CS-E5885 Modeling biological networks
Undirected graphical models and graphical lasso

Harri Lähdesmäki

Department of Computer Science
Aalto University

February 16, 2021

Outline

- Undirected graphical models
- Parametrization
- Gaussian graphical models
- Graphical lasso
- Applications to learn biological networks

- We will follow (loosely) selected subsections of Chapters 19 and 26 from (Murphy, 2012)

Introduction

- Previously we considered directed graphical models which are commonly known as Bayesian networks (BN)
- For some applications, the directionality may be meaningless or difficult to interpret
- An alternative is to consider undirected graphical models (UGM) which are also known as Markov random fields (MRF)

Introduction

- Previously we considered directed graphical models which are commonly known as Bayesian networks (BN)
- For some applications, the directionality may be meaningless or difficult to interpret
- An alternative is to consider undirected graphical models (UGM) which are also known as Markov random fields (MRF)
- Some advantages of MRF over BNs
  - Symmetric and thus more natural for some biological applications
- Some disadvantages of MRF compared to BNs
  - Parameters are less interpretable
  - Underlying graphical model structure does not have as intuitive causal interpretation as directed model has
  - Model inference can be more computationally expensive

Undirected graphs

- An undirected graph $G = (V, E)$ consists of
  - A set of nodes $V = (x_1, \ldots, x_n)$
  - A set of undirected edges $E = \{(s, t) : x_s, x_t \in V\}$
- For undirected graphs, $(s, t) \in E$ iff $(t, s) \in E$

Undirected graphs

- An undirected graph $G = (V, E)$ consists of
  - A set of nodes $V = (x_1, \ldots, x_n)$
  - A set of undirected edges $E = \{(s, t) : x_s, x_t \in V\}$
- For undirected graphs, $(s, t) \in E$ iff $(t, s) \in E$
- Path $s \to t$ is a sequence of undirected edges leading from s to t
- Clique: For an undirected graph a clique is a set of nodes that are all neighbors of each other.
- Maximal clique is a clique which cannot be made any larger without losing the clique propert

Example of an undirected graph

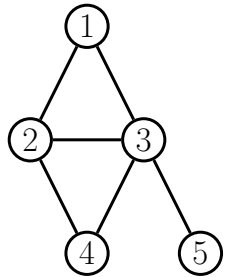


Figure: Figure from (Murphy, 2012)

- $(1, 2)$ is a path, $(1, 2, 3, 5)$ is another path
- $\{1, 2\}$ is a clique, $\{2, 3\}$ is another clique
- Maximal cliques are: $\{1, 2, 3\}$, $\{2, 3, 4\}$ and $\{3, 5\}$

Conditional independence properties of UGMs

- Let us now consider the nodes of an undirected graph as random variables
- For UGMs, the conditional independence (CI) relationships are defined via a simple graph separation
- For sets of nodes $\mathbf{x}_A$, $\mathbf{x}_B$ and $\mathbf{x}_C$, where $A, B, C \subset \{1, \ldots, n\}$ we say

$$\mathbf{x}_A \perp \mathbf{x}_B | \mathbf{x}_C$$

  iff C separates A from B in the graph G
- In terms of the underlying graph structure, this means that when we remove all nodes in C then there are no paths connecting any node in A to any node in B
- This is called the global Markov property

Conditional independence properties of UGMs

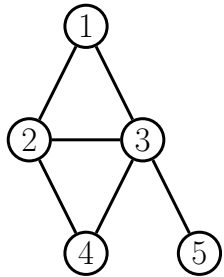


Figure: Figure from (Murphy, 2012)

- $\{1\} \perp \{5\} | \{3\}$
- $\{1\} \perp \{4\} | \{2, 3\}$

Parametrization of MRFs

- Let $\mathbf{y}$ denote random variables corresponding to the nodes of an undirected graph
- The joint distribution is $p(\mathbf{y})$
- How is the underlying UGM related to the joint distribution for $\mathbf{y}$?

