Precision matrix estimation in Gaussian graphical models

Michał Makowski

Wydział Matematyki i Informatyki
Uniwersytet Wrocławski

michalmakowski@outlook.com

4 Października 2019 r.

Plan prezentacji

Modele graficzne

Teoria modeli graficznych uogólnia i opisuje szereg modeli statystycznych

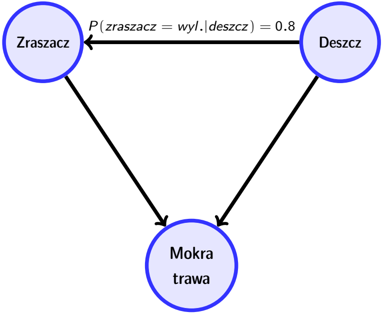
- łańcuchy Markowa/ukryte modele Markowa [*hidden Markov models*]
- sieci bayesowskie
- filtry Kalmana
- sieci neuronowe

Intuicja

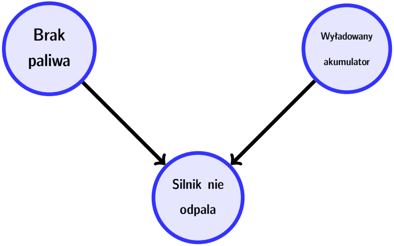
Modele graficzne łączą probabilistykę i teorię grafów

- probabilistyka - niepewność/losowość
- teoria grafów - zależność/korelacja

Graf skierowany



(a) Model 'trawnikowy'



(b) Model 'samochodowy'

Graf nieskierowany

Figures/GGM1.png

Niezależność

Dwie zmienne losowe X, Y są warunkowo niezależne pod warunkiem wektora losowego $\mathbf{Z} = (Z_1, \ldots, Z_n)^T$, wtedy i tylko wtedy, gdy ich rozkłady są niezależne pod warunkiem $\mathbf{Z}$. Relację taką oznaczamy poprzez $\perp\!\!\!\perp$
Formalnie:

$(\mathrm{X} \perp\!\!\!\perp Y) \mid \mathrm{Z} \quad \Longleftrightarrow \quad F_{X,Y|\mathbf{Z}=\mathbf{z}}(x,y) = F_{X\,|\,\mathbf{Z}\,=\,\mathbf{z}}(x) \cdot F_{Y\,|\,\mathbf{Z}\,=\,\mathbf{z}}(y)$
dla każdych x,y,z,

gdzie $F_{X,Y\,|\,\mathbf{Z}\,=\,\mathbf{z}}(x,y) = \Pr(X \leqslant x, Y \leqslant y \mid Z_1 = z_1, \ldots, Z_n = z_n)$ jest warunkową dystrybuantą X oraz Y przy zadanym $\mathbf{Z}$.

Graf

Graf

Grafem nazywamy parę zbiorów $G = (V, E)$, takich, że $E \subset [V]^2$.

Zatem elementami E są dwuelementowe podzbiory V. Aby uniknąć niejasności zawsze zakładamy, że $V \cap E = \emptyset$.
Zbiór wierzchołków grafu G oznaczamy przez $V(G)$, a jego zbiór krawędzi przez $E(G)$ (lub po prostu V oraz E odpowiednio).

Warunkowa niezależność, a graf

Okazuje się, że dla pewnej rodziny rozkładów prawdopodobieństwa można zbudować graf, który odzwierciedla własność warunkowej niezależności pomiędzy pojedynczymi zmiennymi. Zachodzi równoważność

$$\text{wierzchołek } A \text{ jest rozłączny z } B \quad \iff \quad X_A \perp\!\!\!\perp X_B \mid X_{-AB},$$

gdzie poprzez X_{-AB} rozumiemy wszystkie zmienne losowe poza X_A oraz X_B.

Parametryzacja rozkładu normalnego

Każdy wielowymiarowy rozkład normalny $\mathcal{N}(\mu, \boldsymbol{\Sigma})$ może zostać sprowadzony do parametryzacji kanonicznej zadanej przez

$$\gamma = \boldsymbol{\Sigma}^{-1}\mu \quad \text{oraz} \quad \boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1},$$

gdzie macierz Σ nazywamy *macierzą precyzji*.

Warunkowa niezależność, a macierz precyzji

Przedstawienie kanoniczne pozwala zobrazować własność warunkowej niezależności.

Niech $(X_1, \ldots, X_n) \sim \mathcal{N}(\gamma, \boldsymbol{\Theta})$, wtedy

$$X_s \perp\!\!\!\perp X_t \mid X_{-st} \iff \theta_{st} = 0,$$

gdzie poprzez X_{-st} rozumiemy wszystkie zmienne losowe poza X_j oraz X_t.

Graficzne Lasso

Wypukłą relaksacją wcześniejszego problemu jest

$$\mathbb{L}_\lambda(\boldsymbol{\Theta}, \mathbf{X}) = \log\det\boldsymbol{\Theta} - \operatorname{tr}(\boldsymbol{S}\,\boldsymbol{\Theta}) - \lambda\|\,\boldsymbol{\Theta}\,\|_1.$$

gdzie $\|\cdot\|_1$ oznacza normę ℓ_1 elementów macierzy powyżej przekątnej $\|A\|_1 = \sum_{i \geqslant j} |a_{ij}|$.

Graphical Lasso problem

$$\widehat{\boldsymbol{\Theta}} \in \operatorname*{arg\,max}_{\Theta \in S_+^p} \{\log\det\boldsymbol{\Theta} - \operatorname{tr}(\boldsymbol{S}\,\boldsymbol{\Theta}) - \lambda\|\,\boldsymbol{\Theta}\,\|_1\}.$$

Wielokrotne testowanie 1/3

Problem wielokrotnego testowania pojawia się, gdy wykonujemy równocześnie m testów statystycznych, a każdy z nich może dokonać potencjalnego odkrycia.

Wyniki oceniamy przy pomocy macierzy błędów

		Real value	
		(+)	(−)
Test outcome	(+)	True positive	False positive
	(−)	False negative	True negative

.

