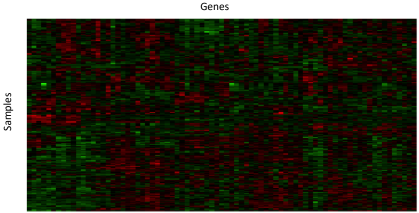
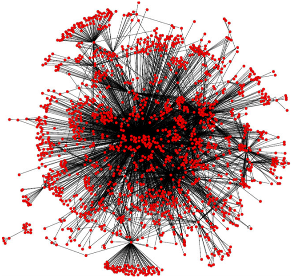
Graphical Models

Motivation: constructing gene regulatory networks from gene expression data





Nodes are genes, and edges represent regulatory interactions.

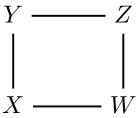
Overview

- Graphical models are a family of multivaiate distributions (e.g. multivariate Gaussian) with certain parsimonious assertions.
- Each variable is represented by a node in a graph.
- Graphical models provide a compact way of representing conditional independence relationships through a graph structure (presence and absence of edges).
- Loosely speaking, nodes are observed whereas the graph is often hidden.
- Learning graph structure is one of the key interests in graphical models.

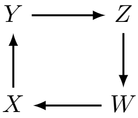
Graphs

A **graph** $\mathcal{G} = (V, E)$ consists of a set of **nodes** $V = \{1, \ldots, p\}$ and a set of **edges** $E \subseteq V \times V$.

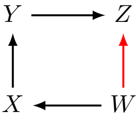
- **Undirected graphs**: edges are undirected $i - j$ or $\{i, j\} \in E$. Also known as **Markov Random Field**.



- **Directed graphs**: edges are directed $i \to j$ or $(i, j) \in E$
    - We will focus on directed **acyclic** graphs (**DAG**): starting from any node, one cannot return to it by following the arrows. Also known as **Bayesian Network**. DAG has potential **causal** interpretations.



not a DAG



DAG

- Field goal percentage (FG%) in a season

$$\frac{\#\text{ of shots made in the season}}{\#\text{ of shots attempted in the season}}$$

Basketball statistic

- Field goal percentage (FG%) in a season

$$\frac{\#\text{ of shots made in the season}}{\#\text{ of shots attempted in the season}}$$

- Compare two players Adam and Bob
  - In the 1st half season, Adams FG% $>$ Bobs
  - In the 2nd half season, Adams FG% $>$ Bobs

Basketball statistic

- Field goal percentage (FG%) in a season

$$\frac{\text{\# of shots made in the season}}{\text{\# of shots attempted in the season}}$$

- Compare two players Adam and Bob
  - In the 1st half season, Adams FG% $>$ Bobs
  - In the 2nd half season, Adams FG% $>$ Bobs
- Q: How does Adams full season FG% compared to Bobs?

- Field goal percentage (FG%) in a season

$$\frac{\#\text{ of shots made in the season}}{\#\text{ of shots attempted in the season}}$$

- Compare two players Adam and Bob
  - In the 1st half season, Adams FG% $>$ Bobs
  - In the 2nd half season, Adams FG% $>$ Bobs
- Q: How does Adams full season FG% compared to Bobs?
- A: Bobs FG% could be better.

Silly example

	Adam	Bob
1st half	?/1	?/1000
2nd half	?/1000	?/1
Total	?/1001	?/1001

Silly example: continued

	Adam	Bob
1st half	1/1	999/1000
2nd half	1/1000	0/1
Total	2/1001	999/1001

Silly example: continued

	Adam	Bob
1st half	1/1	999/1000
2nd half	1/1000	0/1
Total	2/1001	999/1001

- This is known as **Simpsons paradox**.

Admissions to Berkeley by department

Sex	Whether Admitted	
	Yes	No
Male	1198	1493
Female	557	1278

- Two variables S: Sex and A: Admitted?
- Is the admission independent of sex?
  - Recall: two random variables X, Y are independent, denoted by $X \perp\!\!\!\perp Y$, if
$$P(X, Y) = P(X)P(Y)$$
- Females have much lower admission rates than males.

Admissions to Berkeley by department: continued

More detailed table

Department	Sex	Whether Admitted	
		Yes	No
I	Male	512	313
	Female	89	19
II	Male	353	207
	Female	17	8
III	Male	120	205
	Female	202	391
IV	Male	138	279
	Female	131	244
V	Male	53	138
	Female	94	299
VI	Male	22	351
	Female	24	317

- An additional variable D: Department.
- When dealing with complex systems of many random variables, we must have a concept which is more sophisticated, but equally fundamental: that of **conditional independence**.

