Structure estimation for discrete graphical models

Seong-Hwan Jun

June 19, 2015

Objective

- Denote the inverse of a covariance matrix by $\Gamma = \Sigma^{-1}$.
- For Gaussian case, $\Gamma_{st} = 0 \Rightarrow s \perp\!\!\!\perp t |$rest.
- For non-Gaussian case, the relationship is unresolved.
- This paper focusses on establishing number of links between covariance matrices and the edge structure of an underlying graph in the case of discrete-valued random variables.

Review I: Graph Theory Notion

- Graph: $G = (V, E)$, $V = \{1, \ldots, p\}, E \subseteq V \times V$.
- A vertex cutset: $U \subset V$ such that $V \setminus U = V_1 \bigcup V_2$ such that $V_1 \bigcap V_2 = \emptyset$ and $V_1, V_2 \neq \emptyset$.
- A clique: $C \subseteq V$ such that $(s,t) \in E$ for all $s, t, \in C$
- A clique C is maximal if there does not exist a clique C' such that $C \subset C'$.

Review II: Undirected Graphical Model

- Each node represent a variable, denoted X_s for each $s \in V$
- $X_s \in \mathscr{X}$ (e.g., $\mathscr{X} = \{0, 1\}$ for binary-valued case and $\mathscr{X} = \{0, 1, \ldots, m-1\}$ for multinomial-valued case)
- $X_A = \{X_s : s \in A\}$, $A \subseteq V$, refers to a set of random variables (or a vector of random variables) indexed by nodes $s \in A$
- Markov property: $X_A \perp\!\!\!\perp X_B | X_U$ whenever U is a vertext cutset of A and B.
- Factorization property: $p(x_1, \ldots, x_p) \propto \prod_{C \in \mathscr{C}} \psi_C(x_C)$ where $\mathscr{C}$ denotes the set of cliques and ψ is a positive-real valued function referred to as clique compatibility function.
- By Hammerseley-Clifford theorem, Markov property and factorization property are equivalent for any strictly positive distribution, e.g., exponential family.

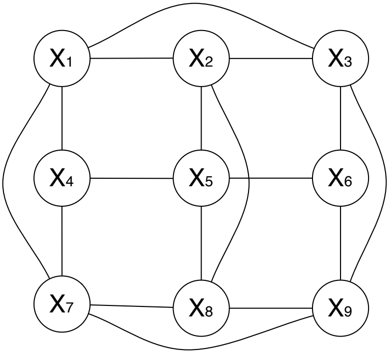
Exponential Family: Binary Variables

- For a binary random vector $X \in \{0,1\}^p$ and for each clique C, define a sufficient statistic $1_C(x_C) = \prod_{s \in C} x_s$.
- In other words, $1_C(x_C) = 1$ if all of $x_s = 1$ for $s \in C$ and 0 otherwise.
- Denote the natural parameters of the exponential familiy by $\theta_C \in \mathbb{R}$.
- The factorization property can be expressed as,

$$p_\theta(x_1, \ldots, x_p) = \exp\{ \sum_{C \in \mathscr{C}} \theta_C 1_C(x_C) - \Psi(\theta)\}$$

where $\Psi(\theta) = \log \sum_{x \in \{0,1\}^p} \exp(\sum_{C \in \mathscr{C}} \theta_C 1_C(x_C))$ is the normalization constant.

Example: Ising Model



$$p_\theta(x_1,\ldots,x_p) = \exp\{\sum_{s\in V}\theta_s x_s + \sum_{(s,t)\in E}\theta_{st}x_s x_t - \Phi(\theta)\}$$

- All two-non adjacent nodes are conditionally indpendent given the rest.

Exponential Family: Multinomial Variables

Denote $\mathscr{X}_0 = \mathscr{X} \setminus \{0\} = \{1, \ldots, m-1\}$ the sufficient statistics are defined as,

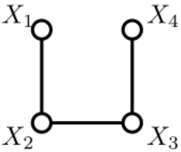
$$1_{C;J}(x_C) = \begin{cases} 1 & \text{if } x_C = J \\ 0 & \text{otherwise} \end{cases}$$

where $J \in \mathscr{X}_0^{|C|}$. The factorization is,

$$p_\theta(x_1, \ldots, x_p) = \exp\{\sum_{C \in \mathscr{C}} < \theta_C, 1_C > -\Phi(\theta)\}$$

where $< \theta_C, 1_C >= \sum_{J \in \mathscr{X}_0^{|C|}} \theta_{C;J} 1_{C;J}(x_C)$.

