CS 722 MACHINE LEARNING
COURSE PROJECT REPORT (FALL 2019)

Semiparametric Differential Graph Models

Kiran Teja Sarvamtotha
Hari Chandana Chintha

ABSTRACT:

Network Analysis plays a key role to study network variations under different conditions than an individual static network. A novel graphical model, Latent Differential Graph Model is proposed. Two semi parametric elliptical distributions are used to represent network under two different conditions and their latent precision matrices difference is used to find variation of these two networks. An estimator for differential graph based on quasi likelihood maximization with nonconvex regularization, which attains faster statistical rate in parameter estimation than the state-of-the-art methods.

INTRODUCTION:

Network analysis has been widely used in various fields to characterize the interdependencies between a group of variables, such as molecular entities including RNAs and proteins in genetic networks. Networks could be modeled as graphical models. A Gaussian graphical model utilized in representing different genes as nodes and the regulation between genes as edges in the graph. In particular, two genes are conditionally independent given the others if and only if the corresponding entry of the precision matrix of the multivariate normal distribution is zero.

PROBLEM STATEMENT:

The gene expression values obtained from high-throughput method, even after being normalized, do not follow a normal distribution. This leads to the inaccuracy in describing the dependency relationships among genes. In order to address this problem, various semiparametric Gaussian graphical models are proposed to relax the Gaussian distribution assumption. Thus by gene network analysis, accurate dependency relationships among genes is achieved.

SIGNIFICANCE OF PROBLEM:

The interactions in many types of networks can change under various environmental and experimental conditions. For example in genetic networks, two genes may be positively conditionally dependent under some conditions but negatively conditionally dependent under others. Therefore, in most of the cases, major attention is attracted not only by a particular individual network but rather by whether and how the network varies with genetic and environmental difference. There comes differential networking analysis, which has emerged as an vital method in differential expression analysis of gene regulatory networks.

EXISTING METHOD:

There are few existing methods to approach this problem and they are as follows:

- **METHOD 1:** It deals with Gaussian Distribution.

  Gene expression values from high throughput method, even after normalization do not follow normal distribution.

- **METHOD 2:** It deals with existing estimators such as:
  - Graphical Lasso (algorithm to estimate precision matrix).
  - Node-wise Regression

- **METHOD 3:** It deals with few of the assumptions such as:
  - Common structural patterns
  - Joint likelihood maximization with group lasso penalty or group bridge penalty.

- **DRAWBACK:** Too restrictive, doubled parameter observations leading to inaccurate dependency relations among genes are considered as the major drawbacks for these existing methods.

PROPOSED METHOD:

The proposed method is as follow:

- **Semi parametric graph models:** To deal with gaussian distribution assumptions.
- **Latent Differential graph model:**
  - Networks under two different conditions are represented by trans elliptical distributions.
  - Connectivity of individual network is encoded by latent precision matrix.
  - Ultimate goal is to estimate parameters.

SEMI PARAMETRIC DIFFERENTIAL GRAPH MODEL:

- **Trans elliptical Distribution:** Gaussian distribution extends to paranormal distribution, likewise a semiparametric extension of elliptical distribution.
- **Kendall's tau Statistic:** Rank based unbiased estimator to find correlation matrix.

- **Latent Differential graph models:** Differential graph is defined with difference between two latent precision matrices.
- **Estimator:** SCAD [Smoothly clipped absolute deviation] and MCP [Minimum Concave Penalty].

ANALYSIS:

There are few analysis drawn by this approach and they are based on few assumptions as below:

- Assumption 1: True covariance matrices and true precision matrices have bounded L1 norm.
- Assumption 2: True difference matrix is bounded to L1,1 norm.
- Assumption 3: MCP and SCAD
- Assumption 4: M-Estimator.
- Assumption 5: Frobenius norm.
- Assumption 6: Oracle estimator.

