Graphical Models for high-dimensional data

Kartik Ravisankar

UMD, AMSC

December 4, 2024

Outline

1. Introduction
   - Factorization
   - Conditional Independence
   - Hammersley Clifford Equivalence
2. Estimation of Graphical Models
   - Introduction
   - Gaussian Graphical Models
3. Summary
   - Topics covered
   - Other topics
   - Thanks!

Introduction

- Consider an undirected graph $\mathcal{G} := (V, E)$ which consists of a set of vertices $V = \{1, 2, \ldots, d\}$ connected by a set of edges E. An edge (j, k) is an undirected edge joining vertices j and k respectively.
- Associate each vertex $j \in V$, a random variable X_j, taking values in $\mathcal{X}_j$, and consider the joint probability distribution $\mathcal{P}$ of the d-dimensional random vector $X = (X_1, X_2, \ldots, X_d)$.

We are interested in the connection between the structure of $\mathcal{P}$ and the underlying graph $\mathcal{G}$. There are two ways to connect the probabilistic and graphical structures: one based on *factorization*, and the second based on *conditional independence* properties.

Hammersley Clifford Theorem

Factorization = conditional independence

Defintions

Clique /klēk,klik/

a small group of people, with shared interests or other features in common, who spend time together and do not readily allow others to join them.

Clique - the useful version

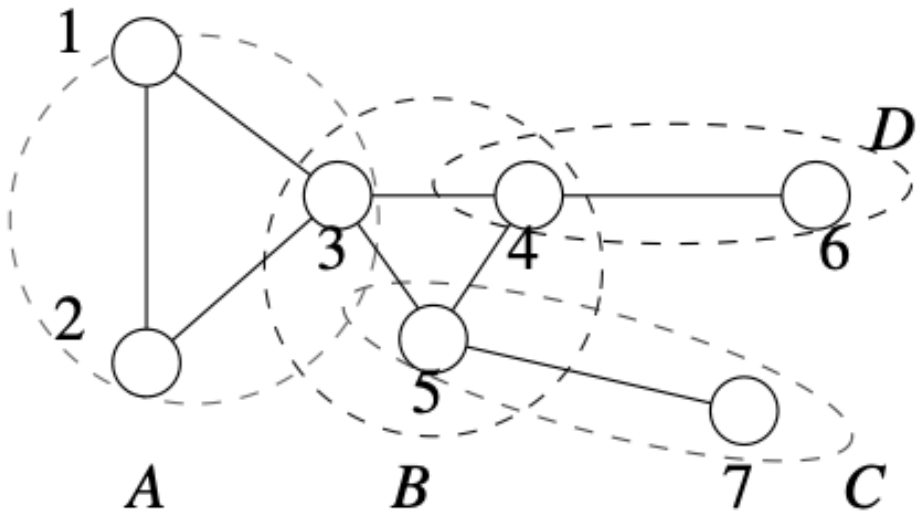
A subset of vertices that are all joined by edges. For all distinct vertices $(j, k) \in C$, there exists a $(j, k) \in E$. A maximal clique is a clique that is not a subset of any other clique.

We denote $\mathscr{C}$ to denote the set of all cliques in $\mathcal{G}$. Each vertex, by definition, is a singleton clique.

Example

- $A = \{1, 2, 3\}$
- $B = \{3, 4, 5\}$
- $C = \{5, 7\}$
- $D = \{4, 6\}$

A,B,C, and D are all maximal cliques!



(a)

Factorization

For all $C \in \mathscr{C}$, we use ψ_C to denote a function of the subvector $x_C := \{x_j : j \in C\}$. We call ψ_C to be a *clique compatibility function*, whose inputs are the cartesian product space $\mathcal{X}_c := \bigotimes_{j \in C} \mathcal{X}_j$, and returns a non-negative real number.