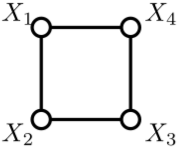
Motivation I: Chain Graph



$$\Gamma_{\text{chain}} = \begin{bmatrix} 9.80 & -3.59 & 0 & 0 \\ -3.59 & 34.30 & -4.77 & 0 \\ 0 & -4.77 & 34.30 & -3.59 \\ 0 & 0 & -3.59 & 9.80 \end{bmatrix}$$

- The Gaussian theory is that $\Gamma_{st} = 0$ if and only if $(s,t) \notin E$.
- The authors noted that it turns out that this is also the case for the chain graph with binary variables.
- The above example is computed using $\theta_s = 0.1$ for $s \in V$ and $\theta_{st} = 2$ for all $(s,t) \in E$.

Motivation II: 4 cycle



$$\Gamma_{\text{loop}} = \begin{bmatrix} 51.37 & -5.37 & -0.17 & -5.37 \\ -5.37 & 51.37 & -5.37 & -0.17 \\ -0.17 & -5.37 & 51.37 & -5.37 \\ -5.37 & -0.17 & -5.37 & 51.37 \end{bmatrix}$$

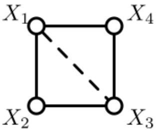
- There are no 0's in the inverse covariance matrix.
- So the Gaussian theory does not hold for this case.

Augmented Random Vector

- It turns out to be useful to consider the interaction between the variables.
- The augmented random vector refers to the random vector consisted of the variables X_s as well as the higher order interaction terms.
- Example: $(X_1, \ldots, X_4, X_{13})$ is an augmented random vector.

Generalized Covariance Matrix

- The authors refer to the generalized covariance matrix as the covariance matrix on the augmented random vector.
- The inverse of the covariance matrix of the augmented random vector $(X_1, \ldots, X_4, X_1X_3)$:



$$\Gamma_{\text{aug}} = 10^3 \times \begin{bmatrix} 1.15 & -0.02 & 1.09 & -0.02 & -1.14 \\ -0.02 & 0.05 & -0.02 & 0 & 0.01 \\ 1.09 & -0.02 & 1.14 & -0.02 & -1.14 \\ -0.02 & 0 & -0.02 & 0.05 & 0.01 \\ -1.14 & 0.01 & -1.14 & 0.01 & 1.19 \end{bmatrix}$$

Note: the entry corresponding to the edge $(2,4)$ is equal to 0!

Triangulation

- Chordless cycle: sequence of nodes, $\{s_1, \ldots, s_l\}$ such that
  - $(s_i, s_{i+1}) \in E$ for all $1 \leq i \leq l-1$ and $(s_l, s_1) \in E$.
  - no other nodes in the cycle are connected by an edge.
- Example: 4-cycle
- Triangulation: Given $G = (V, E)$, a triangulation is an augmented graph $\tilde{G} = (V, \tilde{E})$ that contains no chordless cycle of length greater than 3.
- Example: Tree is trivially triangulated.
- Example: Adding edge $(1, 3)$ is a triangulation of 4-cycle.

Main Result: Notation

Let $\mathscr{S} \subseteq \mathscr{C}$, define the random vector

$$\Psi(X;\mathscr{S}) = \{1_{C;J}, J \in \mathscr{X}_0^{|C|}, C \in \mathscr{S}\}$$

- Let $\Gamma = \text{cov}(\Psi(X;\tilde{\mathscr{C}}))^{-1}$ where $\tilde{\mathscr{C}}$ denotes the cliques of the triangulated graph.
- For any $A, B \in \tilde{\mathscr{C}}$, denote by $\Gamma(A,B)$ the sub-block of Γ indexed by all indicator statistics on A, B.
- Note: $\Gamma(A,B)$ has dimension $(m-1)^{|A|} \times (m-1)^{|B|}$.

