1. Import Data

Left click “data” square button, create a “data” icon on canvas.

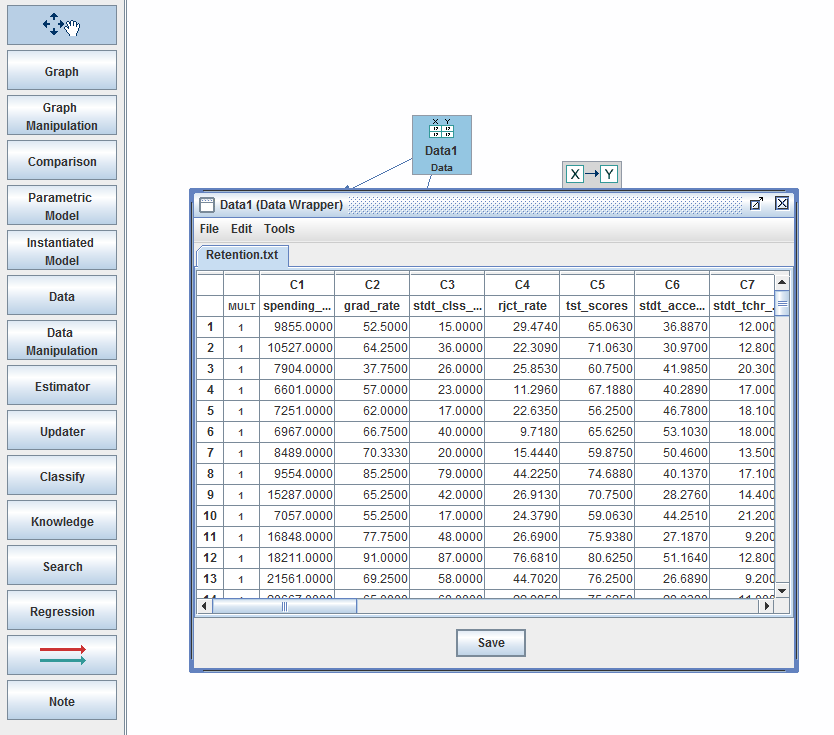
Double left click the “data” icon on canvas, choose “data wrap” option, click ”OK” to confirm.

In the prompted data table window, click “file”, then click “load data”.

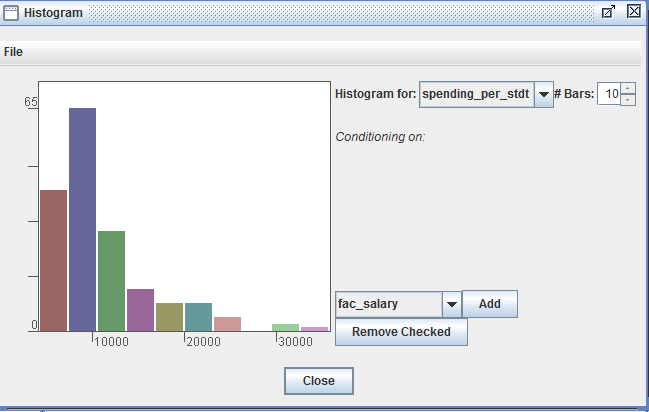
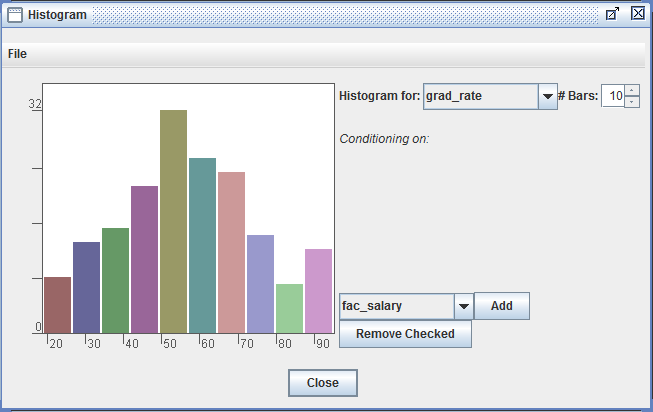
In the prompted file explorer, find saved data “retention.txt” and choose it.