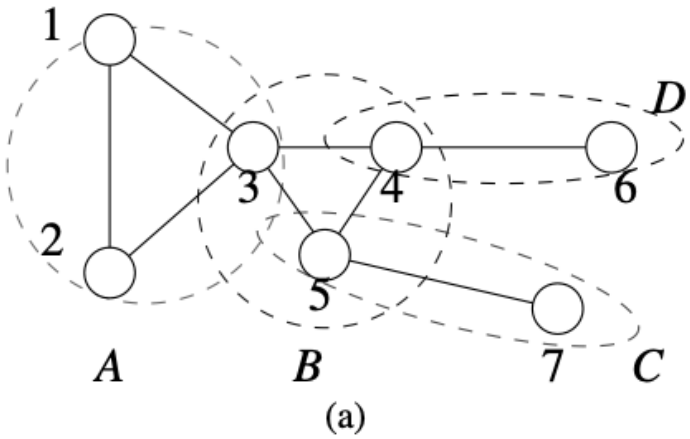
Factorization

The random vector $(X_1, X_2, \ldots X_d)$ factories according to graph G if its density function P can be represented as:

$$p(x_1, \ldots, x_d) \propto \prod_{C \in \mathscr{C}} \psi_C(x_C) \tag{1}$$

for some collection of clique compatibility functions $\psi_C : \mathcal{X}_C \to 0 \cup \mathcal{R}^+$

Factorization



(a)

In the above example, any density that factorizes according to graph G must have the following form

$$p(x_1, \ldots, x_7) \propto \psi_{123}(x_1, x_2, x_3)\psi_{345}(x_3, x_4, x_5)\psi_{46}(x_4, x_6)\psi_{57}(x_5, x_7) \quad (2)$$

WLOG, the product of cliques can always be restricted to the set of all maximal cliques (by redefining the clique compatibility function as necessary).

Multivariate Gaussian Factorization

Consider a d-dimensional non-degenerate gaussian distribution with zero mean (WLOG) can be parametrized by the precision matrix $\Theta := \Sigma^{-1}$ as follows.

$$p(x_1, \ldots, x_d; \Theta) = \frac{\sqrt{det(\Theta)}}{2\pi^{d/2}} exp\{-\frac{1}{2} x^T \Theta x\} \tag{3}$$

Upon expanding the quadratic form,

$$exp\{-\frac{1}{2} \sum_{(j,k)\in E} \Theta_{jk} x_j x_k\} = \prod_{(j,k)\in E} exp\{-\frac{1}{2} \Theta_{jk} x_j x_k\} \tag{4}$$

Gaussian factorization

If we define $\psi_{j,k} = e^{-\frac{1}{2}\Theta_{jk} x_j x_k}$, any zero-mean Gaussian distribution can be factorized in terms of functions on edges, or cliques of size two, **even if the underlying graph has higher order cliques**.

Ising Model factorization

Ising is a theoretical model in statistical physics, that was originally developed to describe ferromagnetism. (100 year anniversary this year)

https://doi.org/10.1038/s44260-024-00012-0

The Ising model celebrates a century of interdisciplinary contributions

Given an undirected graph $G = (V, E)$, we have a factorization of the form,

$$p(x_1, x_2, \ldots, x_d; \theta^*) = \frac{1}{Z(\theta^*)} exp\left\{ \sum_{j \in V} {\theta_j}^* x_j + \sum_{(j,k) \in E} {\theta_{jk}}^* x_j x_k \right\} \quad (5)$$

Conditional Independence

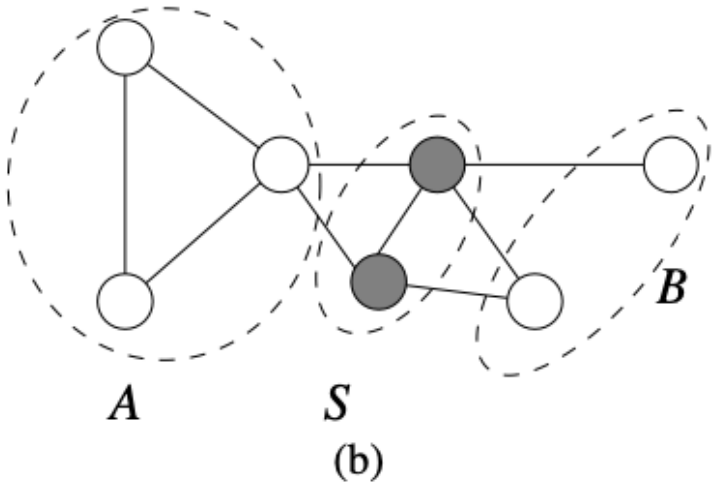
Conditional Independence is an alternate approach to connect the probabilistic and graphical structures, using certain conditional independence statements defined by the graph.

