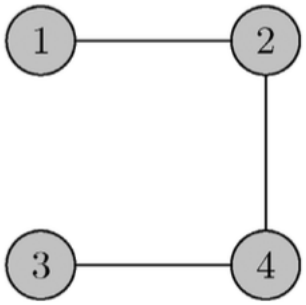
NATIONAL RESEARCH UNIVERSITY

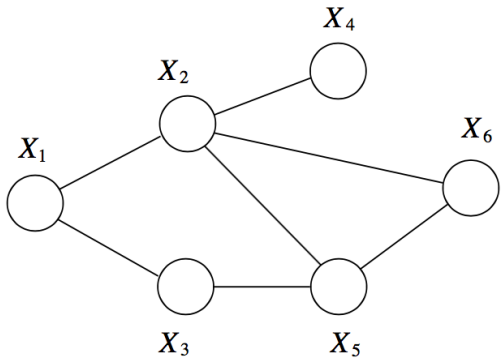
Comparison of different methods of graphical models identification

Ivan Grechikhin

Laboratory of Algorithms, Technologies and Networks Analysis

- G = (V,E) − graph, reflecting the structure of a network. |V| = N
- $X_1,\cdots,X_N$ − random variables, corresponding to the vertices, supposedly from a multivariate normal distribution $N(\mu, \Sigma=\Pi^{-1})$.
- Π - precision matrix, matching the graph, non-zero element means presence of a corresponding edge.
- Zero element in Π means conditional independence of two random variables.
- The problem is a recovery of an underlying graph structure, using observations of given random variables, also the quality and accuracy of this recovery.





- The distribution of sample correlation values is known asymptotically or exactly if true correlation is zero:

$$\sqrt{n-3}\left(z(r_{ij}) - z(\rho_{ij})\right) \to N(0,1) \quad n \to \infty$$

Where *z(r)* is Fisher variance stabilizing z-transform.

- Knowing the distribution of zero elements, we can obtain p-values for every sample correlation value.
- 4 variants of p-value adjustment that control FWER:
  - Bonferroni (simple and step-down Holm procedure)
  - Sidak (simple and step-down Holm procedure)

Drton M., Perlman M.D. (2007). *Multiple Testing and Error Control in Gaussian Graphical Model Selection*. Statistical Science 2007, Vol 22, No.3, 430-449.

Heuristic procedures

- Using distributions of zero and non-zero elements in sample correlation matrix P, we can estimate the likelihood of the matrix P. Using heuristic algorithms we maximize the likelihood function. Through this process, we obtain which sample correlations we consider as non-zero.
- Non-zero elements are considered as distributed uniformly on [-1,1].
- Greedy algorithm:
  - Starts with some initial disposition of zero and non-zero elements.
  - Changes the status of one sample correlation, which improves the likelihood function the most after changing the status (from zero to non-zero or vice versa).
  - The algorithm stops when there can be no simple improvements.

Schafer J., Strimmer K. (2005) *An empirical Bayes approach to inferring large-scale gene association networks*. Bioinformatics, Vol 21, no. 6 2005, pages 754–764

Quality measures

TP – True Positive, sample and true correlations are non-zero
TN – True Negative, sample and true correlations are zero
FP – False Positive, sample correlation is non-zero whereas true correlation is zero (Type I Error)
FN – False Negative, sample correlation is zero whereas true correlation is non-zero (Type II Error)

Measures:

- FWER: E(I(FP > 0))
- Type I Error Number: E(FP)
- Type II Error Number: E(FN)
- False Discovery Rate: FP/(TP+FP)
- False Non-Discovery Rate: FN/(TN+FN)

Experiments

- The precision matrix is generated as $\Pi = A^T A$, where A is matrix with random placements for non-zero elements, each non-zero element is generated from uniform distribution on [-1,1].
- The number of variables is 25.
- The number of observations in experiments are: 100, 200, 300, 400, 500.
- The number of experiments for each case is 1000; for heuristic is 100.
- Different densities of matrices are analyzed: 0 (diagonal), 0.2, 0.6, 0.95.
- The significance level alpha for statistical procedures is 0.1 for all experiments.

Results

q=0.2, Type I Error Number	100	200	300	400	500
Bonferroni	0.076	0.088	0.072	0.066	0.082
Bon. Holm	0.078	0.09	0.076	0.071	0.085
Sidak	0.079	0.096	0.085	0.077	0.091
Sidak Holm	0.08	0.1	0.092	0.084	0.092
Heuristic	**44.66**	**14.26**	**5.62**	**3.13**	**1.75**
q=0.2, Type II Error Number	**100**	**200**	**300**	**400**	**500**
Bonferroni	36.522	28.451	25.676	23.846	22.574
Bon. Holm	36.363	28.336	25.574	23.716	22.425
Sidak	36.414	28.401	25.645	23.805	22.515
Sidak Holm	36.284	28.284	25.529	23.671	22.361
Heuristic	**20.75**	**22.2**	**20.94**	**20.1**	**19.5**

Results

q=0.6, Type I Error Number	100	200	300	400	500
Bonferroni	0.032	0.044	0.031	0.045	0.044
Bon. Holm	0.033	0.052	0.044	0.068	0.061
Sidak	0.032	0.046	0.033	0.05	0.047
Sidak Holm	0.034	0.056	0.046	0.071	0.063
Heuristic	**18.6**	**5.47**	**1.55**	**0.83**	**0.42**
q=0.6, Type II Error Number	**100**	**200**	**300**	**400**	**500**
Bonferroni	157.201	129.713	114.537	105.175	98.14
Bon. Holm	156.8	128.633	113.328	103.769	96.631
Sidak	156.959	129.413	114.299	104.945	97.895
Sidak Holm	156.554	128.347	113.071	103.523	96.375
Heuristic	**88.85**	**90.22**	**90.23**	**88.55**	**86.75**

Results

q=0.95,Type I Error Number	100	200	300	400	500
Bonferroni	0.004	0.002	0.002	0.006	0.009
Bon. Holm	0.004	0.004	0.01	0.008	0.01
Sidak	0.004	0.002	0.002	0.006	0.009
Sidak Holm	0.004	0.005	0.01	0.009	0.01
Heuristic	**2.15**	**0.54**	**0.14**	**0.06**	**0.01**
q=0.95,Type II Error Number	**100**	**200**	**300**	**400**	**500**
Bonferroni	262.147	223.957	198.573	180.204	166.712
Bon. Holm	261.634	221.822	195.288	175.954	162.004
Sidak	261.815	223.51	198.107	179.762	166.256
Sidak Holm	261.278	221.341	194.786	175.395	161.472
Heuristic	**152.84**	**161.34**	**162.22**	**161.6**	**160.72**

Conclusion

- Statistical procedures control FWER at level 0.1, which was expected
- The number of Type II errors decreases with increasing number of observations. For smaller graphs (7x7), previous experiments showed that this number is achieving almost zero much earlier, than in the current case. The percent of false negatives relative to the number of true edges is quite high even for 500 observations.
- All statistical algorithms show practically the same result, where Sidak is slighly better than Bonferronni and step-down procedure is slightly better than simple adjustment.
- The heuristic algorithm does not control FWER. The number of Type I errors for this algorithm decreases with increasing number of observations. The heuristic works better in high density matrices considering number of Type I errors.
- The heuristic algorithm seems to stabilize the number of Type II errors, which is at the same time smaller than for statistical algorithms for given number of observations. The heuristic works better considering the sum of Type I and Type II errors.

NATIONAL RESEARCH
UNIVERSITY

Thank you
for your attention!