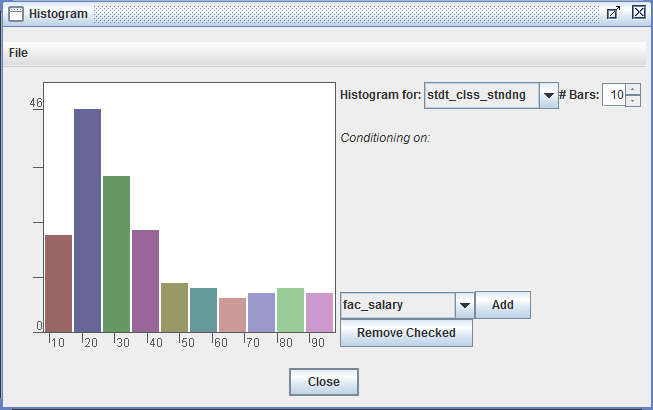
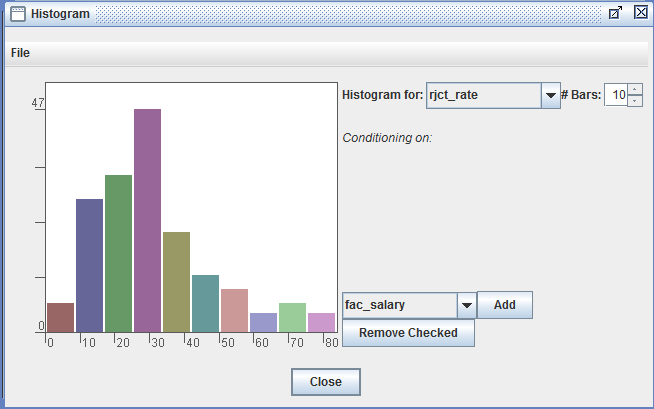
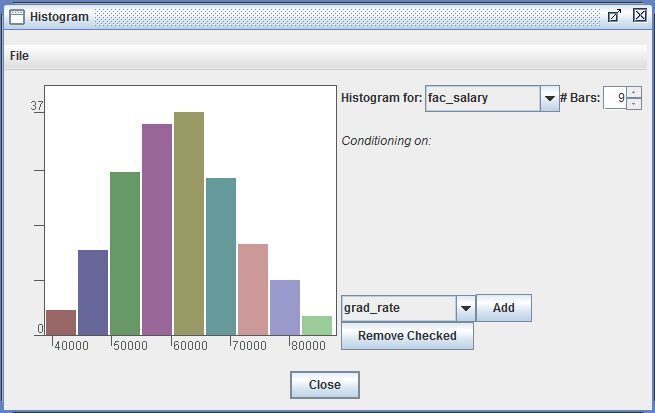
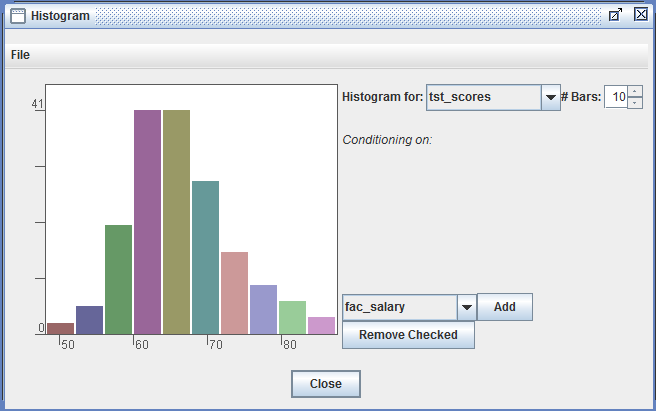
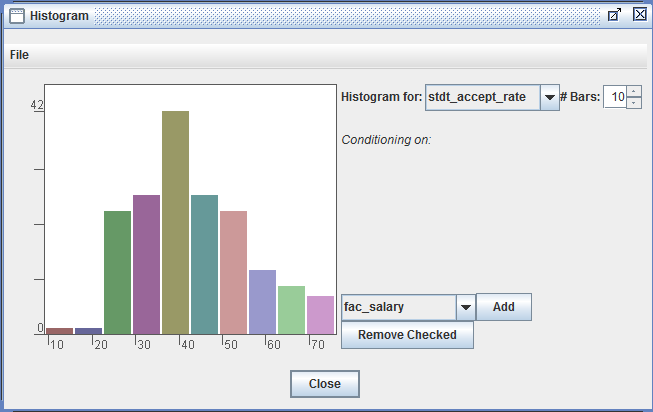
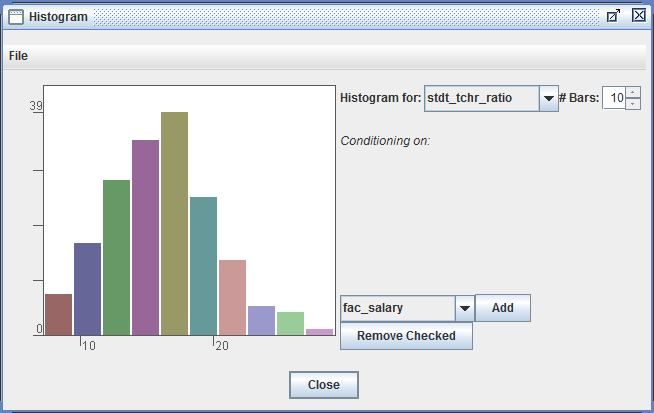
In the prompted import option window, preview the data, and choose “load”, then “save”.