Vertex cut set

Let S be the set of vertices, which when removed results in a vertex-induced subgraph $G(V \backslash S)$ with vertices $V \backslash S$ and edges $E(V \backslash S) := \{(j,k) \in E | j,k \in V \backslash S\}$. The set S is a vertex cutset if the residual graph $G(V \backslash S)$ consists of two or more disconnected non-empty components.

Conditional Independence

S is a vertex cut set in the graph below, as $V \backslash S$ results in two disjoint components A and B respectively.



Conditional Independence

For any subset $A \subseteq V$, let $X_A := (X_j, j \in A)$. For any three disjoint subsets, say A, B, and S, of the vertex set V, we use $X_A \perp\!\!\!\perp X_B | X_S$ to mean that the subvector X_A is conditionally independent of X_B given X_S.

Markov Property

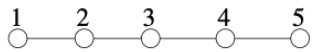
Based on what we have learnt from cut sets, we define if a random vector X is Markov with respect to graph G or not.

Markov property

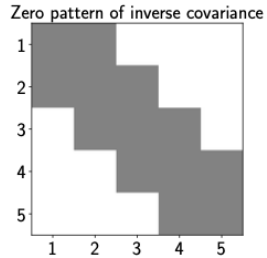
A random vector $X = (X_1, \ldots, X_d)$ is **Markov** with respect to graph G, if for all vertex cutsets S breaking the graph into two disjoint components (A and B), the conditional independence $X_A \perp\!\!\!\perp X_B | X_S$ hold true.

Examples - Markov Chain

The Markov chain provides the simplest example. Consider a chain graph on the vertex set $V = \{1, 2, \ldots, d\}$ containing the edges $(j, j+1)$ $\forall j = 1, 2, \ldots, d-1$. Each vertex $j \in \{2, 3, \ldots, d-1\}$ acts as a cut set for



(a)



(b)

such a chain graph.
In the above figure,

- **2** splits the graph into 1 and 3,4,5.
- **3** splits the graph into 1,2 and 4,5.
- **4** splits the graph into 1,2,3 and 5.

These singleton cut sets essentially define the Markov property of a Markov time series model, breaking the model into *past* and *future* states. Given a current state X_j, the future states X_F are conditionally independent of the past states X_P.

Examples - Neighborhood-based cuts

For any vertex $j \in V$, we define a neighbourhood set $N_j = \{k \in V | (j,k) \in E\}$.

- **N(j) is always a vertex cut set!** A non-trivial one as long as j is not connected to every other vertex.
- It splits the graph into two disjoint components $A = \{j\}$ and $B = \{V \backslash N(j) \cup j\}$.

Theorem

HC Equivalence - Thm 11.8

For a given undirected graph and any random vector $X = \{X_1, X_2, \ldots, X_d\}$ with strict positive density p, the following properties are equivalent.

1. The random vector X factories according to the structure of the graph G as referenced earlier.
2. The random vector X is Markov with respect to the graph G

Proof

Proof is done one-way. Factorization property $\implies$ *Markov property*. We prove this by defining three subsets of cliques. $\mathscr{C}_A = \{C \in \mathscr{C} | C \cap A \neq \phi\}$, $\mathscr{C}_B = \{C \in \mathscr{C} | C \cap B \neq \phi\}$, and $\mathscr{C}_S = \{C \in \mathscr{C} | C \subset S\}$. We claim that these 3 subsets form a disjoint partition of $\mathscr{C}$.

