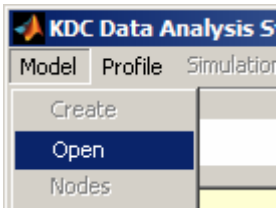


Step 3: Model is Loaded into Workspace

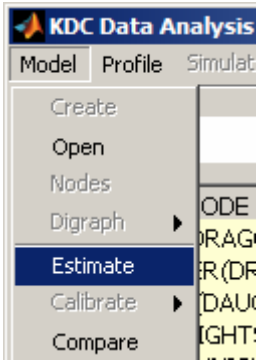


How to Estimate Model Parameters

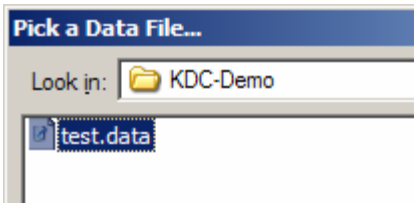
Step 1: Load a Model into the Workspace



Step 2: Initiate Estimation Procedure



Step 3: Select Data File



Step 4: View Results Display

```
Model: "test.model"
Data Sample (Nr of Propositions = 17): "test.data"

** OVERALL STATISTICAL INFERENCE RELIABILITY (details below): Acceptable

********************************* RESULTS ******************************
DIGRAPH              WEIGHT     STD.ERROR          Z         Pr(Type 1)
THEORYA               2.4509        0.2206      11.1120        0.0000
THEORYB               3.3741        0.7721       4.3702        0.0000
```

Sample Size (N) = 17

"Theory A weight" = 2.4509±0.2206

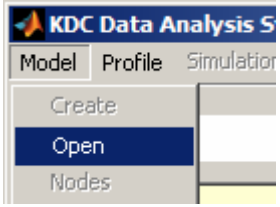
"Theory B weight" = 3.3741±0.7721

Note: Both Contribution Weights are significantly different from zero at 0.05 since both p-values are less than 0.05

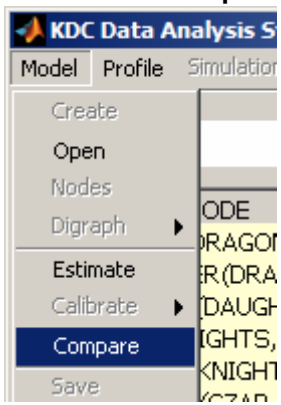
Indicates that Assumptions of Statistical Analysis Appear Valid

How to Compare Models

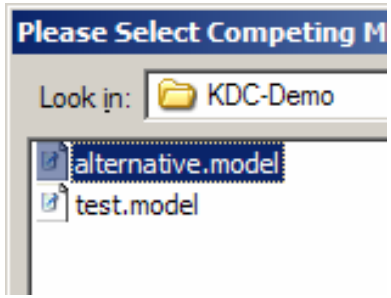
Step 1: Load First Model into Workspace



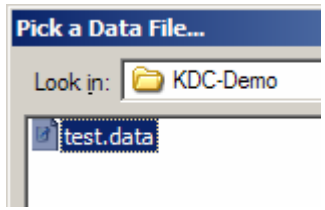
Step 2: Initiate Model Comparison



Step 3: Select Competing Model



Step 4: Select Data File



Assumptions for statistical inferences appear to NOT be valid. That is, quality of model/data is NOT sufficient to support reliable inferences.

Ideally, autocorrelation SHOULD be different from its critical value for the analysis to be valid.

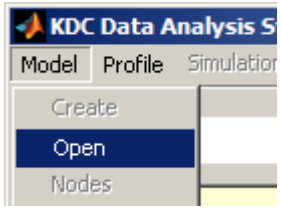
Step 5: View Results Display

Model 1: "test.model"
Model 2: "alternative.model"
Data Sample (n = 17): "test.data"
** OVERALL STATISTICAL INFERENCE QUALITY (details below): Poor
Autocorrelation (R= 0.11288) different from critical value of 0.16667.
WARNING! R matrix multicollinearity level (Condition No. = 17651839198056.19900) too large!
********** DISCREPANCY RISK MODEL SELECTION TEST RESULTS **********
Model 1 Fit = 1.21407, Model 2 Fit = 1.52728,
Model 1 and Model 2 provided equally effective fits to the data.
No significant difference between likelihood per observation for Model 1 ("test.model") and likelihood per observation for Model 2 ("alternative.model") (Discrepancy Variance = 1.3115
p = 0.796222 which is greater than 0.050000 significance level.