Data imported.



1. Data Overview &Normality Test

In data wrapper window, click ”Tools” to get options: Calculator, Histogram, Scatter plots, Q-Q plots, Normality tests, and Descriptive statistics. Descriptive statistics demonstrates no missing values.



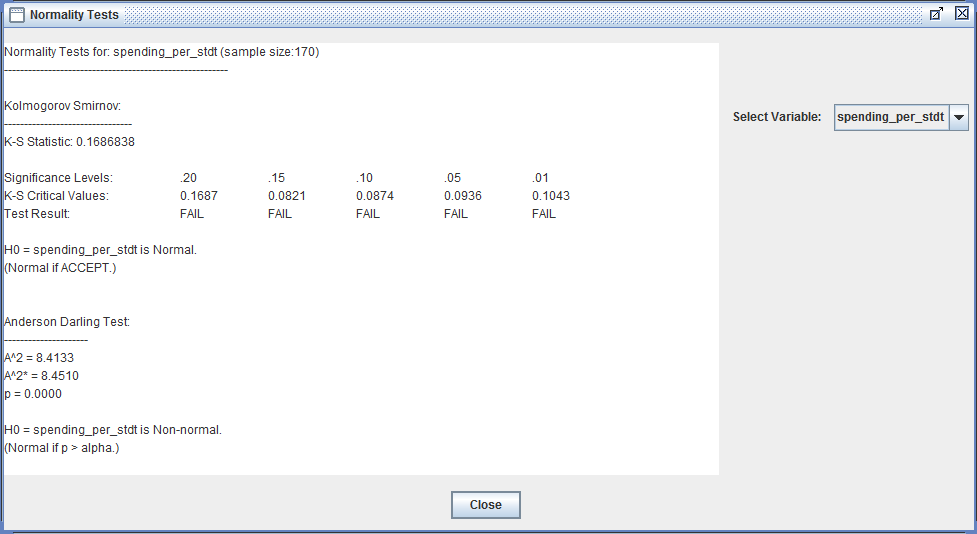
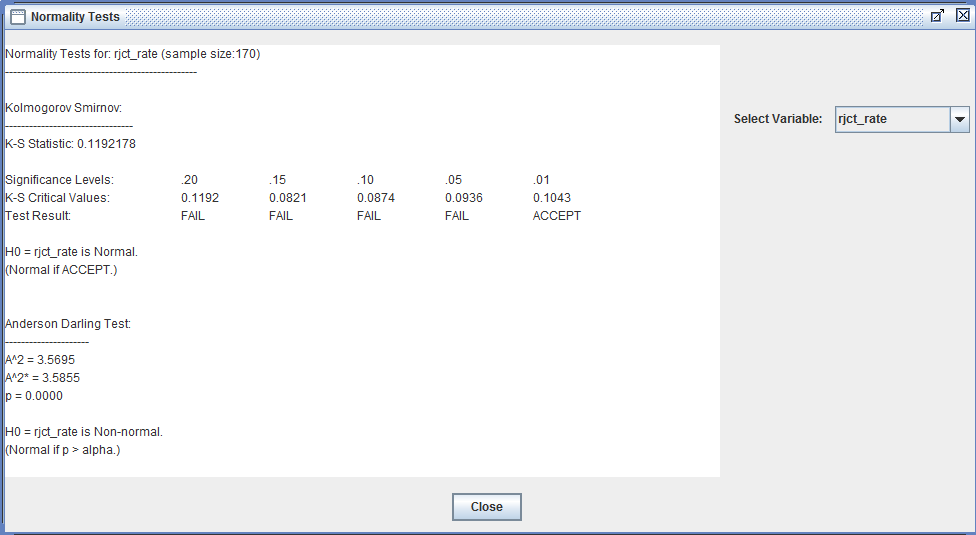
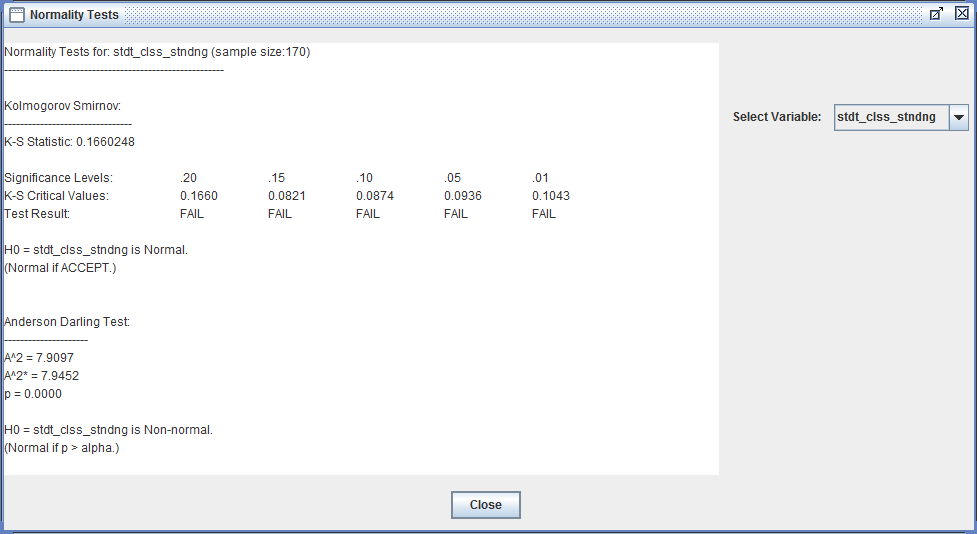
From histograms, we can observe data distribution.

spending\_per\_stdnt, stdnt\_clss\_stnding and rjct\_rate show the tendency of normality violation. Histograms of the three variable are all positive/right skewness, suggesting upper-bound outliers and positive correlation. This tendency is different from Druzdzel &Glymour(1994)’s work, where spending per student and student teacher ration suffers from positive skewness problem.