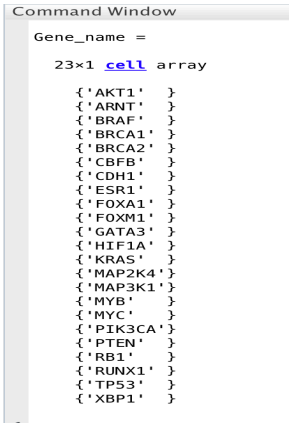
EXPERIMENTS AND RESULTS:

DATASET:

The experiment is carried out on the below mentioned dataset and results are observed.

- Simulated Input DataSet: It contains the below mentioned files and information.
  - hSigma_1.txt - Correlation values (23*23 matrix)
  - hSigma_2.txt - Correlation values (23*23 matrix)
  - TFs.txt – Gene Names (23 Genes)

INPUT GENE NAMES: The input genes which are considered in the experiment.



INPUT CORRELATION MATRIX: the correlation matrix as an input would be the very basic step in the experiment.

Columns 1 through 5

1	-0.3058331	-0.1126909	-0.2243108	-0.1809194
-0.3058331	1	0.1343607	0.06236434	0.09711339
-0.1126909	0.1343607	1	0.2456214	0.3178067
-0.2243108	0.06236434	0.2456214	1	0.546658
-0.1809194	0.09711339	0.3178067	0.546658	1
-0.03344168	-0.09711156	0.005831474	0.02708271	0.2492237
0.05481696	-0.1304048	0.1866509	0.1467408	0.1026204
-0.07113664	0.09321986	0.1768625	0.2186414	0.08337544
0.07481783	0.1604524	0.1419789	0.1574501	0.07356827
0.1626092	-0.1760432	0.174767	0.4522109	0.4155998
0.01511346	0.1088468	0.2501771	0.07281972	-0.05786521
-0.004252637	0.02231357	0.0612721	-0.04589496	0.1387949
-0.2274318	0.2417079	0.1461709	0.1961104	0.3195185
-0.1199922	0.2364502	0.07735402	0.1459936	0.09656309
-0.3055848	0.3435062	0.1545638	0.03677765	0.2360427
-0.06331924	0.1822191	0.09089532	0.07541077	0.008169985
-0.0677077	-0.1536511	-0.0007879837	-0.004719605	-0.1099099
-0.3415557	0.2481597	0.2906183	0.2216469	0.3492643
-0.1129922	0.2084229	0.1289383	0.0495807	0.1685915
-0.1076211	0.2138806	0.1269302	0.08143344	0.4371549
-0.1097556	0.2486442	0.07578413	-0.05424747	0.02937108
0.06629815	0.1089747	-0.09468314	-0.02480759	0.02566813
-0.005269436	0.0180405	0.06339311	0.03410124	-0.0913181

Columns 6 through 10

-0.03344168	0.05481696	-0.07113664	0.07481783	0.1626092
-0.09711156	-0.1304048	0.09321986	0.1604524	-0.1760432
0.005831474	0.1866509	0.1768625	0.1419789	0.174767
0.02708271	0.1467408	0.2186414	0.1574501	0.4522109
0.2492237	0.1026204	0.08337544	0.07356827	0.4155998
1	0.4482548	-0.1233973	-0.1039155	0.1089912
0.4482548	1	0.2185254	0.1172983	0.1347476
-0.1233973	0.2185254	1	0.6130597	0.08942778
-0.1039155	0.1172983	0.6130597	1	0.1139618
0.1089912	0.1347476	0.08942778	0.1139618	1
-0.2295821	0.0563673	0.3204046	0.5002253	0.1031225
0.2063733	0.03350126	-0.08974542	-0.1540338	-0.01052267
0.1809649	0.004185858	0.05000255	0.08536073	0.0756001
-0.1216052	-0.03046367	0.36264	0.3357439	-0.06719264
-0.004692627	-0.1386376	-0.06236924	0.05122916	-0.1770155
-0.1441533	0.03971206	0.491017	0.451435	-0.01969561
-0.05803058	-0.1503735	-0.1806295	-0.2242628	0.07484302
0.17445	0.1784923	0.02946789	-0.04268711	-0.005558652
0.1441017	-0.02695689	-0.05078078	0.1133414	-0.03426118
0.4318841	0.1538562	0.1172524	0.1776297	-0.03126391
0.05558256	-0.08168802	-0.2892593	-0.2001054	-0.1742697
-0.08323663	-0.0953455	-0.04958347	-0.02523993	0.03866298
-0.1416357	0.09600501	0.4279337	0.471108	-0.0163162