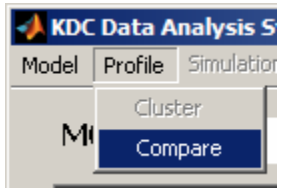
Note that even though Model 1 appears to fit the data more effectively than Model 2, there is not evidence this is a reliable difference.

How to Compare Independent Data Sets with respect to a Model (e.g., between-subjects experimental design)

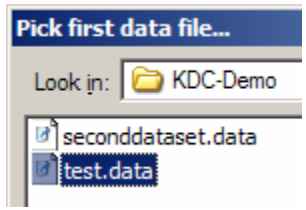
Step 1: Load First Model into Workspace



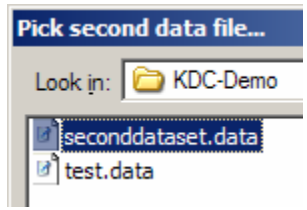
Step 2: Initiate Compare Profile



Step 3: Select First Data File



Step 4: Select Second Data File



Step 5: View Results

Model: "test.model"
Data Sample 1 (n = 17): "test.data"
Data Sample 2 (n = 13): "seconddataset.data"

** OVERALL STATISTICAL INFERENCE QUALITY (details below): Acceptable
********** RESULTS (By Propositions) **********
Contribution weights estimated using Data Sample 1 ("test.data")
differed from weights estimated using Data Sample 2
("seconddataset.data"), CHI-SQUARE(2) = 10.5, p = 0.0052, p < 0.0500.

DIGRAPH	WEIGHT ("Sample 1")	WEIGHT ("Sample 2")	STD.ERROR	Z	P(Type 1)
THEORYA	2.45094	0.62807	0.56226	3.24203	0.00119
THEORYB	3.37412	3.37567	1.48987	-0.00104	0.99917

Statistical assumptions are valid.

Planned comparison shows contribution weight patterns between data sets is significantly different at 0.05 level.

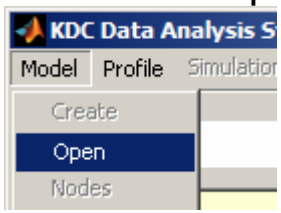
Theory A Contribution Weight for Data Sample 1 (“test data”) significantly larger than Theory A Contribution Weight for Data Sample 2 (“seconddataset.data”) (p=0.0019)

Theory B Contribution Weight for Data Sample 1 (“test.data”) is not significantly different from Theory B Contribution Weight for Data Sample 2 (“seconddataset.data”)

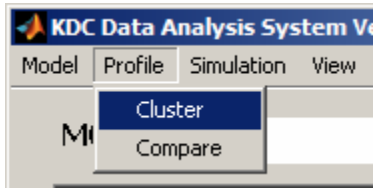
CLUSTER DATA ANALYSIS (useful for examining individual differences)

Step 1: Select `Help→User Level→Expert` from Main Menu Bar

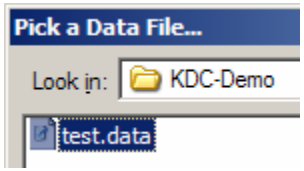
Step 2: Load a Model into the Workspace



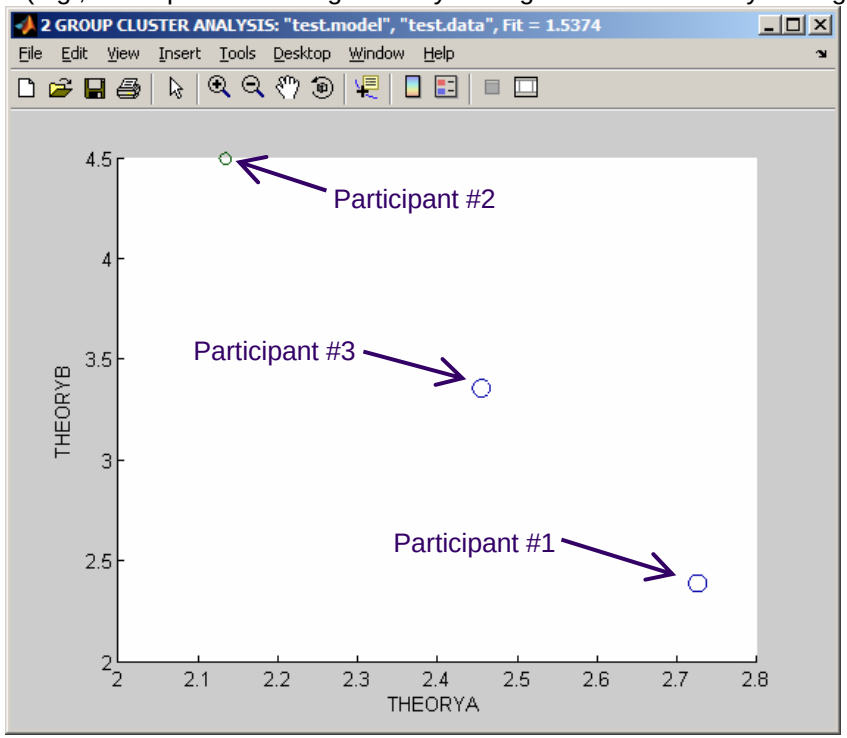
Step 3: Select `Profile→Cluster` from Main Menu Bar



