2017年6月16日　統計数理研究所　オープンハウス

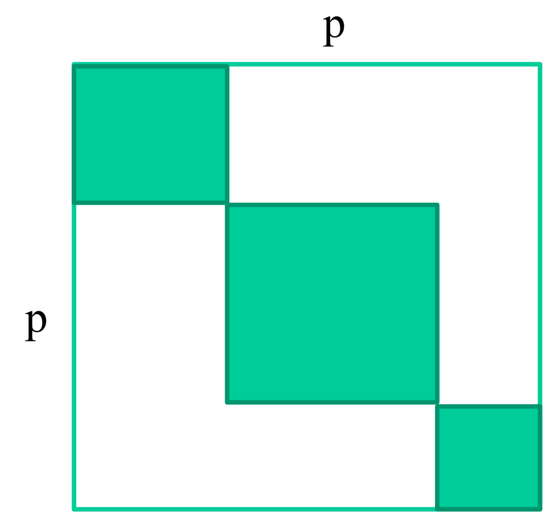
Variable Clustering With the Gaussian Graphical Model

Andrade Daniel　　総合研究大学院大学　統計科学専攻　５年一貫制博士課程４年

Goal:

- **Cluster objects according to their pair-wise correlations. Mean is not useful for distinguishing group of objects.**
- **Model-based clustering including principled methods for selection of number of clusters.**

Precision matrix X



Applications:

- **Detecting independent groups of stocks, sensors, genes, costumers, politicians,…**

Assumptions:

- **Data is generated iid from Normal distribution with mean 0 and block-diagonal precision matrix X.**
- **Number of blocks m is unknown.**

Proposed Approach:

$$\underset{X \succ 0}{\text{minimize}} - \log det(X) + trace\Big(XS\Big)$$

subject to

X is block sparse with exactly m blocks .

S is the sample covariance matrix

→ Express constraint using (unnormalized) Laplacian L

$$\underset{X \succ 0}{\text{minimize}} - \log det(X) + trace\Big(XS\Big)$$

subject to

$$L_{ii} = \sum_{k \neq i} |X_{ik}|^q \,, \quad (1)$$

$$L_{ij} = -|X_{ij}|^q \text{ for } i \neq j, \quad (2)$$

$$rank(L) = p - m \,, \qquad q \in \{1, 2\}$$

↓ Convex relaxation of rank constraint

$$\underset{X \succeq 0}{\text{minimize}} - \log det(X) + trace\Big(XS\Big) + \lambda_m ||L||_*$$

subject to

$$L_{ii} = \sum_{k \neq i} |X_{ik}|^q \,,$$

$$L_{ij} = -|X_{ij}|^q \text{ for } i \neq j \,.$$

←

$$X^* := \underset{X \succeq 0}{\arg\min} - \log det(X) + trace\Big(XS\Big) + \lambda_m \sum_{i \neq j} |X_{ij}|^q \quad (3)$$

q = 1: Problem is equivalent to Graphical Lasso (GL)

q = 2: Squared (SQR) penalty on partial correlations.
Solvable via a new efficient fix-point iteration algorithm.

Model Selection:

$$BIC_{\lambda,m} = -2\mathcal{L}(\hat{X}; S, \mathcal{C}_{\lambda,m}) + \log n \cdot \sum_{C \in \mathcal{C}_{\lambda,m}} \frac{1}{2}(|C|^2 - |C|) \quad (4)$$

where $L(\hat{X}\,;\,S, \mathcal{C}_{\lambda,m})$ is the unpenalized log-likelihood,
and $\mathcal{C}_{\lambda,m}$ is the partition of the variables (found by spectral clustering)

Summary

Algorithm 1 Proposed method for the estimation of variable clusters.

J := set of values for the regularization parameter of the Laplacian L.
K_{max} := maximum number of considered clusters.
for $\lambda \in J$ **do**
  X^* := solve the optimization problem from Equation (3).
  $(\mathbf{e}_1, \ldots, \mathbf{e}_{K_{max}})$:= determine the eigenvectors corresponding to the K_{max} lowest eigenvalues of the Laplacian L (as defined in Equations (1) and (2) with X^*).
  for $k \in \{2, \ldots, K_{max}\}$ **do**
    $\mathcal{C}_{\lambda,k}$:= cluster all variables into k partitions using k-means with $(\mathbf{e}_1, \ldots, \mathbf{e}_k)$.
    $BIC_{\lambda,k}$:= evaluate BIC using Equation (4).
  end for
end for
$\lambda^*, k^* := \arg\max_{\lambda \in J, k \in \{2,\ldots,K_{max}\}} BIC_{\lambda,k}$
return clustering $\mathcal{C}_{\lambda^*,k^*}$

Experiments:

Evaluation of clustering results for "ideal" and "noise-corrupted" synthetic data with 4 clusters, p = 400, and n in {100, 200, 400, 800, 1600}.
Shows the adjusted normalized mutual information (ANMI) and the number of clusters (Clusters).