W naszym problemie podejmujemy decyzję o połączeniu, lub nie, dwóch wierzchołków estymowanego grafu.

Wielokrotne testowanie 2/3

Familywise error rate [FWER]

$$\text{FWER} = \mathbb{P}(\text{type I error})$$

Family-wise error rate (FWER) to prawdopodobieństwo popełnienia przynajmniej jednego fałszywego odkrycia.

False discovery rate (FDR)

$$\text{FDR} = \mathbb{E}\left[\frac{\#[\text{False positive}]}{\#[\text{False positive}] + \#[\text{True positive}]}\right]$$

Local false discovery rate (localFDR)

$$\text{localFDR} = \mathbb{E}\left[\frac{\#[\text{False positive outside the component}]}{\#[\text{False positive}] + \#[\text{True positive}]}\right]$$

Wielokrotne testowanie 3/3

Korekta Bonferonniego

Odrzucamy hipotezę zerową dla każdego testu, dla którego $p_i \leqslant \frac{\alpha}{m}$.
Korekta Bonferonniego kontroluje FWER na poziomie α, tj. FWER $\leqslant \alpha$.
Nie są wymagane żadne dodatkowe założenia

Metoda Holma

Porządkujemy p-wartości rosnąco, a następnie szukamy pierwszej hipotezy dla której zachodzi $p_{(k)} > \frac{\alpha}{m+1-k}$ i odrzucamy wszystkie wcześniejsze.
Metoda Holma kontroluje FWER na poziomie α.

Metoda Benjaminiego-Hochberga

Porządkujemy p-wartości rosnąco, a następnie szukamy ostatniej hipotezy dla której zachodzi $p_{(k)} \leqslant \alpha \frac{k}{m}$ i odrzucamy wszystkie wcześniejsze.
Metoda BH kontroluje FDR na poziomie α.

Wybór parametrów w gLasso

Banerjee - lambda dla graficznego Lasso

$$\lambda^{\text{Banerjee}}(\alpha) = \max_{i<j}(s_{ii}, s_{jj}) \frac{\text{qt}_{n-2}(1 - \frac{\alpha}{2p^2})}{\sqrt{n - 2 + \text{qt}^2_{n-2}(1 - \frac{\alpha}{2p^2})}} \tag{1}$$

Banerjee et al. w swojej pracy udowodniła następujące twierdzenie

Twierdzenie [BEd08]

Używając (1) jako parametru kary w problemie graficznego Lasso, dla każdego ustalonego α mamy

$$\mathbb{P}(\text{FWER}) \leqslant \alpha.$$

SLOPE

W pracy [Bog+15] Bogdan et al. zaproponowali nowe podejście do problemu regularyzacji. SLOPE używa normy *OL1* zamiast *L1* do wyboru współczynników w problemie regresji liniowej.

Norma *OL1*

Regularyzator oparty o posortowana normę ℓ_1 (znaną jako *OL1*, *OWL* lub *OSCAR*) dla $\beta \in \mathbb{R}^p$ oraz $\lambda \in \mathbb{R}^p, \lambda_1 \geqslant \ldots \geqslant \lambda_p$ zadany jest przez

$$J_\lambda(\beta) = \sum_{i=1}^{p} \lambda_i |\beta|_{(i)}.$$

Udowodniono, że pod pewnymi założeniami oraz przy konstrukcji ciągu λ opartego o procedure BH zaproponowana metoda kontroluje kontroluje FDR w modelu regresji wielorakiej.

gSLOPE

Zamieniając operator ℓ_1 na OL1 otrzymujemy graficzne SLOPE.

Problem graficznego SLOPE [gSLOPE]

$$\widehat{\mathbf{\Theta}} \in \underset{\Theta \in S_{+}^{p}}{\arg\max} \{\log\det \mathbf{\Theta} - \mathrm{tr}\left(\boldsymbol{S}\,\mathbf{\Theta}\right) - \mathrm{J}_{\lambda}(\mathbf{\Theta})\}$$

W pracy [Sob19] P. Sobczyk pokazał, że wykorzystanie OL1 w problemie estymacji macierzy precyzji daje obiecujące rezultaty w kontekście kontroli FDR.

Wybór parametru w gSLOPE (1/2)

Ciąg parametrów zbudowany w oparciu o procedurę Holma

$$m = \frac{p(p-1)}{2},$$

$$\lambda_k^{\text{Holm}} = \frac{\text{qt}_{n-2}(1 - \frac{\alpha}{m+1-k})}{\sqrt{n - 2 + \text{qt}_{n-2}^2(1 - \frac{\alpha}{m+1-k})}},$$

$$\lambda^{\text{Holm}} = \{\lambda_1^{\text{Holm}}, \lambda_2^{\text{Holm}}, ..., \lambda_m^{\text{Holm}}\}.$$

Wybór parametru w gSLOPE (2/2)

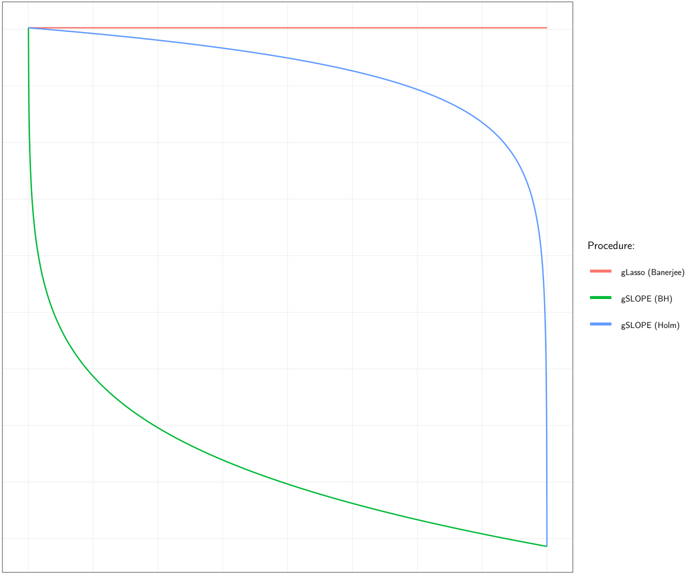
Ciąg parametrów zbudowany w oparciu o procedurę BH

$$m = \frac{p(p-1)}{2},$$

$$\lambda_k^{\mathrm{BH}} = \frac{\mathrm{qt}_{n-2}(1-\frac{\alpha k}{m})}{\sqrt{n-2+\mathrm{qt}_{n-2}^2(1-\frac{\alpha k}{m})}},$$

$$\lambda^{\mathrm{BH}} = \{\lambda_1^{\mathrm{BH}}, \lambda_2^{\mathrm{BH}}, ..., \lambda_m^{\mathrm{BH}}\}.$$