- Given any clique C, it is either contained entirely in S or it has a non-trivial intersection with either A or B, thus proving the union property.
- It is obvious that $\mathscr{C}_S$ is disjoint from $\mathscr{C}_A$ and $\mathscr{C}_B$. If $\mathscr{C}_A \cap \mathscr{C}_B = C$, then $\exists (a, b) \in E$ such that $a \in A$ and $b \in B$ with $(a, b) \in C$. This is not possible as S is a **vertex cutset!**

Proof (contd)

$$p(x_A, x_B, x_S) = \frac{1}{Z}\Big\{\prod_{C\in\mathscr{C}_A}\psi_C(x_C)\Big\}\Big\{\prod_{C\in\mathscr{C}_B}\psi_C(x_C)\Big\}\Big\{\prod_{C\in\mathscr{C}_S}\psi_C(x_C)\Big\} \quad (6)$$

Define $Z_A(x_S) = \sum_{x_A}\psi_A(x_A, x_S)$ and $Z_B(x_B, x_S) = \sum_{x_B}\psi_A(x_B, x_S)$. We can obtain the marginals as follows:

$$p(x_S) = \frac{Z_A(x_S)Z_B(x_S)}{Z}\psi_S(x_S) \quad (7)$$

$$p(x_A, x_S) = \frac{Z_B(x_S)}{Z}\psi_A(x_A, x_S)\psi_S(x_S) \quad (8)$$

Proof (contd)

$$\frac{p(x_A, x_B, x_S)}{p(x_S)} = \frac{\psi_A(x_A, x_S)\psi_B(x_B, x_S)}{Z_A(x_S)Z_B(x_S)} \tag{9}$$

$$\frac{p(x_A, x_S)}{p(x_S)} = \frac{\psi_A(x_A, x_S)}{\psi_S(x_S)} \tag{10}$$

$$\frac{p(x_B, x_S)}{p(x_S)} = \frac{\psi_B(x_B, x_S)}{\psi_S(x_S)} \tag{11}$$

Final result

$$p(x_A, x_B|x_S) = \frac{p(x_A, x_B, x_S)}{p(x_S)} = \frac{p(x_A, x_S)}{p(x_S)}\frac{p(x_B, x_S)}{p(x_S)} = p(x_A|x_S)p(x_B|x_S)$$

The other direction can be proven quite similarly but is not presented in Wainwright.

Typical problems

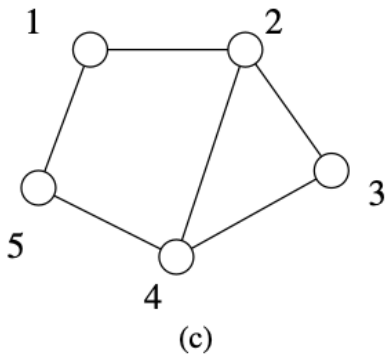
Consider samples $x_1, \ldots, x_n$ where each $x_i = (x_{i1}, \ldots, x_{id})$, drawn from a graph-structured probability distribution.

Graphical model quandary

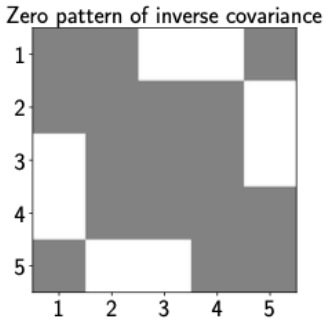
- Graph structure G is known, and we want to estimate the compatibility function $\psi_C, C \in \mathscr{C}$.
- Graph structure is unknown, we need to estimate G (or specifically E) and the clique compatibility functions.

Problem Setting

For a Gaussian graphical model, the factorization property can be parametrized by the inverse covariance or precision matrix $\Theta^* = \Sigma^{-1}$. Based on HC equivalence, it ensures that $\Theta^*_{jk} = 0$ where $(j, k) \notin E$.



(c)



(d)

Since the mean can be easily estimated, we assume it to be 0 without loss of generality. The problem then is the estimation of Θ^*. In cases, where we need to estimate the graphical model, the goal is to estimate E. If we denote the estimate as $\hat{E}$ and the corresponding covariance matrix $\hat{\Theta}$.

Unrestricted MLE

Using the precision matrix parametrization of multivariate gaussian distribution, we get the negative log likelihood to be as follows: (upon simplification and rescaling)

$$L_n(\Theta) = \langle\langle \Theta, \hat{\Sigma} \rangle\rangle - log \text{ det } (\Theta) \tag{12}$$

where $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T$

Unrestricted MLE $\hat{\Theta} = \hat{\Sigma}^{-1}$, but in a high-dimensional setting, Σ is always rank deficient and thus, **MLE does not exist!**.