Parametrization of MRFs

- Let $\mathbf{y}$ denote random variables corresponding to the nodes of an undirected graph
- The joint distribution is $p(\mathbf{y})$
- How is the underlying UGM related to the joint distribution for $\mathbf{y}$?
- Nodes do not have any topological ordering as they do with BNs, so the chain rule or conditional probability distributions cannot be used
- Instead, we associate a potential function with each maximal clique
- Potential function can be any non-negative function
- The joint distribution is then proportional to the product of potential functions of all maximal cliques

Hammersley-Clifford theorem[1]

- A positive distribution $p(\mathbf{y}) > 0$ satisfies the CI properties of an undirected graph G iff $p(\mathbf{y})$ can be represented as a product of potentials, one for each maximal clique

$$p(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \prod_{C \in \mathcal{C}} \psi_C(\mathbf{y}_C|\boldsymbol{\theta}_C),$$

where $\mathcal{C}$ contains all maximal cliques of G and $Z(\boldsymbol{\theta})$ is the partion function (here for discrete-valued $\mathbf{y}$)

$$Z(\boldsymbol{\theta}) = \sum_{\mathbf{y}} \prod_{C \in \mathcal{C}} \psi_C(\mathbf{y}_C|\boldsymbol{\theta}_C),$$

which ensures that the distribution sums to 1

[1] Theorem 19.3.1 in (Murphy, 2012)

An example of probability factorization

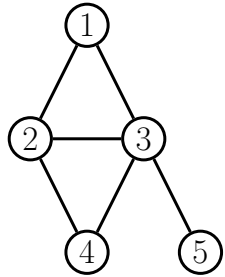


Figure: Figure from (Murphy, 2012)

- If p satisfies the CI properties of this graph, then the joint distribution factorizes as

$$p(\mathbf{y}|\boldsymbol{\theta}) \propto \psi_{1,2,3}(y_1, y_2, y_3|\boldsymbol{\theta}_{1,2,3})\psi_{2,3,4}(y_2, y_3, y_4|\boldsymbol{\theta}_{2,3,4}) \times \psi_{3,5}(y_3, y_5|\boldsymbol{\theta}_{3,5})$$

Pair-wise MRFs

- We can freely choose the parametrization for the potential and partition functions
- E.g. widely used is a pairwise MRF where the parametrization is restricted to the edges of the graph (instead of maximal cliques)
- Pair-wise MRF for the undirected graph on the previous slide

$$
\begin{aligned}
p(\mathbf{y}|\boldsymbol{\theta}) &\propto \psi_{1,2}(y_1, y_2)\psi_{1,3}(y_1, y_3)\psi_{2,3}(y_2, y_3) \\
&\quad \times \psi_{2,4}(y_2, y_4)\psi_{3,4}(y_3, y_4)\psi_{3,5}(y_3, y_5) \\
&\propto \prod_{s,t \in E} \psi_{s,t}(y_s, y_t)
\end{aligned}
$$

Pair-wise MRFs

- We can freely choose the parametrization for the potential and partition functions
- E.g. widely used is a pairwise MRF where the parametrization is restricted to the edges of the graph (instead of maximal cliques)
- Pair-wise MRF for the undirected graph on the previous slide

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\theta}) &\propto \psi_{1,2}(y_1, y_2)\psi_{1,3}(y_1, y_3)\psi_{2,3}(y_2, y_3) \\ &\quad \times \psi_{2,4}(y_2, y_4)\psi_{3,4}(y_3, y_4)\psi_{3,5}(y_3, y_5) \\ &\propto \prod_{s,t \in E} \psi_{s,t}(y_s, y_t) \end{aligned}$$

- All node pairs included in each maximal clique form the clique-wise potentials (some node pairs can be in several cliques)
- We can also include terms corresponding to individual nodes, $\psi_t(y_t)$

Gaussian MRFs

- An undirected graphical Gaussian model, also called Gaussian Markov random field, is a pair-wise MRF that has the following form

$$
\begin{aligned}
p(\mathbf{y}|\boldsymbol{\theta}) &\propto \prod_{s,t\in E} \psi_{s,t}(y_s, y_t) \prod_{t\in V} \psi_t(y_t) \\
\psi_{s,t}(y_s, y_t) &= \exp\left(-\frac{1}{2} y_s \Lambda_{st} y_t\right) \\
\psi_t(y_t) &= \exp\left(-\frac{1}{2}\Lambda_{tt} y_t^2 + \eta_t y_t\right)
\end{aligned}
$$

Gaussian MRFs

- An undirected graphical Gaussian model, also called Gaussian Markov random field, is a pair-wise MRF that has the following form

$$
\begin{aligned}
p(\mathbf{y}|\boldsymbol{\theta}) &\propto \prod_{s,t \in E} \psi_{s,t}(y_s, y_t) \prod_{t \in V} \psi_t(y_t) \\
\psi_{s,t}(y_s, y_t) &= \exp\left(-\frac{1}{2} y_s \Lambda_{st} y_t\right) \\
\psi_t(y_t) &= \exp\left(-\frac{1}{2}\Lambda_{tt} y_t^2 + \eta_t y_t\right)
\end{aligned}
$$