Porównanie ciągów lambda



Alternating direction method of multipliers

Do rozwiązania problemu gSLOPE posłużyliśmy się algorytmem *ADMM*, pozwala on rozwiązywać problemy optymalizacji wypukłej postaci

$$\begin{aligned} \text{minimum} \quad & f(x) + g(y) \\ \text{pod warunkiem} \quad & Ax + By = c. \end{aligned}$$

Rozszerzony operator Lagranga z parametrem $\rho > 0$ zdefinowany jest jako

$$\mathcal{L}_\rho(x, y, \nu) = f(x) + g(y) + \nu^T(Ax + By - c) + \frac{\rho}{2}\|Ax + By - b\|^2.$$

`Figures/ADMM.png`

ADMM dla gSLOPE

Dla graficznego SLOPE problem optymalizacyjny jest postaci

$$\begin{aligned} \text{minimum} \quad & -\log\det \boldsymbol{\Theta} + \operatorname{tr}(\boldsymbol{S}\,\boldsymbol{\Theta}) + \mathbb{I}[\boldsymbol{\Theta} \succeq 0] + J_\lambda(Y) \\ \text{pod warunkiem} \quad & Y = \boldsymbol{\Theta}\,. \end{aligned}$$

Rozszerzony operator Lagranga $\mathcal{L}_\rho : \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p} \to \mathbb{R}$ z parametrem $\rho > 0$ jest zadany przez

$$\mathcal{L}_\rho(X, Y, N) = -\log\det \boldsymbol{\Theta} + \operatorname{tr}(\boldsymbol{S}\,\boldsymbol{\Theta}) + \mathbb{I}[\boldsymbol{\Theta} \succeq 0] + J_\lambda(Y) +$$
$$\rho\langle N, \boldsymbol{\Theta} - Y\rangle_F + \frac{\rho}{2}\|\,\boldsymbol{\Theta} - Y\|_F^2.$$

Środowisko

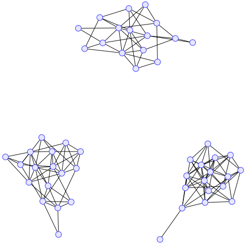
- Implementacja w R, pakiet **huge** do tworzenia grafów.

Środowisko

- Implementacja w R, pakiet **huge** do tworzenia grafów.
- Trzy rodzaje grafów:

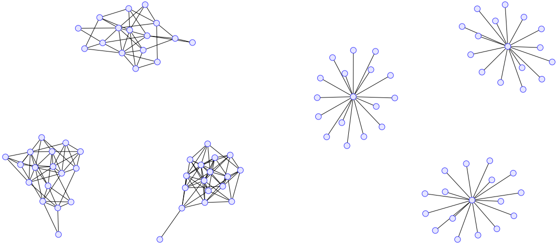
Środowisko

- Implementacja w R, pakiet **huge** do tworzenia grafów.
- Trzy rodzaje grafów: klastrowe



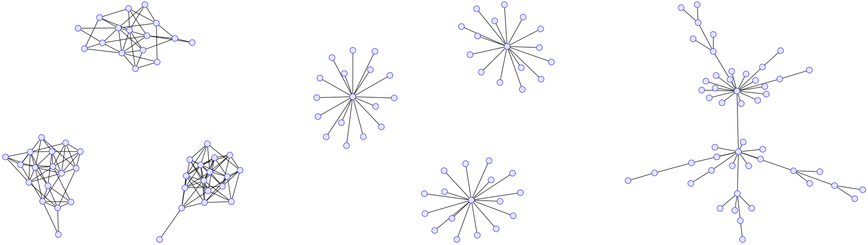
Środowisko

- Implementacja w R, pakiet **huge** do tworzenia grafów.
- Trzy rodzaje grafów: klastrowe, hub



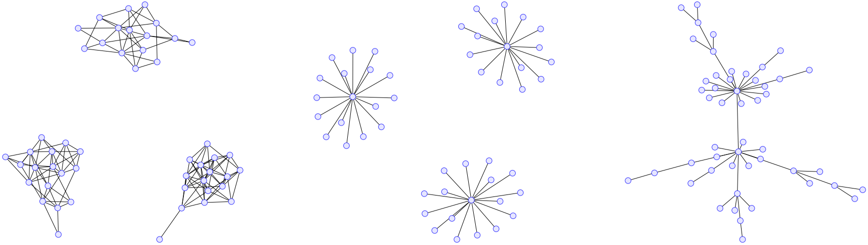
Środowisko

- Implementacja w R, pakiet **huge** do tworzenia grafów.
- Trzy rodzaje grafów: klastrowe, hub, and scale-free.



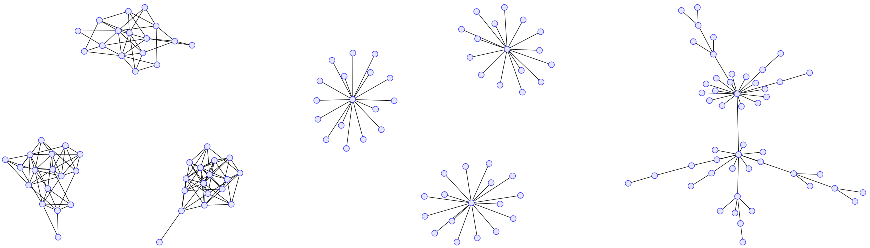
Środowisko

- Implementacja w R, pakiet **huge** do tworzenia grafów.
- Trzy rodzaje grafów: klastrowe, hub, and scale-free.
- Dane: $p = 100$, $n \in \{50, 100, 200, 400\}$; zmienny stosunek wartości poza przekątną do wartości na przekątnej, zmienna rzadkość grafu i zmienna wielkość składowych spójnych.