Regularized L_1 Graphical Lasso

When the graph G has few edges (sparse adjacency matrix), a natural form of regularization is to impose an L_1 constraint on the entries of Θ. (Recall that while L_0 is natural, L_1 is used as a convex surrogate)

$$\hat{\Theta} = argmin_{\Theta \in S^{d \times d}} \left\{ \langle\langle \Theta, \hat{\Sigma} \rangle\rangle - log \text{ det } (\Theta) + \lambda_n \|\Theta\|_{1,\text{off}} \right\} \tag{13}$$

Here, the l_1 norm is applied to the off-diagonal elements of Θ. One could also imagine penalizing the diagonal entries of Θ, but since they must be positive for any non-degenerate inverse covariance, doing so introduces additional bias.

Can we find bounds?

Can we find bounds for $\|\hat{\Theta} - \Theta^*\|_F$?

Frobenius norm bounds for graphical lasso

Suppose the matrix Θ^* has at most m non zero entries per row, and $\lambda_n = 8\sigma^2\left(\sqrt{\frac{logd}{n}} + \delta\right)$ for some $\in (0, 1]$. Then, as long as $6(\|\Theta^*\|_2 + 1)^2\lambda_n\sqrt{md} < 1$, then the graphical lasso estimate $\hat{\Theta}$ satisfies

$$\|\Theta^* - \hat{\Theta}\|_F \leq \frac{9md\lambda_n^2}{(\|\Theta^*\|_2 + 1)^4} \text{ wp } \geq 1 - 8e^{\frac{-n\delta^2}{16}} \tag{14}$$

A sample covariance matrix $\hat{\Theta}$ formed by n iid samples of a zero-mean random vector in which each coordinate has σ sub Gaussian tails.

Key Results from Chapter 9

Subspace Lipschitz Constant

For any subspace S of R^d , the subspace Lipschitz constant with respect to the pair $(\phi, \|\cdot\|)$ is given by:

$$\psi(S) = sup_{u \in S \setminus \{0\}} \frac{\phi(u)}{\|u\|} \tag{15}$$

Restricted Strong Convexity

For a given norm $\|\cdot\|$ and a regularizer $\phi(\cdot)$, the cost function satisfies restricted strong convexity with radius R and curvature κ and tolerance τ_n^2 if

$$\epsilon_n(\Delta) \geq \frac{\kappa}{2}\|\Delta\|^2 - \tau_n^2 \phi^2(\Delta) \; \forall \Delta \in \mathcal{B}(R) \tag{16}$$

Key Results from Chapter 9

Bounds for general models - 9.20

In addition to *certain* regularity conditions, the optimal parameter θ^* belongs to M. Then any optimal solution $\hat{\theta}$ to the optimization problem satisfies the bounds

$$\phi\left(\hat{\theta} - \theta^*\right) \leq 6\frac{\lambda_n}{\kappa}\psi^2(\bar{M}) \tag{17}$$

$$\|\hat{\theta} - \theta^*\|^2 \leq \frac{9\lambda_n^2\psi^2(\bar{M})}{\kappa^2} \tag{18}$$

Key Results from Chapter 9

Regularity conditions

- Cost function is convex and satisfies the local restricted strong convexity (RSC) condition with curvature κ, radius R and tolerance $\hat{\tau_n}$ wrt an inner product induced norm $\|\cdot\|$
- $\exists$ a pair of subspace $M \subseteq \bar{M}^{\perp}$ such that the regularizer decomposes over $(M, M^{\perp})$

Corollary 9.20 holds for the regularity conditions and is conditioned on the "good" event of the regularizer where the score function is not large in terms of the dual norm ϕ^*

$$G(\lambda_n) = \left\{ \phi^*(\nabla L_n(\theta^*)) \leq \lambda_n/2 \right\} \tag{19}$$

Proof - Verify Strong Convexity

Let $\mathscr{B}_F(1) = \{\Delta \in S^{d\times d} | \|\Delta\|_F \leq 1\}$ denote the set of symmetric matrices with Frobenius norm at most one.