I conducted normality tests of all variables with “Tools-Normality tests”.

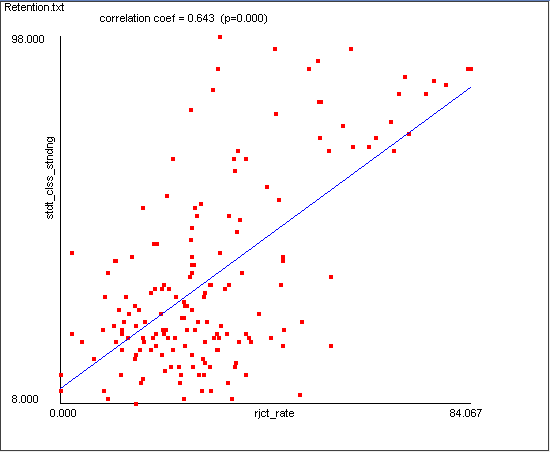
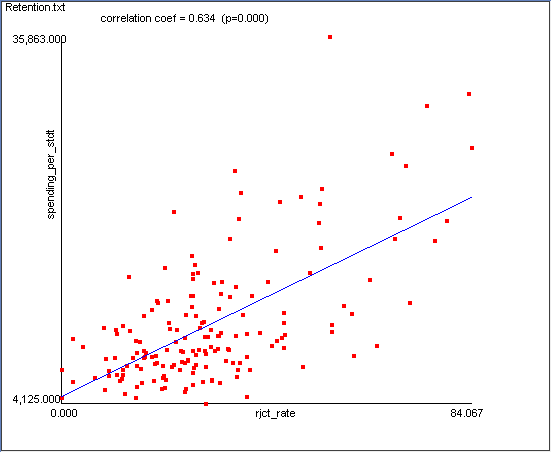
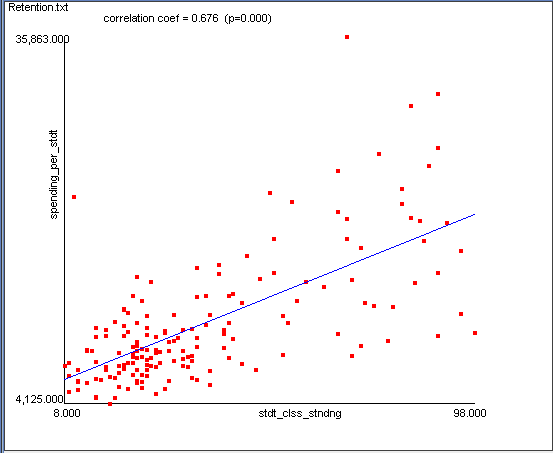
PC algorithm is designed for all-iid-normal-continuous data or all-iid-categorical data. But it’s suggested to be robust to slight violation of normality in its manual. Therefore, I focus on more tolerate indicator.

The result illustrates that spending\_per\_stdnt, stdnt\_clss\_stnding cannot pass normality test and rjct\_rate only passes at the 0.01 level.



1. Preprocessing

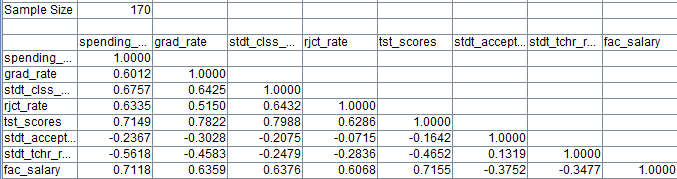
Deleting scatter pattern outliers cannot fix the non-normal problem.

Scatter plots of those three variables are presented. Scatter plots comply with tendency guess and show some pattern outliers. However, deleting 6 outliers as method described in the reference paper has no impact on single variable distribution. Because the skewness is continuous.