Step 4: Select Data File



Step 5: View Graphical Results
(e.g., Participant 1 has large Theory A weight but small Theory B weight)



Step 6: View Text Results

```
** CLUSTER ANALYSIS RESULTS:
Clustering Fit =
GROUP 1:
     2
GROUP 2:
     1  3
```

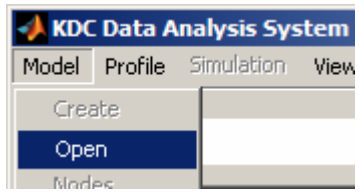
Cluster analysis of contribution weight pattern for each individual participant resulted in grouping Participants 1 and 3 into “Group 2”.

Participant 2 was assigned “Group 1”

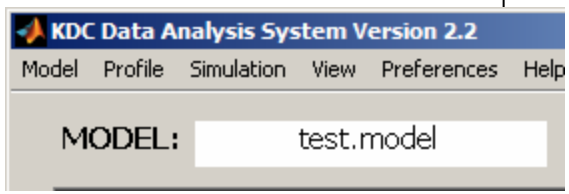
How to Generate Calibrated Digraphs (i.e., digraphs with weighted links)

Step 1: Select `Help→User Level→Expert` from Main Menu Bar

Step 2: Load a Model into the Workspace

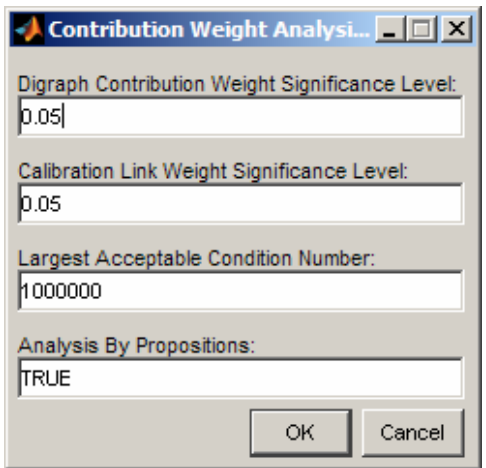


(requesting a loading of model file)

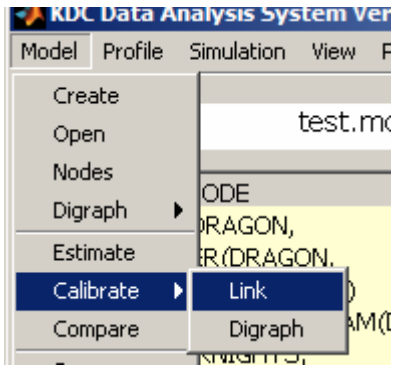


(after loading file "`test.model`")

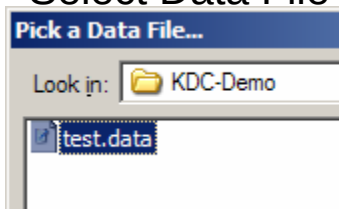
Step 3: Select Significance Level for Pruning Links using `Preferences→Analyis→ Weights.` If the type 1 Error probability that a link Is significantly different from Zero is greater than the Calibration Link Weight Significance Level, then that link is deleted.