Twice differentiable loss function

$\nabla L_n(\Theta) = \hat{\Sigma} - \Theta^{-1}$ and $\nabla^2 L_n(\Theta) = \Theta^{-1} \bigotimes \Theta^{-1}$

For any $\Delta \in \mathscr{B}_F(1)$, a Taylor series expansion (intermediate version) leads to

$$L_n(\Theta^* + \Delta) - L_n(\Theta) = \langle\langle \nabla L_n(\Theta^*), \Delta \rangle\rangle + \frac{1}{2} vec(\Delta)^T \nabla^2 L_n(\Theta^* + t\Delta) vec(\Delta) \tag{20}$$

$$\epsilon_n(\Delta) = \frac{1}{2} vec(\Delta)^T \nabla^2 L_n(\Theta^* + t\Delta) vec(\Delta) \tag{21}$$

$$\epsilon_n(\Delta) \geq \frac{1}{2} \gamma_{min}\{\nabla^2 L_n(\Theta^* + t\Delta)\} \|vec(\Delta)\|^2 \tag{22}$$

Proof- Verifying Strong Convexity

For symmetric invertible matrix A, $\|A^{-1} \bigotimes A^{-1}\|_2 = \frac{1}{\|A\|_2}$

$$\epsilon_n(\Delta) \geq \frac{1}{2} \frac{\|\Delta\|_F^2}{\|\Theta^* + t\Delta\|_F^2} \tag{23}$$

Employing the $\triangle$ inequality in conjunction with $t\|\Delta\|_2 \leq t\|\Delta\|_F \leq 1$, we get

$$\epsilon_n(\Delta) \geq \frac{1}{2(\|\Theta\|_2 + 1)^2} \|\Delta\|_F^2 \tag{24}$$

24 proves that the loss function satisfies the RSC condition shown in 16 with $\kappa = (\|\Theta\|_2 + 1)^{-2}$ and $\tau_n = 0$

Proof - Subspace Lipschitz Constant

Next we introduce a subspace suitable for the application of 17 and 18 to the graphical Lasso. Letting S denote the support set of Θ^*, we define the subspace

$$M(S) = \{\Theta \in \mathcal{R}^{d \times d} | \Theta_{jk} = 0 \ \forall \ (j,k) \notin E\} \tag{25}$$

With this choice of $M(\cdot)$, we get

$$\psi^2(M(S)) = sup_{\Theta \in M(S)} \frac{(\sum_{j \neq k} |\Theta_{jk}|)^2}{\|\Theta\|_F^2} \leq |S| \leq md \tag{26}$$

The last inequality holds as we allow for **at most m non-zero entries per row.**

Proof - Verifying Event

Next, we need to verify the stated choice of the regularization parameter λ_n satisfied the conditions of 17 and 18 with high probability.
Recall that $\nabla L_n(\theta^*) = \hat{\Sigma} - \Sigma$. The dual norm defined by $\|\cdot\|_{1,off}$ is the L_∞ norm on off-diagonal elements which we will denote as $\|\cdot\|_{max,off}$.
Using lemma 6.26 (without proof), we have

$$P(\|\hat{\Sigma} - \Sigma\|_{max,off} \geq \sigma^2 t) \leq 8e^{\frac{-n}{16} min(t,t^2)+2logd} \ \forall t > 0 \tag{27}$$

Setting $t = \lambda_n/\sigma^2$, we can verify that $G(\lambda_n)$ is highly probable.
Proposition 9.13 implies that the error matrix $\hat{\Delta}$ satisfies the bound $\|\Delta_{S^c}\|_1 \leq 3\|\hat{\Delta_S}\|_1$, and hence

$$\|\hat{\Delta}\|_1 \leq \|\Delta_S\|_1 \leq 4\sqrt{md}\|\hat{\Delta}\|_F \tag{28}$$

The second inequality holds true as $|S| \leq md$ (recall the number of non-zero rows)

Localizing the error matrix

Recall from equation 24, we know that $-L_n(\Theta^*) + L_n(\Theta^* + \Delta) + \langle\langle \nabla L_n(\Theta^* + \Delta), -\Delta \rangle\rangle \geq \frac{\kappa}{2}\|\Delta\|_F^2$. Also,