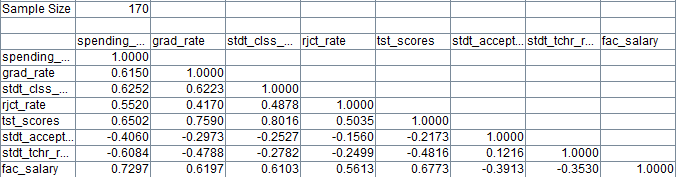
If we delete outliers according to histograms, that would be 30+ cases, a big loss in information.

I choose the “data manipulation-Non parameter transform” to deal with the data, fixing the non-normal problem without influence correlation matrix too much.

Correlation Matrix of Raw Data



Correlation Matrix of Non-Parameter Transformed Data



1. Repeat Druzdzel &Glymour(1994) suggested model

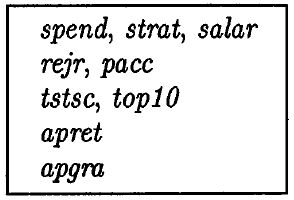
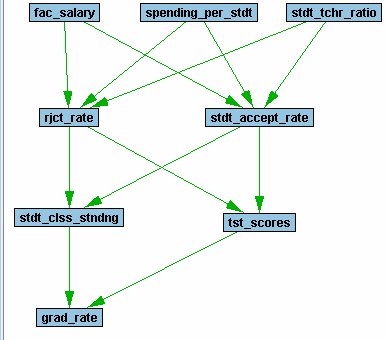
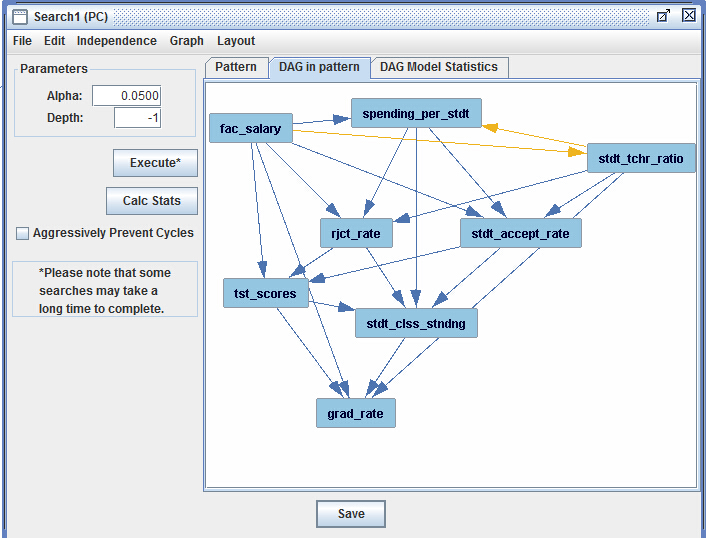
Druzdzel &Glymour(1994) constructed a time sequential tiers of all variables. (“knowledge” box）

Illustration of Druzdzel &Glymour(1994)’s model

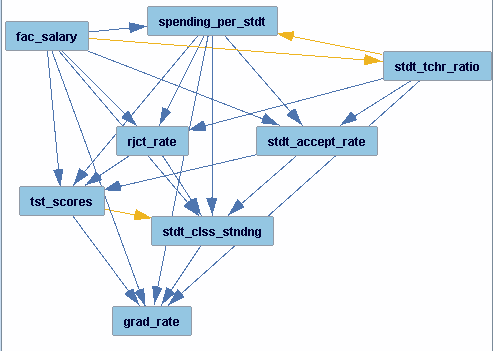
Make Non-parameter transformed data manipulation box and knowledge box as parent nodes, add search box as child node.

Click “Independence-Fisher’s Z”, default alpha level is .05, default depth (The depth value specifies the maximum size of subsets of variables which Tetrad can condition on when testing for independence) is -1.



Degrees of Freedom = 8; Chi-Square = 79.3026; P Value = 0; BIC Score = 38.2162

After trying different alpha (.05, .01) and depth value (1,2,3,4), the best model I got is alpha=.05, depth=1.