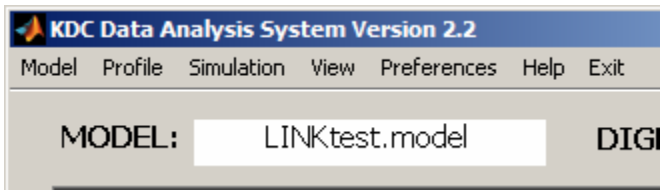
Step 4: Select `Model→ Calibrate→ Link`



Step 5: Select Data File



Step 6: "Calibrated" model is created with suffix ".model" and with the prefix "LINK". This model is a new text file in the project folder and contains "weighted" links where the weights are optimally chosen using maximum likelihood estimation to "best-fit" the data in the data file selected in Step 4.



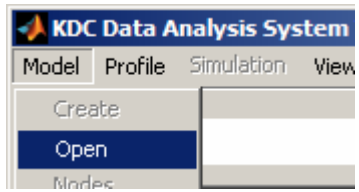
(the newly create file "`LINKtest.model`" with weighted links!)

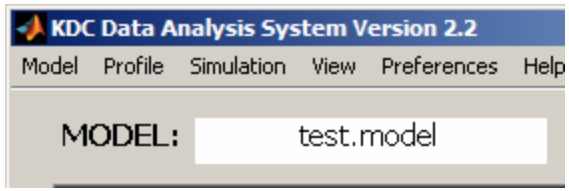
How to Generate Simulated Data

- Simulated human subject data may be generated using either parametric or non-parametric bootstrap methodologies. Use `Preferences` to select bootstrap methodology choice.
- Simulated human subject data files may be compared with actual human subject data to evaluate model quality.
- By estimating the contribution weight(s) for several simulated human subject data sets, the standard error(s) of that contribution weight across data sets may be compared with the analytical formulas in software to evaluate the large sample approximations.

Step 1: Select `Help→User Level→Expert` from Main Menu Bar

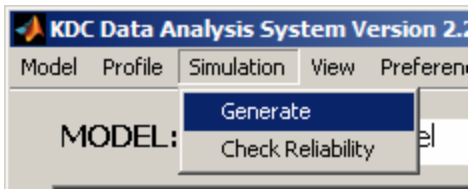
Step 2: Load a Model into the Workspace



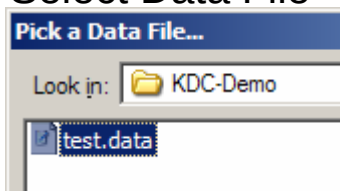


(after loading file "`test.model`")

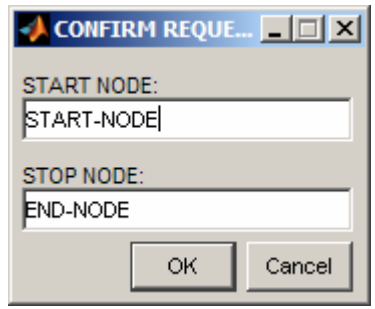
Step 3: Select `Simulation→Generate`



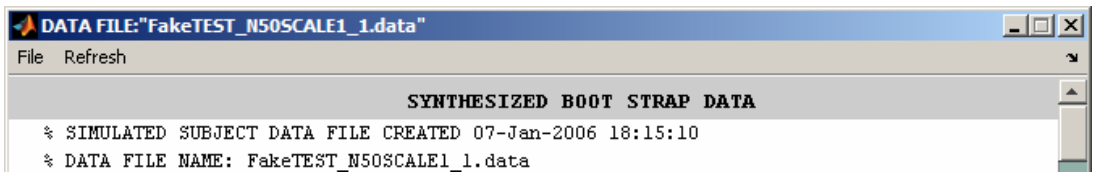
Step 4: Select Data File



Step 5: A "START NODE" (first node mentioned by model) and a "STOP NODE" (last node mentioned by model) from the NODE INTERPRETATION LIST in the model file must be defined.



Step 6: A data file is created which contains "simulated data" generated by the model. This has the same format as a human subject data file.

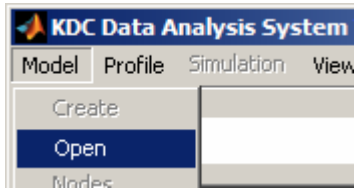


Generating Operating Curves for Evaluation of Statistical Test Performance Using Simulated Data (Setting Up Simulation Runs)

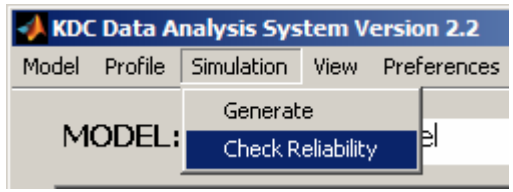
Step 1: Select `Help→User Level→Expert` from Main Menu Bar

Note: Fine-grained adjustments to the simulation runs may be accessed via the `Preferences menu`.