Implications of strong convexity

If f is κ strongly convex around x:
$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\kappa}{2}\|y - x\|_2^2$ holds for all vectors z in a ball B_2 centered at x. $B_2(x) = \{z$
—z-x$\|_2 \leq 1\}$, then

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \frac{\kappa}{2}\|y - x\|_2 \tag{29}$$

Using the 29, we get

$$\langle\langle L_n(\Theta^* + \Delta) - \nabla L_n(\Theta^*), \Delta \rangle\rangle \geq \kappa\|\Delta\|_F \ \forall \Delta \in S^{d\times d} \backslash B_F(1) \tag{30}$$

Localizing the error matrix

Since $\hat{\Theta}$ is optimal, we have $\langle\langle \nabla L_n(\Theta^* + \hat{\Delta}) + \lambda_n \hat{Z}, \hat{\Delta} \rangle\rangle = 0$, where $\hat{Z} \in \partial \|\Theta\|_{1,off}$ is a subgradient matrix for the elementwise L_1 norm. By adding and subtracting terms, we can find that

$$\langle\langle \nabla L_n(\Theta^* + \hat{\Delta}) - \nabla L_n(\Theta^*), \hat{\Delta} \rangle\rangle \leq \lambda_n |\langle\langle \hat{Z}, \hat{\Delta} \rangle\rangle + |\langle\langle \nabla L_n(\Theta^*), \hat{\Delta} \rangle\rangle| \quad (31)$$

If $\|\hat{\Delta}\|_F \geq 1$, then from the LB of 29, we get

$$\langle\langle \nabla L_n(\Theta^* + \hat{\Delta}) - \nabla L_n(\Theta^*), \hat{\Delta} \rangle\rangle \leq \left\{ \lambda_n + \|\nabla L_n(\Theta^*)\|_{max} \right\} \|\hat{\Delta}\|_1 \quad (32)$$

Since $\|\nabla L_n(\Theta^*)\|_{max} \leq \lambda_n/2$ under $G(\lambda_n)$, the RHS is atmost $3\frac{\lambda_n}{2}\|\hat{\Delta}\|_1 \leq 6\lambda_n\sqrt{md}\|\Delta\|_F$ from 28.
From the 29, we get $\kappa\|\hat{\Delta}\|_F \leq \frac{3\lambda_n}{2}\|\hat{\Delta}\|_1 \leq 6\lambda_n\sqrt{md}\|\hat{\Delta}\|_F$.
This leads to a contradiction whenever $\frac{6\lambda_n\sqrt{md}}{\kappa} < 1$, thus proving that $\|\hat{\Delta}\|_F \leq 1$

Frobenius Norm Bound of Graphical Lasso

We were able to prove the frobenius norm bound of the gaussian graphical lasso from corollary 9.20 in the text book, by verifying the necessary regularity conditions hold true.

Frobenius Norm bound

$$\|\Theta^* - \hat{\Theta}\|_F \leq \frac{9md\lambda_n^2}{(\|\Theta^*\|_2 + 1)^4} \text{ wp } \geq 1 - 8e^{\frac{-n\delta^2}{16}}$$

This is a crude result, and it only guarantees $\hat{\Theta}$ is close to Θ^* with respect to the Frobenius norm. Nothing is said about the edge structure of the graph. The result precludes the setting $n << d$, as the proof implies that imply that n must be lower bounded by a constant multiple of $md\ log(d)$, which is larger than d.

More practical result

- We now turn to a more refined type of result, namely one that allows for high-dimensional scaling ($d >> n$), and moreover guarantees that the graphical Lasso estimate $\hat{\Theta}$ correctly selects all the edges of the graph.
- Such an edge selection result can be guaranteed by first proving that $\hat{\Theta}$ is close to the true precision matrix Θ^* in the element wise L_∞ norm on the matrix elements (denoted by $\|\cdot\|_{max}$).
- This problem can also be converted to bounds on the L_2 matrix operator/spectral norm.
- **Edge selection in a Gaussian graphical model $\approx$ variable selection in a sparse linear model**

More practical result

- Chugang spoke on Chapter 7 earlier this semester where we discussed **incoherence conditions** which limits the influence of irrelevant variables on relevant ones.
- In the least squares regression setting, the incoherence condition was imposed on the design matrix, or the hessian of the objective function.
- We will follow the latter approach by imposing the condition on the hessian of the loss function. $\nabla^2 L_n(\Theta) = \Theta^{-1} \bigotimes \Theta^{-1}$. The incoherence condition must be satisfied by the d^2 dimensional matrix $\Gamma^* = \nabla^2 L_n(\Theta^*)$.