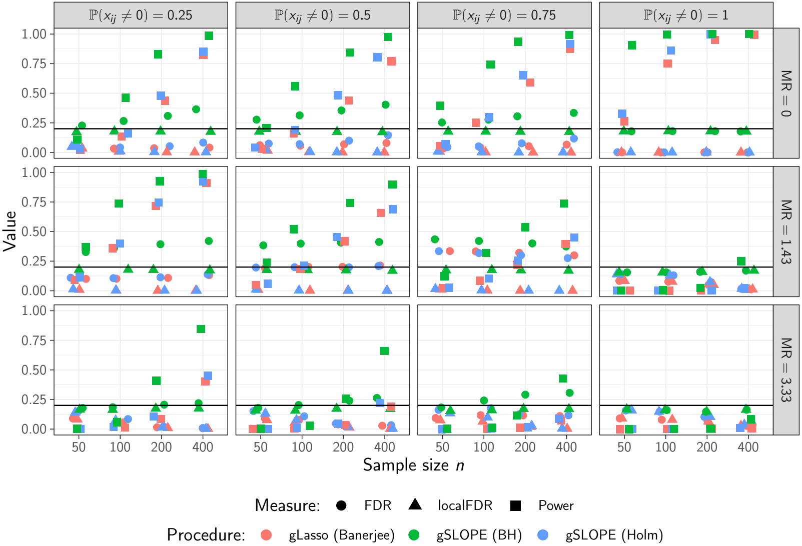
Środowisko

- Implementacja w R, pakiet **huge** do tworzenia grafów.
- Trzy rodzaje grafów: klastrowe, hub, and scale-free.
- Dane: $p = 100$, $n \in \{50, 100, 200, 400\}$; zmienny stosunek wartości poza przekątną do wartości na przekątnej, zmienna rzadkość grafu i zmienna wielkość składowych spójnych.
- Dwa poziomy pożądanej kontroli FDR: 0.05 and 0.2 .



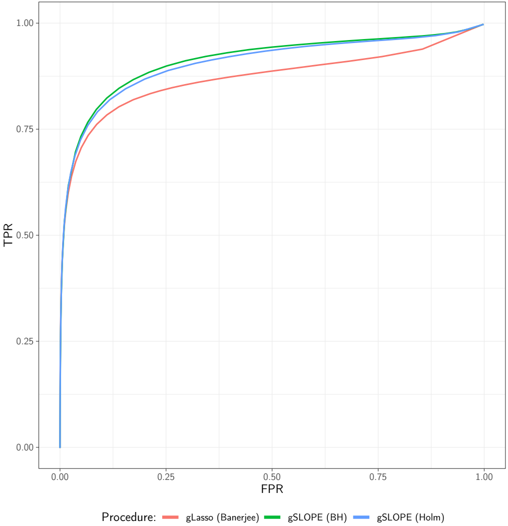
Wyniki dla grafów klastrowych

Setup: $\alpha = 0.2$, 100 variables, 10 components, cluster graph.

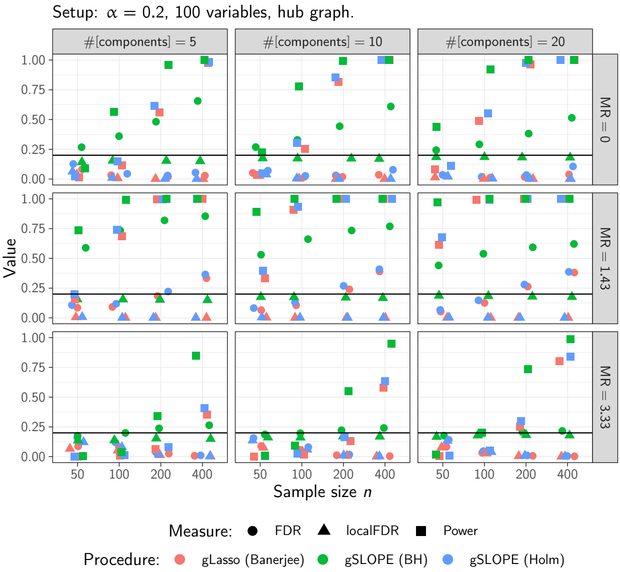


Krzywa ROC dla grafów klastrowych

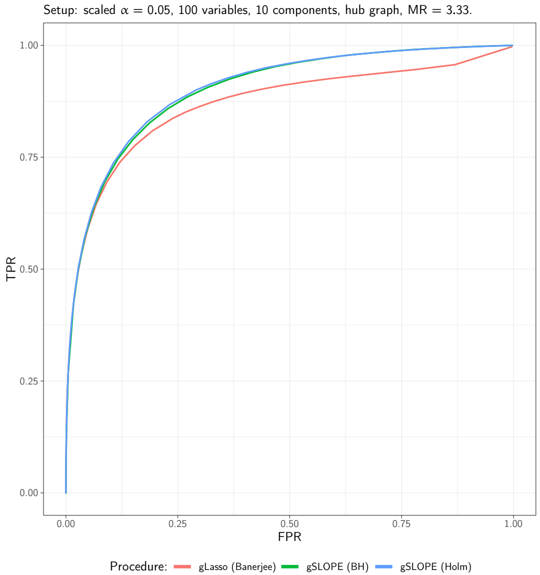
Setup: scaled $\alpha = 0.05$, 100 variables, 10 components, cluster graph, MR $= 0$, $\mathbb{P}(x_{ij} \neq 0) = 0.5$.