Columns 11 through 15

0.01511346	-0.004252637	-0.2274318	-0.1199922	-0.3055848
0.1088468	0.02231357	0.2417079	0.2364502	0.3435062
0.2501771	0.0612721	0.1461709	0.07735402	0.1545638
0.07281972	-0.04589496	0.1961104	0.1459936	0.03677765
-0.05786521	0.1387949	0.3195185	0.09656309	0.2360427
-0.2295821	0.2063733	0.1809649	-0.1216052	-0.004692627
0.0563673	0.03350126	0.004185858	-0.03046367	-0.1386376
0.3204046	-0.08974542	0.05000255	0.36264	-0.06236924
0.5002253	-0.1540338	0.08536073	0.3357439	0.05122916
0.1031225	-0.01052267	0.0756001	-0.06719264	-0.1770155
1	-0.1907526	-0.1170828	0.1970181	-0.125253
-0.1907526	1	0.1279575	-0.07540129	-0.0393146
-0.1170828	0.1279575	1	0.05775454	0.2965795
0.1970181	-0.07540129	0.05775454	1	0.07001134
-0.125253	-0.0393146	0.2965795	0.07001134	1
0.1873318	-0.1310109	0.004911598	0.1402393	0.246916
0.01186502	0.1148001	-0.1039363	-0.1484816	-0.08584669
-0.1397474	0.195687	0.3025795	0.02730123	0.2794567
-0.1481189	0.07033601	0.247436	0.1279574	0.194215
-0.08173362	0.2227595	0.3285539	0.1747863	0.2538832
-0.2026384	0.1182692	0.1652717	-0.04273409	0.3195686
0.03803278	0.02764758	-0.06365088	0.1569678	0.02879049
0.2380601	-0.1204518	-0.02691913	0.3742258	0.005227485

Columns 16 through 20

-0.06331924	-0.0677077	-0.3415557	-0.1129922	-0.1076211
0.1822191	-0.1536511	0.2481597	0.2084229	0.2138806
0.09089532	-0.0007879837	0.2906183	0.1289383	0.1269302
0.07541077	-0.004719605	0.2216469	0.0495807	0.08143344
0.008169985	-0.1099099	0.3492643	0.1685915	0.4371549
-0.1441533	-0.05803058	0.17445	0.1441017	0.4318841
0.03971206	-0.1503735	0.1784923	-0.02695689	0.1538562
0.491017	-0.1806295	0.02946789	-0.05078078	0.1172524
0.451435	-0.2242628	-0.04268711	0.1133414	0.1776297
-0.01969561	0.07484302	-0.005558652	-0.03426118	-0.03126391
0.1873318	0.01186502	-0.1397474	-0.1481189	-0.08173362
-0.1310109	0.1148001	0.195687	0.07033601	0.2227595
0.004911598	-0.1039363	0.3025795	0.247436	0.3285539
0.1402393	-0.1484816	0.02730123	0.1279574	0.1747863
0.246916	-0.08584669	0.2794567	0.194215	0.2538832
1	-0.04090442	0.006839401	-0.07690242	0.03973906
-0.04090442	1	-0.006403996	-0.142378	-0.1585893
0.006839401	-0.006403996	1	0.3092991	0.2852123
-0.07690242	-0.142378	0.3092991	1	0.3053111
0.03973906	-0.1585893	0.2852123	0.3053111	1
0.05871721	0.06772995	0.1913326	0.2487007	0.05217658
0.1526597	0.1135845	-0.1926334	-0.1118879	-0.006602763
0.2280252	-0.07876378	-0.0975313	0.03230164	0.07501388