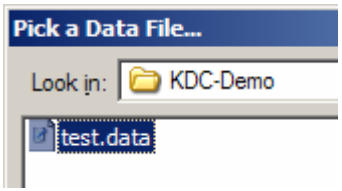
Step 2: Load a Model into the Workspace



Step 3: Select `Simulation→Check Reliability`

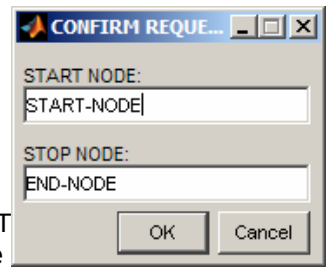


Step 4: Select Data File



Step 5:

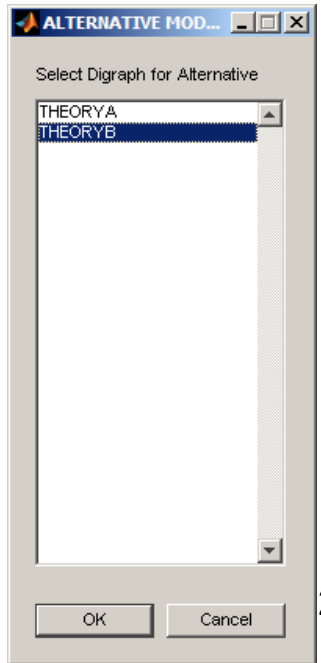
A "START NODE" (first node mentioned by model) and a "STOP NODE" (last node mentioned by model) from the NODE INTERPRETATION LIST in the model file must be defined.



Step 6: Select Digraph for Alternative Model

Compare Model ROC: The digraph which is selected is defined as the model which is "competing" with the null model comprised of the remaining digraphs.

Profile Compare ROC: The model with the remaining digraphs is the null model. One data set is generated from that model and one data set is generated from the model which consists of only the single selected digraph.