Wyniki dla grafów typu hub



Krzywe ROC dla grafów typu *hub*



Bibliografia

Onurena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. *Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data*. 2008.

Małgorzata Bogdan et al. *SLOPE - Adaptive variable selection via convex optimization*. 2015.

Stephen Boyd et al. *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*. 2010.

Emmanuel Candes. *Advanced Topics in Convex Optimization*.

Trevor Hastie et al. *Statistical Learning with Sparsity: The Lasso and Generalizations*. 2015.

Piotr Sobczyk. *Identifying low-dimensional structures through model selection in high-dimensional data*. 2019.

Pytania?

Dziękuję za uwagę.

Factorization theorem

Compatibility function

Let $G = (V, E)$ be a graph with a vertex set $V = 1, 2, \ldots, p$ and $\mathfrak{C}$ be its clique set. Let $\mathbb{X} = (X_1, \ldots, X_p)$ be a random vector defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, indexed by the graph nodes.

Definition (Compatibility function)

Let $C \in \mathfrak{C}$ be a clique of the graph G and let $\mathbb{X}_C$ be a subvector of the vector $\mathbb{X}$ indexed by the elements of the clique C, that is $\mathbb{X}_C = (X_s, s \in C)$. A real-valued function ψ_C of the vector $\mathbb{X}_C$ taking positive real values is called a *compatibility function*.

Factorization property

Definition (Factorization)

Let $C \in \mathfrak{C}$ be a clique of the graph G and let $\mathbb{X}_C$ be a subvector of the vector $\mathbb{X}$ indexed by the elements of the clique C, that is $\mathbb{X}_C = (X_s, s \in C)$. A real-valued function ψ_C of the vector $\mathbb{X}_C$ taking positive real values is called a *compatibility function*.

Given a collection of compatibility functions, we say that probability distribution $\mathbb{P}$ *factorizes over G* if it has decomposition

$$\mathbb{P}(x_1, \dots, x_n) = \frac{1}{Z} \prod_{C \in \mathfrak{C}} \psi_C(x_C), \tag{2}$$

where Z is the normalizing constant, known as the *partition function*. It is given by

$$Z = \sum_{x} \prod_{C \in \mathfrak{C}} \psi_C(x_C), \tag{3}$$

where the sum goes over all possible realizations of $\mathbb{X}$.

Markov property

Consider a cut set S of the given graph and let introduce a symbol $\perp\!\!\!\perp$ to denote the relation *is conditionally independent of*. With this notation, we say that the random vector $\mathbb{X}$ is Markov with respect to G if

$$\mathbb{X}_A \perp\!\!\!\perp \mathbb{X}_B \mid \mathbb{X}_S \qquad \text{for all cut sets } S \subset V, \tag{4}$$

where $\mathbb{X}_A$ denotes the subvector indexed by the subgraph A.

Canonical formulation

Canonical formulation

Any nondegenerated multivariate normal distribution $\mathcal{N}(\mu, \mathbf{\Sigma})$ can reparametrized into canonical parameters of the form

$$\gamma = \mathbf{\Sigma}^{-1}\mu \quad \text{and} \quad \mathbf{\Theta} = \mathbf{\Sigma}^{-1}.$$

Then density function is given by

$$\mathbb{P}_{\gamma,\mathbf{\Theta}}(x) = \exp\left\{\sum_{s=1}^{p} \gamma_s x_s - \frac{1}{2}\sum_{s,t=1}^{p} \theta_{st} x_s x_t - A(\gamma, \mathbf{\Theta})\right\},$$

where $A(\gamma, \mathbf{\Theta}) = -\frac{1}{2}\left(\det[(2\pi)^{-1}\,\mathbf{\Theta}] + \gamma^T\,\mathbf{\Theta}^{-1}\,\gamma\right)$.

Canonical formula derivation

$$\mathbb{P}_{\mu,\boldsymbol{\Sigma}}(x) = \left(\sqrt{\det[2\pi\boldsymbol{\Sigma}]}\right)^{-1} \exp\left\{(-\frac{1}{2}(x-\mu)^T\boldsymbol{\Sigma}^{-1}(x-\mu)\right\}$$

Canonical formula derivation

$$
\begin{aligned}
\mathbb{P}_{\mu,\boldsymbol{\Sigma}}(x) &= \left(\sqrt{\det[2\pi\boldsymbol{\Sigma}]}\right)^{-1} \exp\left\{(-\frac{1}{2}(x-\mu)^T\boldsymbol{\Sigma}^{-1}(x-\mu)\right\} \\
&= \left(\sqrt{\det[(2\pi\boldsymbol{\Sigma})^{-1}]}\right) \exp\left\{-\frac{1}{2}x^T\boldsymbol{\Sigma}^{-1}x + x^T\boldsymbol{\Sigma}^{-1}\mu - \frac{1}{2}\mu^T\boldsymbol{\Sigma}^{-1}\mu\right\}
\end{aligned}
$$

Canonical formula derivation

$$\begin{aligned}
\mathbb{P}_{\mu,\boldsymbol{\Sigma}}(x) &= \left(\sqrt{\det[2\pi\boldsymbol{\Sigma}]}\right)^{-1} \exp\left\{(-\frac{1}{2}(x-\mu)^T\boldsymbol{\Sigma}^{-1}(x-\mu)\right\} \\
&= \left(\sqrt{\det[(2\pi\boldsymbol{\Sigma})^{-1}]}\right) \exp\left\{-\frac{1}{2}x^T\boldsymbol{\Sigma}^{-1}x + x^T\boldsymbol{\Sigma}^{-1}\mu - \frac{1}{2}\mu^T\boldsymbol{\Sigma}^{-1}\mu\right\} \\
&= \left(\sqrt{\det[(2\pi)^{-1}\,\boldsymbol{\Theta}]}\right)^{-1} \exp\left\{-\frac{1}{2}x^T\,\boldsymbol{\Theta}\,x + x^T\gamma - \frac{1}{2}\gamma^T\,\boldsymbol{\Theta}^{-1}\,\gamma\right\}
\end{aligned}$$