A first look at conditional independence

- For three variables it is of interest to see whether independence holds for fixed value of one of them, e.g. is the admission independent of sex for every department separately?
- We denote this as $A \perp\!\!\!\perp S|D$ and display it graphically as

$$A — D — S$$

- Note that there the two conditions

$$A \perp\!\!\!\perp S \quad \text{and} \quad A \perp\!\!\!\perp S|D$$

are **very different** and will typically not both hold unless we either have $A \perp\!\!\!\perp (D, S)$ or $(A, D) \perp\!\!\!\perp S$, i.e. if one of the variables is completely independent of both of the others.

Back to Admissions Example

Department	Sex	Whether Admitted	
		Yes	No
I	Male	512	313
	Female	89	19
II	Male	353	207
	Female	17	8
III	Male	120	205
	Female	202	391
IV	Male	138	279
	Female	131	244
V	Male	53	138
	Female	94	299
VI	Male	22	351
	Female	24	317

- Apart from Department I, it holds that $A \perp\!\!\!\perp S | D$. In Department I, a higher proportion of females are admitted!

Kidney stone treatment

- Treatment A: open surgical procedures
- Treatment B: percutaneous nephrolithotomy (which involves only a small puncture)

Treatment	Success	Failure	Success rate
A	273	77	78%
B	289	61	83%

- Is the successful rate independent of the choice of treatment?
- If you had a kidney stone, which treatment would you choose?

Kidney stone treatment: continued

Stone size	Treatment A	Treatment B
Small Stones	93%(81/87)	87%(234/270)
Large Stones	73%(192/263)	69%(55/80)
Both	78%(273/350)	83%(289/350)

- If we condition on the stone size, Treatment A is obviously better...

- Two random variables X, Y are **conditionally independent** given a third random variable Z, denoted by $X \perp\!\!\!\perp Y | Z$, if

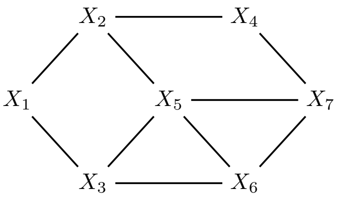
$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

- Equivalently,

$$P(X|Y, Z) = P(X|Z)$$

- Intuitively, knowing Z renders Y irrelevant for predicting X.
- Note that (marginal) independence is a special case of conditional independence ($Z = \emptyset$).
- **Graphical models are all about conditional independence.**
- We are going to discuss:
  - given a graph, how to read off the conditional independence
  - given data, how to estimate a graph

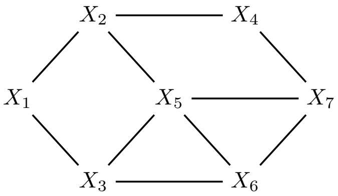
- For several variables, complex systems of conditional independence can for example be described by undirected graphs:



- **Graph separation**: two sets of nodes A and B are separated by a third set C if every path between A and B has to pass C. For example,
  - $A = \{X_1\}, B = \{X_4, X_7\}, C = \{X_2, X_3\}$
  - $A = \{X_1, X_2\}, B = \{X_7\}, C = \{X_4, X_5, X_6\}$
- **Global Markov property (G)**: a set of variables A is conditionally independent of set B, given the values of a set of variables C if C separates A from B.

$$X_A \perp\!\!\!\perp X_B | X_C$$

- E.g. $X_1 \perp\!\!\!\perp (X_4, X_7)|X_2, X_3$ and $(X_1, X_2) \perp\!\!\!\perp X_7|X_4, X_5, X_6$.

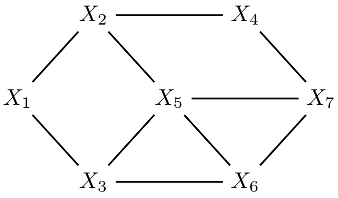


- **Local Markov property (L)**: a variable is conditionally independent of all other variables given its neighbors:

$$X_i \perp\!\!\!\perp X_{rest} | X_{N(i)}$$

  where $N(i)$ is the set of neighbors of node i (i.e. all nodes connected to node i) and *rest* is the set of all other nodes.
- $G \implies L$
- E.g. $X_5 \perp\!\!\!\perp X_1, X_4 | X_2, X_3, X_6, X_7$ and $X_6 \perp\!\!\!\perp X_1, X_2, X_4 | X_3, X_5, X_7$

Pairwise Markov property



- **Pairwise Markov property (P)**: any two non-adjacent variables i and j are conditionally independent given all other variables

$$X_i \perp\!\!\!\perp X_j | X_{rest}$$

- $G \implies P$
- E.g. $X_1 \perp\!\!\!\perp X_5 | X_{rest}$ and $X_2 \perp\!\!\!\perp X_6 | X_{rest}$