"ideal"

		100	200	400	800	1600
SQR-Spectral-BIC	ANMI	0.44 (0.02)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)
	Clusters	13.2 (1.94)	4.1 (0.3)	4.0 (0.0)	4.0 (0.0)	4.0 (0.0)
GL-Spectral-BIC	ANMI	0.38 (0.01)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)
	Clusters	13.6 (1.5)	4.0 (0.0)	4.0 (0.0)	4.0 (0.0)	4.0 (0.0)
ID-Spectral-BIC	ANMI	0.11 (0.02)	0.24 (0.02)	0.96 (0.03)	1.0 (0.0)	1.0 (0.0)
	Clusters	9.3 (2.53)	8.5 (2.8)	7.9 (2.81)	4.0 (0.0)	4.0 (0.0)
DPVC	ANMI	0.65 (0.01)	0.75 (0.01)	0.80 (0.01)	0.84 (0.01)	0.86 (0.01)
	Clusters	25.3 (1.55)	16.7 (1.68)	13.1 (1.14)	10.4 (0.49)	9.3 (1.19)
SLC	ANMI	0.29 (0.22)	0.48 (0.15)	0.43 (0.26)	0.75 (0.19)	0.69 (0.21)
	Clusters	2.7 (0.64)	2.8 (0.75)	2.6 (0.66)	3.8 (1.08)	3.6 (1.2)
ALC	ANMI	0.73 (0.03)	0.74 (0.01)	0.75 (0.02)	0.78 (0.02)	0.77 (0.02)
	Clusters	5.0 (0.0)	5.0 (0.0)	5.0 (0.0)	5.0 (0.0)	5.0 (0.0)

"noise-corrupted"

		100	200	400	800	1600
SQR-Spectral-BIC	ANMI	0.49 (0.03)	0.65 (0.06)	0.80 (0.02)	0.92 (0.01)	0.95 (0.02)
	Clusters	13.9 (1.14)	13.2 (1.54)	13.3 (1.85)	12.2 (1.25)	8.8 (1.89)
GL-Spectral-BIC	ANMI	0.38 (0.02)	0.48 (0.02)	0.83 (0.02)	0.68 (0.03)	0.95 (0.03)
	Clusters	13.6 (1.43)	14.4 (1.02)	12.5 (3.14)	11.3 (1.42)	7.9 (2.34)
ID-Spectral-BIC	ANMI	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
	Clusters	14.7 (0.46)	14.7 (0.46)	14.9 (0.3)	15.0 (0.0)	14.5 (0.81)
DPVC	ANMI	0.35 (0.04)	0.33 (0.01)	0.33 (0.03)	0.32 (0.01)	0.31 (0.01)
	Clusters	12.6 (0.92)	8.6 (1.02)	7.8 (0.6)	6.4 (0.49)	5.5 (0.5)
SLC	ANMI	0.01 (0.01)	0.03 (0.02)	0.05 (0.03)	0.06 (0.02)	0.01 (0.01)
	Clusters	2.0 (0.0)	2.4 (0.49)	2.7 (0.9)	3.2 (0.6)	2.1 (0.3)
ALC	ANMI	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.01)	0.0 (0.0)
	Clusters	2.0 (0.0)	2.4 (0.49)	2.2 (0.4)	2.5 (0.5)	2.1 (0.3)

SQR-Spectral-BIC = Use of proposed Algorithm 1 with q = 2.
GL-Spectral-BIC = Use of proposed Algorithm 1 with q = 1.
ID-Spectral-BIC = Use of proposed Algorithm 1 with $X^* = (S + \lambda I)^{-1}$
DPCV = Dirichlet Process Variable Clustering proposed in [Palla et al. 2012]
SLC = Single Linkage Clustering including model selection proposed in [Tan et al. 2015]
ALC = Average Linkage Clustering including model selection proposed in [Tan et al. 2015]

Average runtime in minutes of algorithms for "ideal" synthetic data with p = 400, n = 1600:

SQR-Spectral-BIC	GL-Spectral-BIC	ID-Spectral-BIC	DPVC	SLC	ALC
0.67 (0.0)	5.35 (0.12)	0.14 (0.0)	2.48 (0.06)	1.71 (0.02)	1.59 (0.02)

Conclusions:

- **Combination of Spectral clustering and BIC for clustering selection is useful, even when model assumptions are violated ("noise-corrupted").**
- **Performance of SQR-Spectral-BIC is comparable or better than GL-Spectral-BIC, while being almost 10 times faster.**
- **SQR-Spectral-BIC and GL-Spectral-BIC can perform considerably better than previously proposed methods (also on real data, results omitted here).**