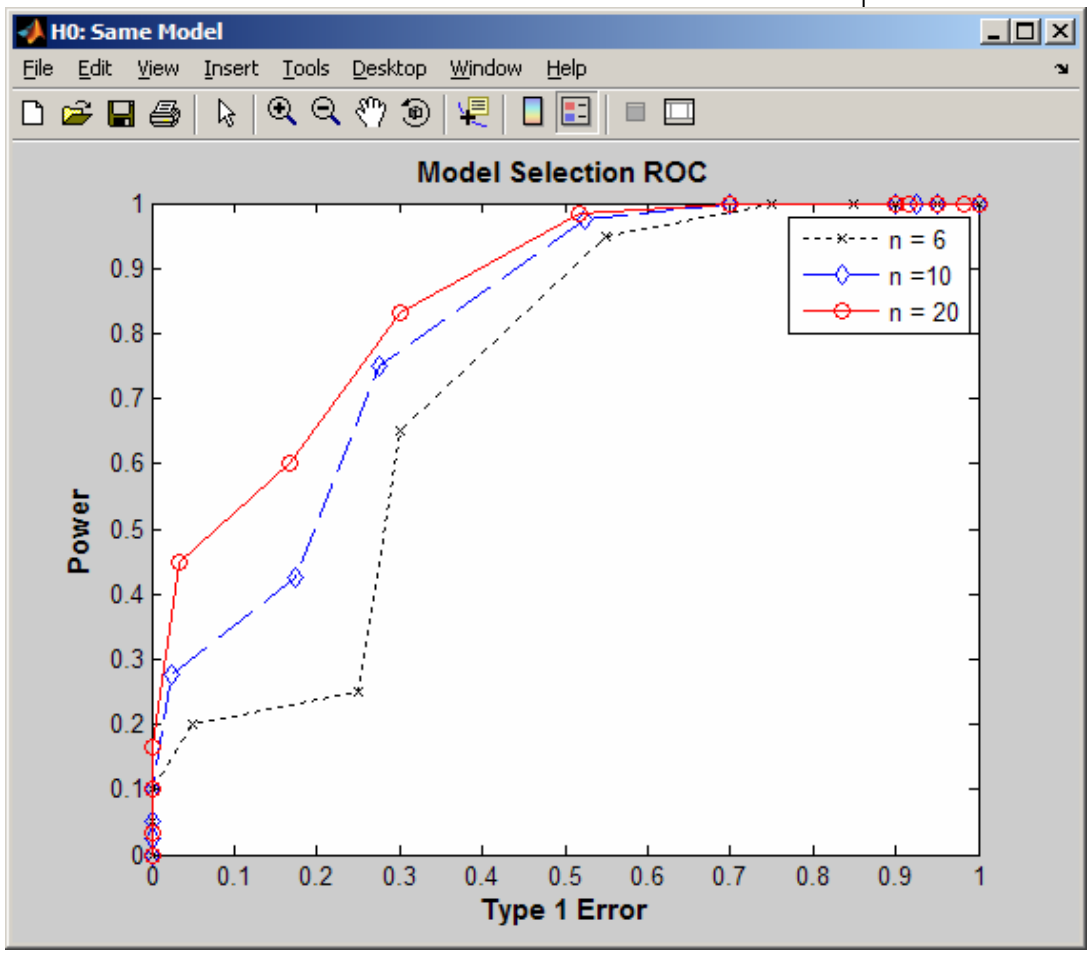


Evaluating Statistical Test Performance Using Simulated Data (Simulation Run Outputs)

Operating characteristic curves for both `Model→Compare` and `Profile→Compare` are generated.

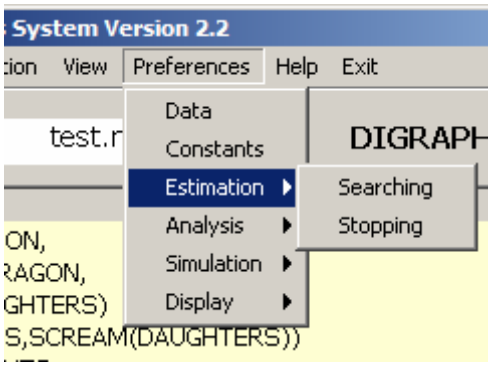
For each significance level, the Type 1 error probability and power may be estimated since the simulated data is generated from a known source.

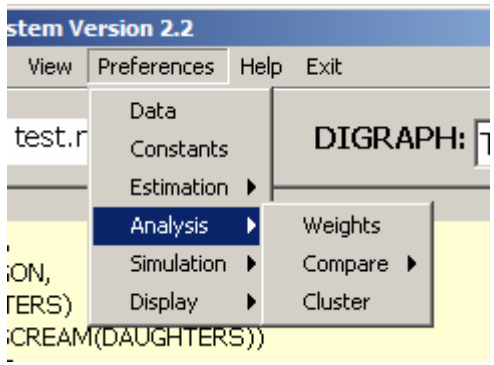
Ideally, the Power should be close to unity when the Type 1 error is equal to a typical significance level (e.g., α=0.01 or α=0.05)

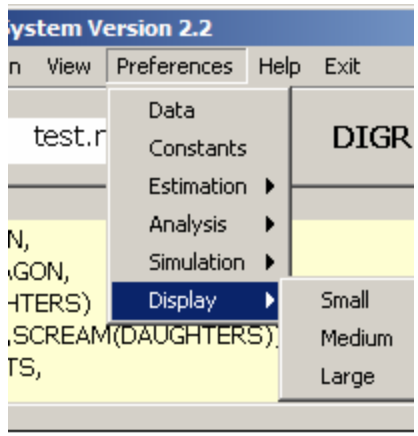


Operating Characteristic Curves for `Model→Compare` Statistical Test. The sample size in this case should probably be increased to reduce the Type 2 error rate.

Preferences Menu







- Step 1: `Help→User Level→Expert`
- Step 2: This gives you access to the Preferences menu which provides additional fine-grained control of the functionality in the KDC software package

References

- Golden, R. M. (in preparation). *Knowledge Digraph Contribution Analysis.*

- Jaynes, C. and Golden, R. M. (2003). Statistical Detection of Local Coherence Relations in Narrative Recall and Summarization Data. In R. Alterman and D. Kirsch (Eds.). *Proceedings of the 25th Annual Conference of the Cognitive Science Society,* Boston, MA: Cognitive Science Society, 3-8.

- Golden, R. M. (1998). Knowledge digraph contribution analysis of protocol data. *Discourse Processes, 25,* 179-210.