- After some straightforward algebra, the joint distribution can be written as

$$
p(\mathbf{y}|\boldsymbol{\theta}) \propto \exp\left(\boldsymbol{\eta}^T \mathbf{y} - \frac{1}{2}\mathbf{y}^T \boldsymbol{\Lambda} \mathbf{y}\right),
$$

which is a multivariate normal written in so-called information form

Gaussian MRFs

- Define $\mathbf{\Lambda} = \mathbf{\Sigma}^{-1}$ and $\boldsymbol{\eta} = \mathbf{\Lambda}\boldsymbol{\mu}$

$$
\begin{aligned}
p(\mathbf{y}|\boldsymbol{\theta}) &\propto \exp\left(\boldsymbol{\eta}^T\mathbf{y} - \frac{1}{2}\mathbf{y}^T\mathbf{\Lambda}\mathbf{y}\right) \\
&\propto \exp\left(\boldsymbol{\mu}^T\mathbf{\Sigma}^{-1}\mathbf{y} - \frac{1}{2}\mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{y}\right) \\
&\propto \exp\left(-\frac{1}{2}\boldsymbol{\mu}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}\right) \cdot \exp\left(\boldsymbol{\mu}^T\mathbf{\Sigma}^{-1}\mathbf{y} - \frac{1}{2}\mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{y}\right) \\
&= \exp\left(-\frac{1}{2}\mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{y} + \boldsymbol{\mu}^T\mathbf{\Sigma}^{-1}\mathbf{y} - \frac{1}{2}\boldsymbol{\mu}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}\right) \\
&= \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right) \\
&\propto \frac{1}{(2\pi)^{n/2}|\mathbf{\Sigma}|^{1/2}} \cdot \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right) \\
&= \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}, \mathbf{\Sigma})
\end{aligned}
$$

Gaussian MRFs

- If $\Lambda_{s,t} = 0$, then $\psi_{s,t}(y_s, y_t) = \exp\left(-\frac{1}{2} y_s \Lambda_{st} y_t\right) = \exp(0) = 1$ and there is no pair-wise potential term for node pair s and t
- So by the conditional probability (or factorization theorem)

$$y_s \perp y_t | \mathbf{y}_{-\{s,t\}} \quad \text{iff} \quad \Lambda_{s,t} = 0$$

- These are called structural zeros since they represent the absent edges in the underlying graph G
  - Recall again that $\mathbf{\Lambda} = \mathbf{\Sigma}^{-1}$

Gaussian MRFs

- If $\Lambda_{s,t} = 0$, then $\psi_{s,t}(y_s, y_t) = \exp\left(-\frac{1}{2} y_s \Lambda_{st} y_t\right) = \exp(0) = 1$ and there is no pair-wise potential term for node pair s and t
- So by the conditional probability (or factorization theorem)

$$y_s \perp y_t | \mathbf{y}_{-\{s,t\}} \quad \text{iff} \quad \Lambda_{s,t} = 0$$

- These are called structural zeros since they represent the absent edges in the underlying graph G
  - Recall again that $\mathbf{\Lambda} = \mathbf{\Sigma}^{-1}$
- In other words, there is one-to-one correspondence between a zero in the inverse covariance methods and lack of edges between the two variables in the underlying G
- Gaussian MRFs thus correspond to sparse precision (inverse covariance) matrices, assuming a sparse underlying graph
- We can use this property when learning sparse biological networks

Multivariate Gaussian distributions

- We now know that Gaussian MRF corresponds to a multivariate Gaussian distribution
- Learning parameters for a Gaussian MRF is more challenging than learning parameters for a general (non-sparse) multivariate normal because for a given graph G we need to maintain the zeros in the inverse covariance matrix that correspond to variable pairs that are not connected by an edge
- Learning the structure of a Gaussian MRF is challenging too
- Lets start by assuming that we want to learn parameters of the standard multivariate Gaussian without any constraints

Multivariate Gaussian distributions

- Given data $D = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the log-likelihood is

$$\ell(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \log p(D|\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \frac{N}{2} \log |\boldsymbol{\Lambda}| - \frac{1}{2} \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Lambda} (\mathbf{x}_i - \boldsymbol{\mu}),$$

  where $|\boldsymbol{\Lambda}|$ denote the determinant of $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$