Canonical formula derivation

$$
\begin{aligned}
\mathbb{P}_{\mu,\boldsymbol{\Sigma}}(x) &= \left(\sqrt{\det[2\pi\boldsymbol{\Sigma}]}\right)^{-1} \exp\left\{(-\frac{1}{2}(x-\mu)^T\boldsymbol{\Sigma}^{-1}(x-\mu)\right\} \\
&= \left(\sqrt{\det[(2\pi\boldsymbol{\Sigma})^{-1}]}\right) \exp\left\{-\frac{1}{2}x^T\boldsymbol{\Sigma}^{-1}x + x^T\boldsymbol{\Sigma}^{-1}\mu - \frac{1}{2}\mu^T\boldsymbol{\Sigma}^{-1}\mu\right\} \\
&= \left(\sqrt{\det[(2\pi)^{-1}\,\boldsymbol{\Theta}]}\right)^{-1} \exp\left\{-\frac{1}{2}x^T\,\boldsymbol{\Theta}\,x + x^T\gamma - \frac{1}{2}\gamma^T\,\boldsymbol{\Theta}^{-1}\,\gamma\right\} \\
&= \exp\left\{-\frac{1}{2}x^T\,\boldsymbol{\Theta}\,x + x^T\gamma - \frac{1}{2}\left(\det[(2\pi)^{-1}\,\boldsymbol{\Theta}] + \gamma^T\,\boldsymbol{\Theta}^{-1}\,\gamma\right)\right\}
\end{aligned}
$$

Canonical formula derivation

$$
\begin{aligned}
\mathbb{P}_{\mu,\boldsymbol{\Sigma}}(x) &= \left(\sqrt{\det[2\pi\boldsymbol{\Sigma}]}\right)^{-1} \exp\left\{(-\frac{1}{2}(x-\mu)^T\boldsymbol{\Sigma}^{-1}(x-\mu)\right\} \\
&= \left(\sqrt{\det[(2\pi\boldsymbol{\Sigma})^{-1}]}\right) \exp\left\{-\frac{1}{2}x^T\boldsymbol{\Sigma}^{-1}x + x^T\boldsymbol{\Sigma}^{-1}\mu - \frac{1}{2}\mu^T\boldsymbol{\Sigma}^{-1}\mu\right\} \\
&= \left(\sqrt{\det[(2\pi)^{-1}\,\boldsymbol{\Theta}]}\right)^{-1} \exp\left\{-\frac{1}{2}x^T\,\boldsymbol{\Theta}\,x + x^T\gamma - \frac{1}{2}\gamma^T\,\boldsymbol{\Theta}^{-1}\,\gamma\right\} \\
&= \exp\left\{-\frac{1}{2}x^T\,\boldsymbol{\Theta}\,x + x^T\gamma - \frac{1}{2}\left(\det[(2\pi)^{-1}\,\boldsymbol{\Theta}] + \gamma^T\,\boldsymbol{\Theta}^{-1}\,\gamma\right)\right\} \\
&= \exp\left\{-\frac{1}{2}x^T\,\boldsymbol{\Theta}\,x + x^T\gamma - A(\gamma,\boldsymbol{\Theta})\right\}
\end{aligned}
$$

Canonical formula derivation

$$
\begin{aligned}
\mathbb{P}_{\mu,\boldsymbol{\Sigma}}(x) &= \left(\sqrt{\det[2\pi\boldsymbol{\Sigma}]}\right)^{-1} \exp\left\{(-\frac{1}{2}(x-\mu)^T\boldsymbol{\Sigma}^{-1}(x-\mu)\right\} \\
&= \left(\sqrt{\det[(2\pi\boldsymbol{\Sigma})^{-1}]}\right) \exp\left\{-\frac{1}{2}x^T\boldsymbol{\Sigma}^{-1}x + x^T\boldsymbol{\Sigma}^{-1}\mu - \frac{1}{2}\mu^T\boldsymbol{\Sigma}^{-1}\mu\right\} \\
&= \left(\sqrt{\det[(2\pi)^{-1}\,\boldsymbol{\Theta}]}\right)^{-1} \exp\left\{-\frac{1}{2}x^T\,\boldsymbol{\Theta}\,x + x^T\gamma - \frac{1}{2}\gamma^T\,\boldsymbol{\Theta}^{-1}\,\gamma\right\} \\
&= \exp\left\{-\frac{1}{2}x^T\,\boldsymbol{\Theta}\,x + x^T\gamma - \frac{1}{2}\left(\det[(2\pi)^{-1}\,\boldsymbol{\Theta}] + \gamma^T\,\boldsymbol{\Theta}^{-1}\,\gamma\right)\right\} \\
&= \exp\left\{-\frac{1}{2}x^T\,\boldsymbol{\Theta}\,x + x^T\gamma - A(\gamma,\boldsymbol{\Theta})\right\} \\
&= \mathbb{P}_{\gamma,\boldsymbol{\Theta}}(x)
\end{aligned}
$$

Log-likelihood derivation

Log-likelihood derivation (1/2)

$$\mathbb{L}(\boldsymbol{\Theta}, \mathbf{X}) = \frac{1}{N} \sum_{i=1}^{N} \log \mathbb{P}_{\boldsymbol{\Theta}}(x_i)$$

Log-likelihood derivation (1/2)

$$\begin{aligned}\mathbb{L}(\boldsymbol{\Theta}, \mathbf{X}) &= \frac{1}{N}\sum_{i=1}^{N} \log \mathbb{P}_{\boldsymbol{\Theta}}(x_i) \\ &= \frac{1}{N}\sum_{i=1}^{N} -\frac{1}{2} x_i^T \, \boldsymbol{\Theta} \, x_i - A(\boldsymbol{\Theta})\end{aligned}$$