Main Result: Theorem 1

Theorem

The generalized covariance matrix cov$(\Psi(X;\tilde{\mathscr{X}}))$ *is invertible and its inverse,* Γ *is block-graph structured:*

- *For any two subsets* $A, B \in \tilde{\mathscr{C}}$ *that are not subsets of the same maximal clique, the block* $\Gamma(A,B)$ *is identically zero.*
- *For almost all parameters* θ*, the entire block* $\Gamma(A,B)$ *is nonzero whenever* A *and* B *belong to a common maximal clique.*

- For example, if $A = \{s\}$, $B = \{t\}$ and if $(s,t) \notin E$ and that they do not belong to the same maximal clique (after triangulation), then $(m-1) \times (m-1)$ sub-block is all identically zero.
- This is consistent with the observation that $\Gamma_{24} = 0$ in the binary variables case for the 4-cycle graph after traingulation (adding edge $(1,3)$).
- Note: trees are already triangulated, meaning that $\Gamma(A,B) = 0$ implies that there is no edge connecting s and t in the underlying graph.

Separator Sets

- Triangulation of G gives rise to junction tree representation of G, where the nodes of the junction tree are the maximal cliques of $\tilde{G}$ and the intersection of any two adjacent cliques C_1, C_2 is referred to as a separator set, $S = C_1 \bigcap C_2$.
- Let $\mathscr{S}$ be a collection of separator sets.
- Define pow$(\mathscr{S}) := \bigcup_{S \in \mathscr{S}}$ pow(S), where pow stands for power set.

Corollary 1

Corollary

Let Γ *be the inverse of cov*$(\Psi(X; V \bigcup \text{pow}(\mathscr{S})))$. *Then,* $\Gamma(\{s\},\{t\}) = 0$ *whenever* $(s,t) \notin \tilde{E}$.

The consequence of this corollary is that it allows to reduce the length of the augmented random vector requried to recover graph structure as $V \bigcup \text{pow}(\mathscr{S}) \subseteq \tilde{\mathscr{C}}$ and is generally much smaller.

Example of Corollary 1

- The 4-cycle triangulated by adding the edge $(1,3)$ has two maximal cliques, $\{1,2,3\}$ and $\{1,3,4\}$. The separator set is $\{1,2,3\}\bigcap\{1,3,4\} = \{1,3\}$.
- Therefore, augmenting the random vector with the sufficient statistic $1_{13}(x_1, x_3) = x_1x_3$ and taking the inverse yields a graph-structured matrix.

$$\Gamma_{\text{aug}} = 10^3 \times \begin{bmatrix} 1.15 & -0.02 & 1.09 & -0.02 & -1.14 \\ -0.02 & 0.05 & -0.02 & 0 & 0.01 \\ 1.09 & -0.02 & 1.14 & -0.02 & -1.14 \\ -0.02 & 0 & -0.02 & 0.05 & 0.01 \\ -1.14 & 0.01 & -1.14 & 0.01 & 1.19 \end{bmatrix}$$

Singleton Set

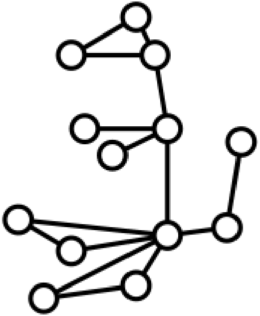
- One setting where Corollary 1 is particularly useful is when the separators sets are all singletons.
- A Singleton set is a set containing exactly one element.
- So in this case, $V \bigcup \text{pow}(\mathscr{S}) = V$.
- Example: This is true for trees. As we saw for the chain graph.

Corollary 2

Corollary

For any graph with singleton spearator sets, the inverse Γ of the covariance matrix cov$(\Psi(X;V))$ is graph-structured.

Example: Dino



Note: The separator sets are all singletons.

Corollary 3: Neighborhood selection

Some notation first:

- Define $N(s) := \{t \in V : (s,t) \in E\}$.
- Define $S(s;d) := \{U \subseteq V \setminus \{s\}, |U| = d\}$.

Corollary

For any node $s \in V$ with deg$(s) \leq d$, the inverse Γ of the matrix cov$(\Psi(X; \{s\} \bigcup \text{pow}(S(s;d))))$ is s-block graph structured. That is, $\Gamma(\{s\}, B) = 0$ whenever $\{s\} \neq B \nsubseteq N(s)$. In particular, $\Gamma(\{s\}, \{t\}) = 0$ for all vertices $t \notin N(s)$.

- Note that pow$(S(s;d))$ is set of all subsets of all possible neighborhoods of s with size $\leq d$.
- This corollary is useful recovering the neighborhood structure, $N(s)$ (but sadly I have not had the time to fully explore the algorithms developed in this paper yet).