Columns 21 through 23

-0.1097556	0.06629815	-0.005269436
0.2486442	0.1089747	0.0180405
0.07578413	-0.09468314	0.06339311
-0.05424747	-0.02480759	0.03410124
0.02937108	0.02566813	-0.0913181
0.05558256	-0.08323663	-0.1416357
-0.08168802	-0.0953455	0.09600501
-0.2892593	-0.04958347	0.4279337
-0.2001054	-0.02523993	0.471108
-0.1742697	0.03866298	-0.0163162
-0.2026384	0.03803278	0.2380601
0.1182692	0.02764758	-0.1204518
0.1652717	-0.06365088	-0.02691913
-0.04273409	0.1569678	0.3742258
0.3195686	0.02879049	0.005227485
0.05871721	0.1526597	0.2280252
0.06772995	0.1135845	-0.07876378
0.1913326	-0.1926334	-0.0975313
0.2487007	-0.1118879	0.03230164
0.05217658	-0.006602763	0.07501388
1	0.02281637	-0.09526531
0.02281637	1	-0.09677568
-0.09526531	-0.09677568	1

IMPLEMENTATION:

- The 3 input data files (hSigma_1.txt, hSigma_2.txt and TFs.txt) are loaded.
- Semi parametric graph models**:** To deal with gaussian distribution.
  - Multi variate Normal Distribution
  - Calculate covariance
  - Calculate Frobenius norm
  - Obtain a Covariance matrix

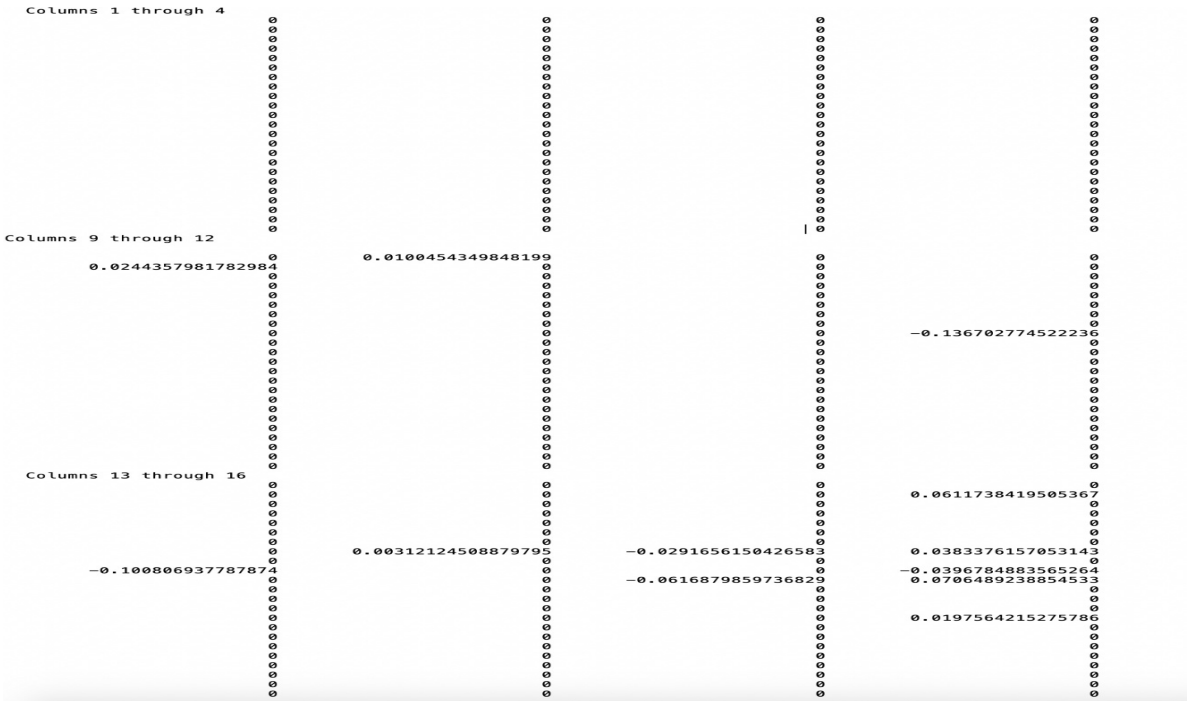
LATENT DIFFERENTIAL GRAPH MODEL:

- Differential Network analysis.
- Calculate Q, Kronecker tensor product for all possible products of Sigma values.
- Calculate b, difference of two latent precision matrices.
- Calculate two trans elliptical distributions and obtain F and G values.
  - F = 1/2 *w'*Q*w - w'*b;
  - G = Q*w - b;

LDGM-L1:

- Process Input: F & G values, W Initial matrix, Lambda.
- Handle case insensitive for Gene Names.
- Compute evaluate functions.
- Outputs the W, Gene regulartory matrix.
- On further refinement we obtain Thetha values.
- Thetha values used to plot the gene regulatory network.

CORRELATION MATRIX:

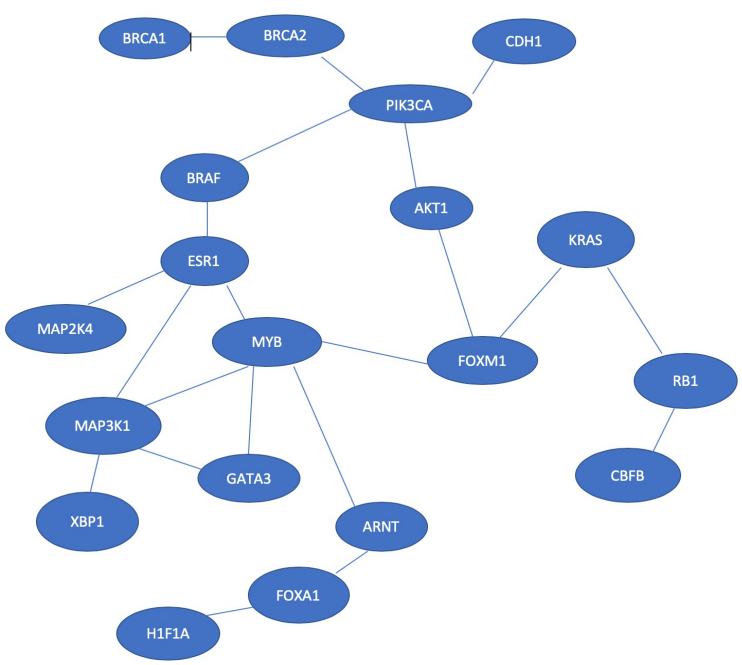


RESULTING GENE IDS AND LIST: The gene ids along with the gene edge list is shows as output which would be a base to form the gene regulatory network.

```
Command Window

edgelist =

     4     5
     3     8
     2     9
     1    10
     9    12
    10    13
     8    14
     8    15
    11    15
     2    16
     8    16
    10    16
    11    16
    15    16
     1    18
     3    18
     5    18
     7    18
     6    20
    13    20
    15    23
```