Log-likelihood derivation (1/2)

$$\begin{aligned}
\mathbb{L}(\mathbf{\Theta}, \mathbf{X}) &= \frac{1}{N} \sum_{i=1}^{N} \log \mathbb{P}_{\mathbf{\Theta}}(x_i) \\
&= \frac{1}{N} \sum_{i=1}^{N} -\frac{1}{2} x_i^T \, \mathbf{\Theta} \, x_i - A(\mathbf{\Theta}) \\
&= \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \log \det[(2\pi)^{-1} \, \mathbf{\Theta}] - \frac{1}{2} x_i^T \, \mathbf{\Theta} \, x_i
\end{aligned}$$

Log-likelihood derivation (1/2)

$$
\begin{aligned}
\mathbb{L}(\boldsymbol{\Theta}, \mathbf{X}) &= \frac{1}{N} \sum_{i=1}^{N} \log \mathbb{P}_{\boldsymbol{\Theta}}(x_i) \\
&= \frac{1}{N} \sum_{i=1}^{N} -\frac{1}{2} x_i^T \, \boldsymbol{\Theta} \, x_i - A(\boldsymbol{\Theta}) \\
&= \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \log \det[(2\pi)^{-1} \, \boldsymbol{\Theta}] - \frac{1}{2} x_i^T \, \boldsymbol{\Theta} \, x_i \\
&= \frac{1}{2N} \sum_{i=1}^{N} \log \left((2\pi)^{-N} \det[\boldsymbol{\Theta}] \right) - x_i^T \, \boldsymbol{\Theta} \, x_i
\end{aligned}
$$

Log-likelihood derivation (1/2)

$$
\begin{aligned}
\mathbb{L}(\boldsymbol{\Theta}, \mathbf{X}) &= \frac{1}{N} \sum_{i=1}^{N} \log \mathbb{P}_{\boldsymbol{\Theta}}(x_i) \\
&= \frac{1}{N} \sum_{i=1}^{N} -\frac{1}{2} x_i^T \, \boldsymbol{\Theta} \, x_i - A(\boldsymbol{\Theta}) \\
&= \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \log \det[(2\pi)^{-1} \, \boldsymbol{\Theta}] - \frac{1}{2} x_i^T \, \boldsymbol{\Theta} \, x_i \\
&= \frac{1}{2N} \sum_{i=1}^{N} \log \left((2\pi)^{-N} \det[\boldsymbol{\Theta}] \right) - x_i^T \, \boldsymbol{\Theta} \, x_i \\
&= \frac{1}{2N} \sum_{i=1}^{N} \log \det \boldsymbol{\Theta} - N \log 2\pi - x_i^T \, \boldsymbol{\Theta} \, x_i = \dots
\end{aligned}
$$

Log-likelihood derivation (2/2)

$$\ldots = \frac{1}{2N} \sum_{i=1}^{N} \log \det \mathbf{\Theta} - N \log 2\pi - x_i^T \, \mathbf{\Theta} \, x_i$$

Log-likelihood derivation (2/2)

$$\ldots = \frac{1}{2N} \sum_{i=1}^{N} \log\det \boldsymbol{\Theta} - N \log 2\pi - x_i^T \, \boldsymbol{\Theta} \, x_i$$

$$= \frac{1}{2N} \sum_{i=1}^{N} \log\det \boldsymbol{\Theta} - N \log 2\pi - \operatorname{tr}\left(x_i^T \, \boldsymbol{\Theta} \, x_i\right)$$

Log-likelihood derivation (2/2)

$$
\begin{aligned}
\ldots &= \frac{1}{2N} \sum_{i=1}^{N} \log \det \mathbf{\Theta} - N \log 2\pi - x_i^T \, \mathbf{\Theta} \, x_i \\
&= \frac{1}{2N} \sum_{i=1}^{N} \log \det \mathbf{\Theta} - N \log 2\pi - \operatorname{tr}\left(x_i^T \, \mathbf{\Theta} \, x_i\right) \\
&= \frac{1}{2} \log \det \mathbf{\Theta} - \frac{N}{2} \log 2\pi - \frac{1}{2N} \sum_{i=1}^{N} \operatorname{tr}\left(x_i x_i^T \, \mathbf{\Theta}\right)
\end{aligned}
$$

Log-likelihood derivation (2/2)

$$
\begin{aligned}
\dots &= \frac{1}{2N} \sum_{i=1}^{N} \log \det \boldsymbol{\Theta} - N \log 2\pi - x_i^T \, \boldsymbol{\Theta} \, x_i \\
&= \frac{1}{2N} \sum_{i=1}^{N} \log \det \boldsymbol{\Theta} - N \log 2\pi - \operatorname{tr}\left(x_i^T \, \boldsymbol{\Theta} \, x_i \right) \\
&= \frac{1}{2} \log \det \boldsymbol{\Theta} - \frac{N}{2} \log 2\pi - \frac{1}{2N} \sum_{i=1}^{N} \operatorname{tr}\left(x_i x_i^T \, \boldsymbol{\Theta} \right) \\
&= \frac{1}{2} \log \det \boldsymbol{\Theta} - \frac{N}{2} \log 2\pi - \frac{1}{2} \operatorname{tr}\left(\boldsymbol{S} \, \boldsymbol{\Theta} \right),
\end{aligned}
$$

where $\boldsymbol{S}$ is an empirical covariance matrix given by $\frac{1}{N} \sum_{i=1}^{N} x_i x_i^T$.