More practical result

Let $S = E \cup \{(j,j)|j \in V\}$, the set of edges including $(j,k),(k,j)$ and self edges (j,j). Naturally, let $S^c = \{(V \times V) \backslash S\}$, then we say the matrix Γ^* is α incoherent if

$$max_{e \in S^c} \|\Gamma^*_{es}(\Gamma^*_{ss})^{-1}\|_1 \leq 1 - \alpha \text{ for some } \alpha \in (0,1] \tag{33}$$

Proposition 11.10

Consider a d-dimensional gaussian distribution based α incoherent inverse covariance matrix Θ^*. Given a sample size $n > c_0(1+8\alpha^{-1})^2 m^2 log(d)$ with $\lambda_n = \frac{c_1}{\alpha}\sqrt{\frac{logd}{n}} + \delta$ for some $\delta in(0,1]$. Then with probability $1 - c_2 e^{-c_3 n\delta^2}$, we have

- $\hat{\Theta}_{jk} = 0 \; \forall (j,k) \notin E$
- $\|\hat{\Theta} - \Theta^*\|_{max} \leq c_4 \left\{ (1+8\alpha^{-1})\sqrt{\frac{logd}{n}} + \lambda_n \right\}$

Consequences of proposition 11.10

- The first part guarantees that $\hat{E} = \{(j,k) \in [d] \times [d], j < k, \hat{\Theta}_{jk} \neq 0\}$ is always a subset of true E.
- The second part guarantees that $\hat{\Theta}$ is close to Θ elementwise. Consequently, if we have a lower bound on the minimum non-zero entry of $|\Theta^*|$ —namely the quantity $\tau^*(\Theta^*) = min_{(j,k)\in E}|\Theta^*_{jk}|$, then graphical lasso recovers the full edge set.
- Proof involves the primal dual witness technique used in chapter 7 (Theorem 7.21) - **Not presented in today's lecture**

Operator Norm Bounds

An extension of proposition 11.10 also provides us with operator norm bounds.

Operator Norm Bound

Under the conditions of operator norm bounds, consider the graphical estimate $\hat{\Theta}$ with regularization parameter $\lambda_n = \frac{c_1}{n}\sqrt{\frac{logd}{n}} + \delta$ for some $\delta \in (0, 1]$. With probability $1 - c_2 e^{-c_3 n\delta^2}$, we have

$$\|\hat{\Theta} - \Theta\|_2 \leq c_4 \|A\|_2 \left\{(1 + 8\alpha^{-1})\sqrt{\frac{logd}{n}} + \lambda_n\right\} \tag{34}$$

where A is the adjacency matrix of the graph G (including the ones on the diagonal or self-edges)

Topics covered

- An introduction to the framework of graphical models
- Factorization, Conditional independence and the equivalence between the two definition (HC equivalence)
- Gaussian Graphical Models
- Frobenius Norm Bound - Proof using Corollary 9.20
- Useful results involving sup and operator norms preserving edge structure of G

Other topics in Chapter 11

There are other topics presented in Chapter 11 which are not presented in today's lecture, which are extensions or generalizations of the Gaussian graphical model. I recommend reading Chapter 11 in more detail if anyone is interested in the following topics.

- Neighborhood based methods - Gaussian Graphical Lasso is a **global** method, which estimates the full graph simultaneously. Neighborhood based methods are *local*, as it recovers the neighborhood N of each vertex $j \in V$.
- Non-Gaussian graphical model estimation (more general)
- Graphs with imperfect information - corrupted or hidden variables

Thanks!

I thank **Dr.Eric Slud** and **Dr.Vincet Lyzinski** for organizing the RIT and giving us the opportunity.

EOS