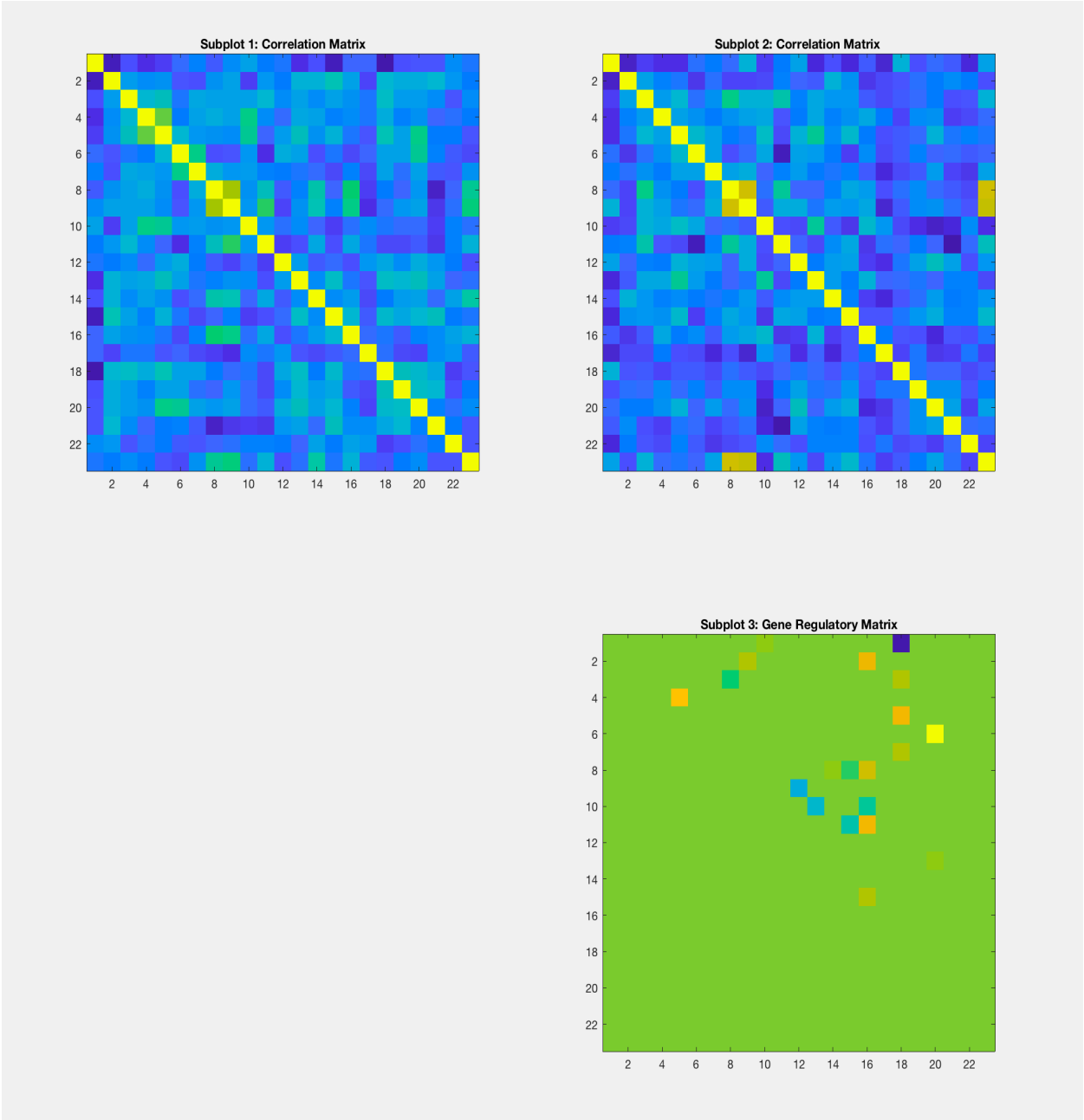
```
Command Window

edgelist_gene =

  21×2 cell array

    {'BRCA1' }    {'BRCA2' }
    {'BRAF'  }    {'ESR1'  }
    {'ARNT'  }    {'FOXA1' }
    {'AKT1'  }    {'FOXM1' }
    {'FOXA1' }    {'HIF1A' }
    {'FOXM1' }    {'KRAS'  }
    {'ESR1'  }    {'MAP2K4'}
    {'ESR1'  }    {'MAP3K1'}
    {'GATA3' }    {'MAP3K1'}
    {'ARNT'  }    {'MYB'   }
    {'ESR1'  }    {'MYB'   }
    {'FOXM1' }    {'MYB'   }
    {'GATA3' }    {'MYB'   }
    {'MAP3K1'}    {'MYB'   }
    {'AKT1'  }    {'PIK3CA'}
    {'BRAF'  }    {'PIK3CA'}
    {'BRCA2' }    {'PIK3CA'}
    {'CDH1'  }    {'PIK3CA'}
    {'CBFB'  }    {'RB1'   }
    {'KRAS'  }    {'RB1'   }
    {'MAP3K1'}    {'XBP1'  }
```

GENE REGULATORY NETWORK: The gene regulatory network is as below which shows the dependency relation exists among the genes.