$G \iff L \iff P$

- We have seen $G \implies L \implies P$.
- If we assume the density of X is continuous and positive, $P \implies G$

Centered Multivariate Gaussian

- Multivariate normal/Gaussian random variables $X = (X_1, \dots, X_p)^T \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean $\boldsymbol{\mu}$ and covariacne $\boldsymbol{\Sigma}$.
- Since we are not interested in the mean $\boldsymbol{\mu}$ in graphical models, we assume $\boldsymbol{\mu} = 0$ throughout. In practice, we center the data.
- With n realizations of X, $x_1, \dots, x_n$, the log-likelihood is given by

$$\ell(\boldsymbol{\Sigma}) = -\frac{n}{2} \log \det \boldsymbol{\Sigma} - \frac{1}{2} x_i^T \boldsymbol{\Sigma}^{-1} x_i$$

  which is equivalent to

$$\ell(\boldsymbol{\Sigma}) = -\frac{n}{2} \log \det \boldsymbol{\Sigma} - \frac{n}{2} \mathrm{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1})$$

  where $\mathbf{S} = \frac{1}{n} \sum_{i=1}^n x_i x_i^T$ and we have used the "trace trick": $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$
- The maximum likelihood estimate of $\boldsymbol{\Sigma}$ is simply $\mathbf{S}$.

Undirected Gaussian Graphical Models

- Let $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$. It is called **precision matrix** or concentration matrix or inverse covariance matrix.
- The log-likelihood in terms of precision matrix $\boldsymbol{\Theta}$

$$\ell(\boldsymbol{\Theta}) = \log\det\boldsymbol{\Theta} - \mathrm{tr}(\mathbf{S}\boldsymbol{\Theta})$$

- **Markov property**: $X_j \perp\!\!\!\perp X_k | \text{rest}$ (the rest of the variables) $\iff$ missing edge between X_j and X_k $\iff \theta_{jk} = 0$
- Example:

$$\begin{tikzcd}
X_3 \arrow[r, no head] \arrow[d, no head] & X_4 \arrow[d, no head] \\
X_2 \arrow[r, no head] & X_1
\end{tikzcd}
\qquad
\boldsymbol{\Theta} = \begin{pmatrix} * & * & 0 & * \\ * & * & * & 0 \\ 0 & * & * & * \\ * & 0 & * & * \end{pmatrix}$$

- Learning graph structure $\iff$ finding a **sparse** solution for $\mathbf{\Theta}$ (i.e. the zero patterns).
    - Neighbourhood selection (Meinshausen and Bühlmann, 2006)
    - Graphical lasso (Friedman, Hastie and Tibshirani, 2008)

Conditional distribution of multivariate Gaussian

- Recall: $X = (X_1, \ldots, X_p)^T \sim N(0, \mathbf{\Sigma})$ and $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$.
- Suppose we partition $X = (Z^T, Y)^T$ where $Z = (X_1, \ldots, X_{p-1})^T$ and $Y = X_p$.
- Partition $\mathbf{\Sigma}$

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{ZZ} & \sigma_{ZY} \\ \sigma_{ZY}^T & \sigma_{YY} \end{pmatrix}$$

- From standard normal theory,

$$Y|Z = z \sim N(z^T \mathbf{\Sigma}_{ZZ}^T \sigma_{ZY}, \sigma_{YY} - \sigma_{ZY}^T \mathbf{\Sigma}_{ZZ}^{-1} \sigma_{ZY})$$

- This is simply a linear regression with $\beta = \mathbf{\Sigma}_{ZZ}^T \sigma_{ZY}$

$$Y = z^T \beta + \epsilon, \quad \epsilon \sim N(0, \sigma_{YY} - \sigma_{ZY}^T \mathbf{\Sigma}_{ZZ}^{-1} \sigma_{ZY})$$

In terms of precision matrix...

- We partition $\mathbf{\Theta}$ in the same way

$$\mathbf{\Theta} = \begin{pmatrix} \mathbf{\Theta}_{ZZ} & \theta_{ZY} \\ \theta_{ZY}^T & \theta_{YY} \end{pmatrix}$$

- Since $\mathbf{\Sigma}\mathbf{\Theta} = \mathbf{I}$, standard formulas for partitioned inverses give

$$\theta_{ZY} = -\theta_{YY}\mathbf{\Sigma}_{ZZ}^{-1}\sigma_{ZY} = -\theta_{YY}\beta$$

  where $1/\theta_{YY} = \sigma_{YY} - \sigma_{ZY}^T\mathbf{\Sigma}_{ZZ}^{-1}\sigma_{ZY} > 0$.
- β and θ_{ZY} are only different by a scale of $-\theta_{YY}$
- Hence, the zeros patterns of β and θ_{ZY} are exactly the same.