ADMM for Graphical SLOPE

ADMM for Graphical SLOPE

Graphical SLOPE problem - ADMM formulation

$$\begin{aligned} &\text{minimize} && -\log\det X + \operatorname{tr}(XS) + \mathbb{I}[X \succeq 0] + J_\lambda(Y) \\ &\text{subject to} && X = Y. \end{aligned}$$

ADMM for Graphical SLOPE

Graphical SLOPE problem - ADMM formulation

$$\begin{aligned} \text{minimize} \quad & -\log\det X + \operatorname{tr}(XS) + \mathbb{I}[X \succeq 0] + J_\lambda(Y) \\ \text{subject to} \quad & X = Y. \end{aligned}$$

Graphical SLOPE problem - Augmented Lagrangian

$$\begin{aligned} \mathcal{L}_\rho(X, Y, N) = & -\log\det X + \operatorname{tr}(XS) + \mathbb{I}[X \succeq 0] \\ & + \lambda\|Y\|_1 + \rho\langle N, X - Y\rangle_F + \frac{\rho}{2}\|X - Y\|_F^2 \end{aligned}$$

X-update (1/3)

We have

$$X_k = \arg\min_{X} \mathcal{L}_\rho(X, Y_{k-1}, N_{k-1}) = \arg\min_{X \succeq 0} \left\{ -\log\det X + \frac{\rho}{2} \left\| X - \tilde{S}_{k-1} \right\|_F^2 \right\},$$

where

$$\tilde{S}_{k-1} = -N_{k-1} + Y_{k-1} - \frac{1}{\rho} S,$$

X-update (1/3)

We have

$$X_k = \arg\min_X \mathcal{L}_\rho(X, Y_{k-1}, N_{k-1}) = \arg\min_{X \succeq 0} \left\{ -\log\det X + \frac{\rho}{2} \left\| X - \tilde{S}_{k-1} \right\|_F^2 \right\},$$

where

$$\tilde{S}_{k-1} = -N_{k-1} + Y_{k-1} - \frac{1}{\rho} S,$$

The X-gradient of the augmented Lagrangian is given by

$$\nabla_X \mathcal{L}_\rho(X, Y_{k-1}, N_{k-1}) = -X^{-1} + \rho X - \rho \tilde{S}_{k-1}.$$

X-update (1/3)

We have

$$X_k = \arg\min_X \mathcal{L}_\rho(X, Y_{k-1}, N_{k-1}) = \arg\min_{X \succeq 0} \left\{ -\log\det X + \frac{\rho}{2} \left\| X - \tilde{S}_{k-1} \right\|_F^2 \right\},$$

where

$$\tilde{S}_{k-1} = -N_{k-1} + Y_{k-1} - \frac{1}{\rho} S,$$

The X-gradient of the augmented Lagrangian is given by

$$\nabla_X \mathcal{L}_\rho(X, Y_{k-1}, N_{k-1}) = -X^{-1} + \rho X - \rho \tilde{S}_{k-1}.$$

As the augmented Lagrangian is convex, it is clear that for some $X^* \succeq 0$

$$\nabla_X \mathcal{L}_\rho(X^*, Y_{k-1}, N_{k-1}) = -(X^*)^{-1} + \rho X^* - \rho \tilde{S}_{k-1} = 0.$$

X-update (2/3)

Rewriting equation as

$$-(X^*)^{-1} + \rho X^* = \rho \tilde{S}_{k-1},$$

we can find a matrix that meets this condition.

X-update (2/3)

Rewriting equation as

$$-(X^*)^{-1} + \rho X^* = \rho \tilde{S}_{k-1},$$

we can find a matrix that meets this condition.

At first, lets take the eigenvalue decomposition of right side

$$\rho \tilde{S}_{k-1} = \rho Q \Lambda Q^T.$$

X-update (2/3)

Rewriting equation as

$$-(X^*)^{-1} + \rho X^* = \rho \tilde{S}_{k-1},$$

we can find a matrix that meets this condition.

At first, lets take the eigenvalue decomposition of right side

$$\rho \tilde{S}_{k-1} = \rho Q \Lambda Q^T.$$

Then by multiplying right and left side by Q and Q^T respectively, we obtain

$$-(\tilde{X}^*)^{-1} + \rho \tilde{X}^* = \rho \Lambda,$$

where $\tilde{X}^* = Q^T X^* Q$.

X-update (3/3)

We have to find positive numbers $\tilde{x}_{ii}^*$ that satisfy

$$(\tilde{x}_{ii}^*)^2 - l_{ii}\tilde{x}_{ii}^* - \frac{1}{\rho} = 0.$$

It is obvious that

$$\tilde{x}_{ii} = \frac{l_i + \sqrt{l_i^2 + 4/\rho}}{2}.$$

Thus X^* is given by $X^* = Q^T \tilde{X}^* Q$. All diagonals are positive since $\rho > 0$.
Define $\mathcal{F}_\rho(\Lambda)$ as

$$\mathcal{F}_\rho(\Lambda) = \frac{1}{2}\,\mathrm{diag}\left\{ l_i + \sqrt{l_i^2 + 4/\rho} \right\}.$$

Since that

$$X^* = Q^T \tilde{X}^* Q = Q^T \mathcal{F}_\rho(\Lambda) Q = \mathcal{F}_\rho(\tilde{S}_{k-1}) = \mathcal{F}_\rho\left(-N_{k-1} + Y_{k-1} - \frac{1}{\rho} S \right),$$

we obtain a formula for updating X_k in each step.

Y-update

A formula for Y_k is different. We have

$$\begin{aligned} Y_k &= \arg\min_{Y} \mathcal{L}_\rho(X_k, Y, N_{k-1}) \\ &= \arg\min_{Y} \left\{ \mathrm{J}_\lambda(Y) + \frac{\rho}{2}\| Y - (X_k + N_{k-1})\|_F^2 \right\} \end{aligned}$$

Y-update

A formula for Y_k is different. We have

$$\begin{aligned} Y_k &= \arg\min_Y \mathcal{L}_\rho(X_k, Y, N_{k-1}) \\ &= \arg\min_Y \left\{ \mathrm{J}_\lambda(Y) + \frac{\rho}{2}\|Y - (X_k + N_{k-1})\|_F^2 \right\} \end{aligned}$$

The last line of Y-update can be represented as a **proximity operator** which has closed form formula for SLOPE

$$\arg\min_Y \left\{ \mathrm{J}_\lambda(Y) + \frac{\rho}{2}\|Y - (X_k + N_{k-1})\|_F^2 \right\} = \mathbf{prox}_{\mathrm{J}_\lambda,\rho}\left(X_k + N_{k-1}\right). \quad (5)$$

ADMM for Graphical SLOPE

Figures/ADMMgSLOPE.png