Multivariate Gaussian distributions

- Given data $D = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the log-likelihood is

$$\ell(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \log p(D|\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \frac{N}{2} \log |\boldsymbol{\Lambda}| - \frac{1}{2} \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Lambda} (\mathbf{x}_i - \boldsymbol{\mu}),$$

  where $|\boldsymbol{\Lambda}|$ denote the determinant of $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$

- The well-known maximum likelihood estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are

$$\begin{aligned} \hat{\boldsymbol{\mu}} &= \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \\ \hat{\boldsymbol{\Sigma}} &= \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T = \mathbf{S} \end{aligned}$$

Structural learning for Gaussian MRF

- The first "quick and dirty" approach for estimating the underlying structure could be obtained by inverting $\mathbf{S}$ and thresholding to identify the strongest precisions
  - If $(\mathbf{S}^{-1})_{i,j} \geq \tau$, then $(i,j) \in E$

Structural learning for Gaussian MRF

- The first "quick and dirty" approach for estimating the underlying structure could be obtained by inverting $\mathbf{S}$ and thresholding to identify the strongest precisions
  - If $(\mathbf{S}^{-1})_{i,j} \geq \tau$, then $(i,j) \in E$
- A better approach could be to explicitly enforce zeros in $\mathbf{\Lambda}$ while simultaneously estimating the other parameters
- We will consider a method called graphical lasso, which uses ℓ_1 penalty for the elements of the precision matrix
  - High values of a penalty term give very sparse matrices (low number of edges in the resulting graph), but less good fit to data
  - Low values of a penalty give denser matrices/graphs but may overfit the data

Graphical lasso

- Assuming the mean of the distribution is already estimated, then the log-likelihood of the multivariate Gaussian for $\mathbf{\Lambda}$ can be written as

$$\begin{aligned}\ell(\mathbf{\Lambda}) &= \frac{N}{2}\log|\mathbf{\Lambda}| - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i-\boldsymbol{\mu})^T\mathbf{\Lambda}(\mathbf{x}_i-\boldsymbol{\mu}) \\ &= \frac{N}{2}\log|\mathbf{\Lambda}| - \frac{1}{2}\mathrm{tr}(\mathbf{S}_\mu\mathbf{\Lambda}),\end{aligned}$$

  where $\mathrm{tr}(\mathbf{A}) = \sum_i(\mathbf{A})_{ii}$ is the trace of matrix $\mathbf{A}$

Graphical lasso

- Assuming the mean of the distribution is already estimated, then the log-likelihood of the multivariate Gaussian for $\mathbf{\Lambda}$ can be written as

$$\begin{aligned}\ell(\mathbf{\Lambda}) &= \frac{N}{2}\log|\mathbf{\Lambda}| - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i-\boldsymbol{\mu})^T\mathbf{\Lambda}(\mathbf{x}_i-\boldsymbol{\mu}) \\ &= \frac{N}{2}\log|\mathbf{\Lambda}| - \frac{1}{2}\mathrm{tr}(\mathbf{S}_\mu\mathbf{\Lambda}),\end{aligned}$$

  where $\mathrm{tr}(\mathbf{A}) = \sum_i(\mathbf{A})_{ii}$ is the trace of matrix $\mathbf{A}$

- The graphical lasso is defined by the following ℓ_1 penalized negative log-likelihood

$$J(\mathbf{\Lambda}) = -\frac{N}{2}\log|\mathbf{\Lambda}| + \frac{1}{2}\mathrm{tr}(\mathbf{S}_\mu\mathbf{\Lambda}) + \rho||\mathbf{\Lambda}||_1,$$

  where ρ is a regularization parameter and $||\mathbf{\Lambda}||_1 = \sum_{k,l}|\lambda_{kl}|$

Graphical lasso

- Graphical lasso does not have a closed form solution but $J(\mathbf{\Lambda})$ needs to be minimized using numerical optimization
  - Gradient-based optimization
  - Coordinate gradient descent: repeatedly optimize one coordinate at a time