VISUALIZATION:

The visualization show the correlation matrix for two input files which contains 23 genes along with the final output i.e Gene Regulatory network.



EXTRA CREDIT:

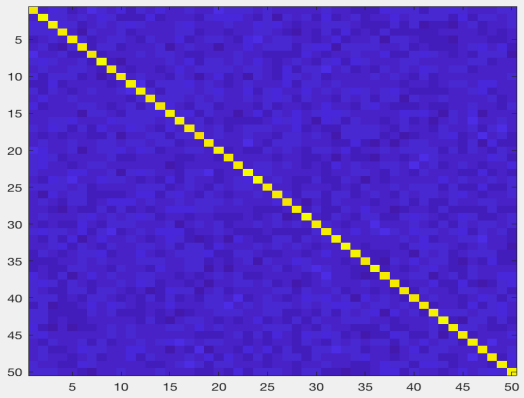
Finding the gene regulatory network on another input data set which contains 50 Genes.

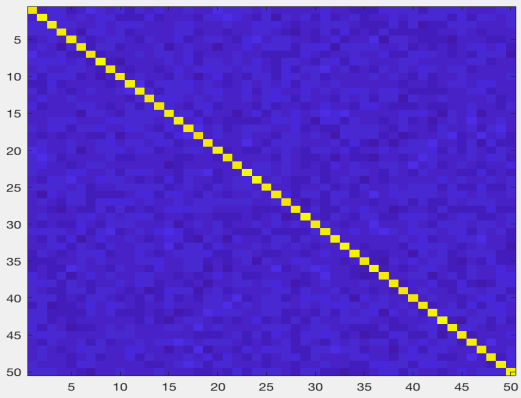
DATASET:

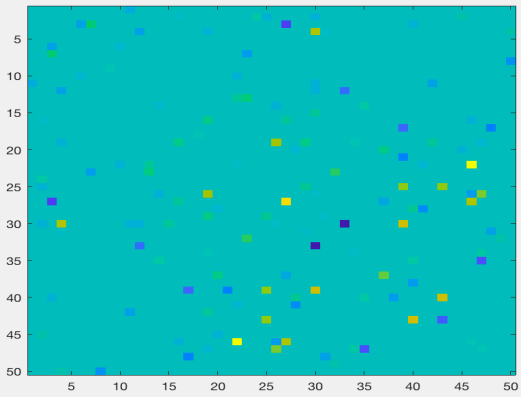
- Simulated Input DataSet:
  - hSigma_1.txt - Correlation values (50*50 matrix)
  - hSigma_2.txt - Correlation values (50*50 matrix)
  - TFs.txt – Gene Names (50 Genes)
  -

VISUALIZATION:

The visualization show the correlation matrix for two input files which contains 50 genes along with the final output i.e Gene Regulatory network.







CONCLUSION:

A semiparametric differential graph model and an estimator for the differential graph based on quasi likelihood maximization are proposed. A nonconvex penalty in our estimator, which results in a faster rate for parameter estimation than existing methods. Experiments on two sets of synthetic data further illustrate our results.

CITATIONS:

- PAPER: https://papers.nips.cc/paper/6529-semiparametric-differential-graph-models.pdf
- https://pdfs.semanticscholar.org/3675/ff7692d5496c17385fddf4123dc2275b64c3.pdf
- BANDYOPADHYAY S, K. D. E. A., MEHTA M (2010). Rewiring of genetic networks in response to dna damage. Science 330 1385–1389
- BARBER, R. F. and KOLAR, M. (2015). Rocket: Robust confidence intervals via kendall's tau for transelliptical graphical models. arXiv preprint arXiv:1502.07641 .
- LOH, P.-L. and WAINWRIGHT, M. J. (2013). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In NIPS
- YUAN, H., XI, R. and DENG, M. (2015). Differential network analysis via the lasso penalized d-trace loss. arXiv preprint arXiv:1511.09188