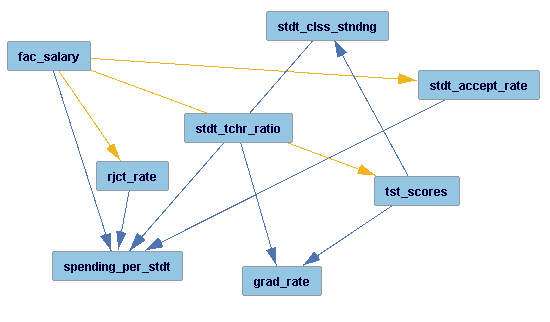


Degrees of Freedom = 5; Chi-Square = 53.6367; P Value = 0; BIC Score = 27.9577

All sequences except stdt\_clss\_stndng are same with hypothesis. But overall, this model is not robust.

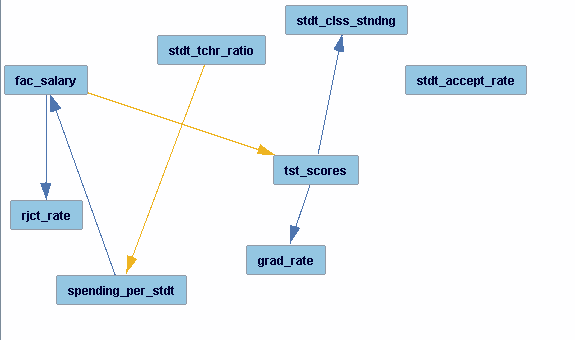
1. Default PC algorithm model

Considering select one presentative from collinear pairs, I conducted “search box” without knowledge input first.



alpha=0.05, depth=-1

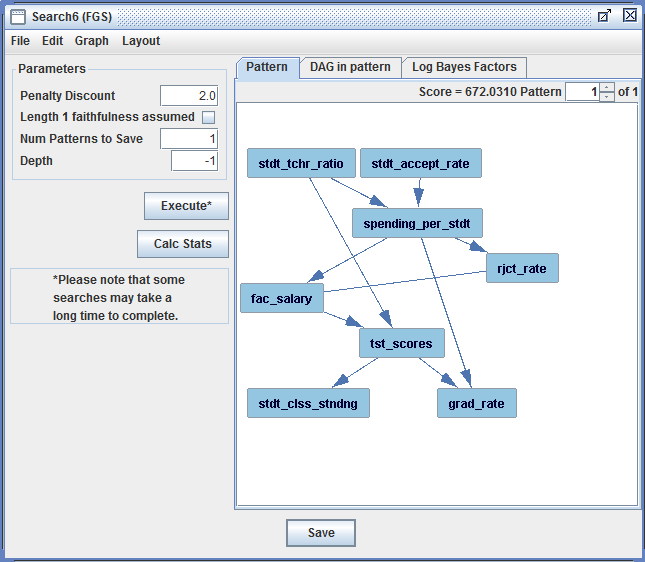
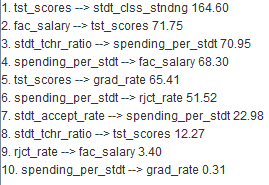
Degrees of Freedom = 17; Chi-Square = 151.0174; P Value = 0; BIC Score = 63.7088



alpha=0.01, depth=-1

Degrees of Freedom = 22; Chi-Square = 239.9391; P Value = 0; BIC Score = 126.9516

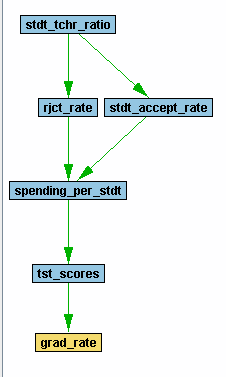
These findings are comply with Druzdzel &Glymour(1994)’s conclusion that test scores is indicator of retention rate.

Also tried the Fast Greedy Search.