Graphical lasso for signaling pathways

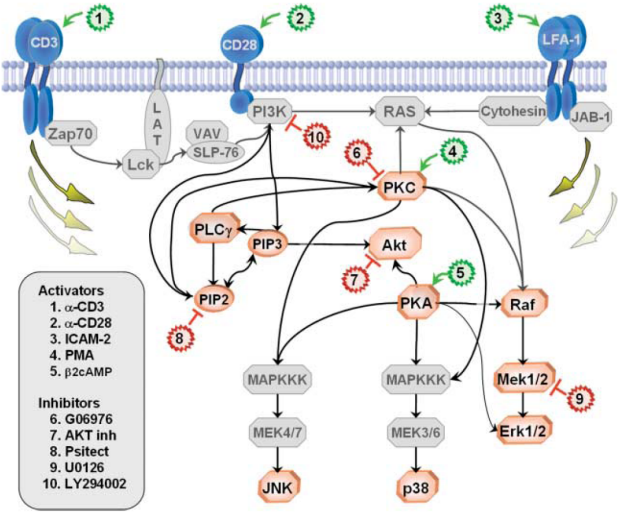


Figure: Figure from (Sachs et al., 2005)

Graphical lasso for signaling pathways

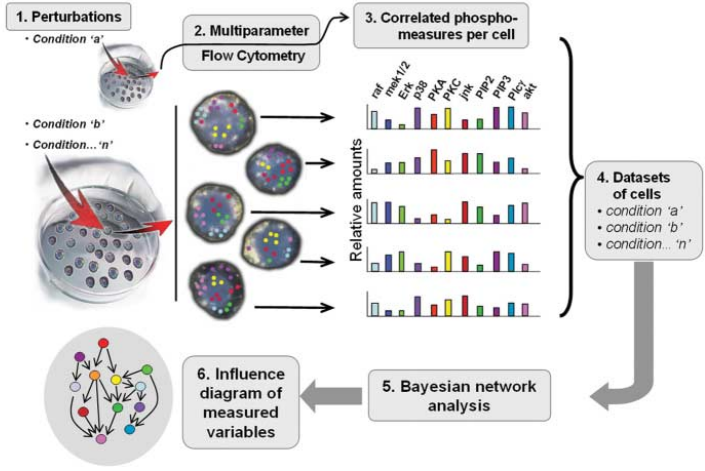


Figure: Figure from (Sachs et al., 2005)

Graphical lasso for signaling pathways

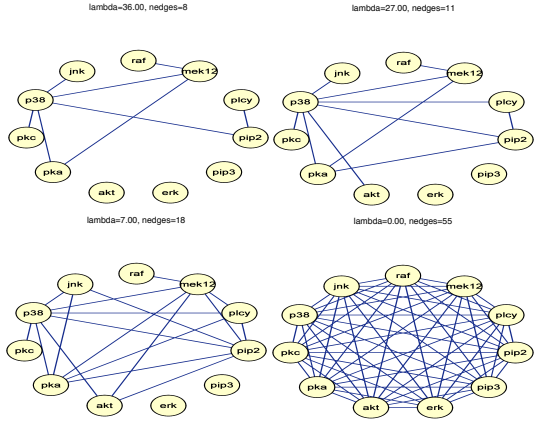
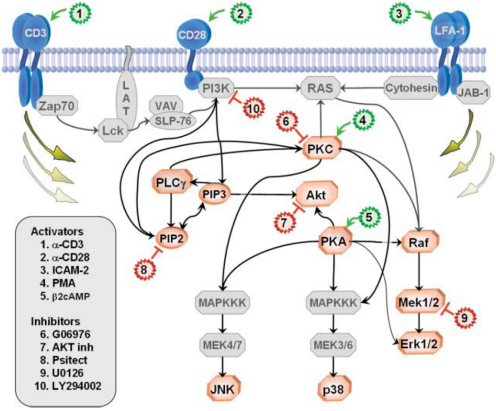


Figure: Figure from (Murphy, 2012)

- Regulazation parameter is often set by the cross-validation

Graphical lasso for signaling pathways



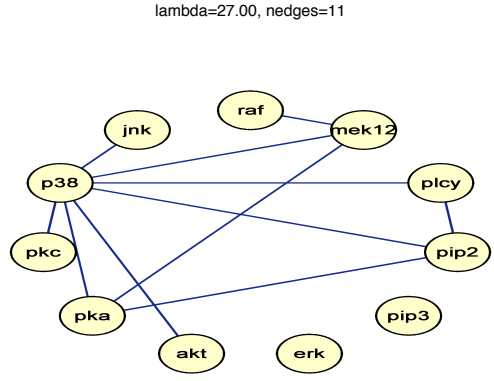


Figure: Figures from (Sachs et al., 2005; Murphy, 2012)

References

- Murphy K (2012) Machine learning: a probabilistic perspective, MIT Press.
- Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., & Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, Vol. 308, No. 5721, pp. 523-529.