1. New Model Exploration

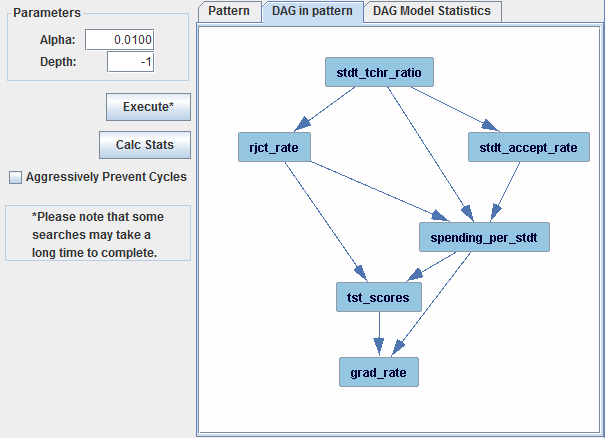
There are high positive correlations in stdt\_clss\_stndng-tst\_scores and spending\_per-stdt-fac\_salary pairs. Exclude one feature from each pair may be better for performing PC algorithm which assumes iid of every variable.

In addition, FGS model suggests variables subscribed by student number quota (student-teacher-ration), like rejection rate and accept rate, will influence budget distribution (spending-per-student and faculty salary). Then, test score will be influenced and finally determine the gradate rate.



Proposed New Model

When test at alpha 0.05 and 0.01, the result is already good.

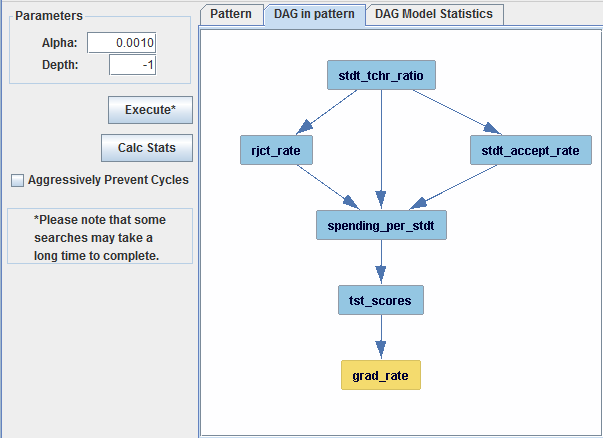


Degrees of Freedom = 6 Chi-Square = 16.0851

P Value = 0.0133

BIC Score = -14.7297

Unlike previous models which decrease alpha to 0.001 will increase BIC, BIC index of this model goes down, suggesting the robustness of this model.



Degrees of Freedom = 8 Chi-Square = 45.1462

P Value = 0

BIC Score = 4.0598

7． Summary

My study support Druzdzel &Glymour(1994)’s conclusion that test score is a main indicator of retention rate (graduate rate in this case). Both repeat Druzdzel &Glymour’s hypothezed model and FGS result comply with the idea. Linear regression shows all the variables explains 62.1% variance where test scores alone explains 57.6%.

My study does not support Druzdzel &Glymour(1994)’s conclusion that student class standing is a main indicator of retention rate (graduate rate in this case). FGS model demonstrates a reversed direction of influence. Considering the high positive correlation between student class standing and test scores, I excluded student class standing variable which also suffering skewness problem in my proposed model.

My study does not support Druzdzel &Glymour(1994)’s guess that spending-per-student, student-teacher-ratio and faculty salary are in the 1st tier; rejection rate and student accept rate are in the 2nd tier. FGS model shows student-teacher-ratio and student accept rate influence spending-per-student. And I build my model by hypothezing that there is a stable budget and student-teacher-ratio policy. Therefore, rejection and acceptance will determine the number of students and then will get resource quota for each student.

Drawbacks of my study, as far as I can numerate, include absent of longitude data, non-normal distribution problem, limited trials on model options and lack of domain knowledge.