- Fit a lasso regression using each variable as the response and the others as predictors

$$n^{-1} \sum_{i=1}^{n} \left(x_{ij} - \sum_{k \neq j} \beta_{jk} x_{ik} \right)^2 + \lambda ||\beta_j||_1$$

  where $\beta_j = (\beta_{j1}, \ldots, \beta_{j,j-1}, \beta_{j,j+1}, \ldots, \beta_{jp})^T$.
- **Asymmetry**: there is no guarantee that $\beta_{jk} \neq 0$ whenever $\beta_{kj} \neq 0$. So need to apply one of the following two rules:
  - **OR** rule: $\theta_{jk} \neq 0$ if $\beta_{jk} \neq 0$ or $\beta_{kj} \neq 0$.
  - **AND** rule: $\theta_{jk} \neq 0$ if $\beta_{jk} \neq 0$ and $\beta_{kj} \neq 0$.
- **Advantage:** Simple and fast.
- **Disadvantage**: it only estimates which components of θ_{jk} are nonzero but doesn't fully estimate $\boldsymbol{\Sigma}$ or $\boldsymbol{\Theta}$.

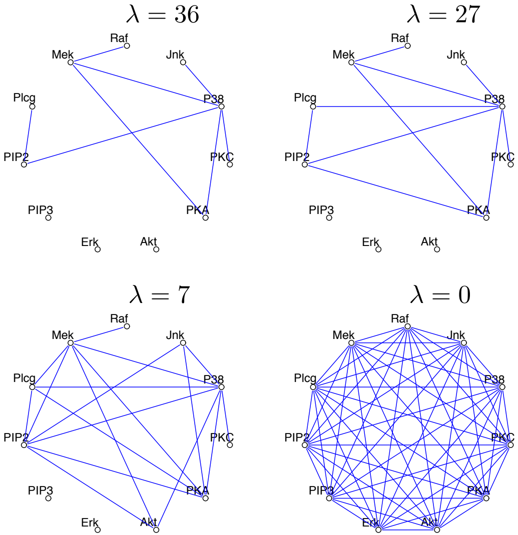
- Graphical lasso (glasso) is a more systematic approach.
- Glasso maximizes the penalized log-likelihood

$$\log \det \mathbf{\Theta} - \text{tr}(\mathbf{S\Theta}) - \lambda ||\mathbf{\Theta}||_1$$

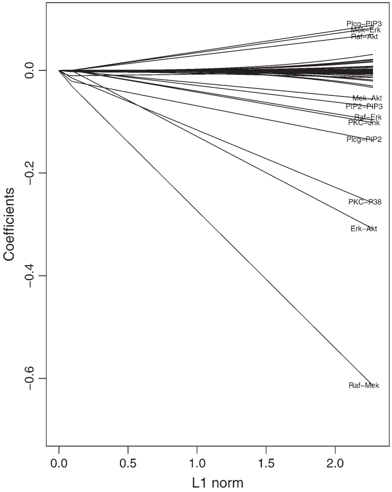
- The negative penalized log-likelihood is convex.
- Neighbourhood selection can be seen as an approximation of glasso.
- Both neighborhood selection and glasso require cross-validation to select λ.

Flow Cytometry

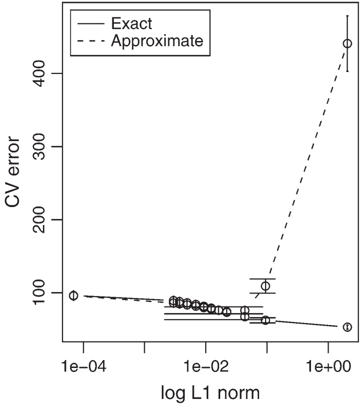
Flow cytometry dataset with $p = 11$ proteins measured on $n = 7,466$ cells.



Solution path of glasso



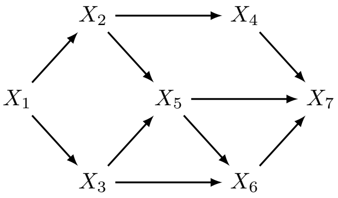
Solution path as the total ℓ_1 norm of the coefficient vector increases, that is as λ decreases. The largest coefficients are labeled with the corresponding pair of proteins.



CV errors of the exact graphical lasso approach versus the approximation by neighborhood selection. For lightly regularized models, the exact approach has a clear advantage.

Directed Acyclic Graphs

- DAGs are also natural models for conditional independence:

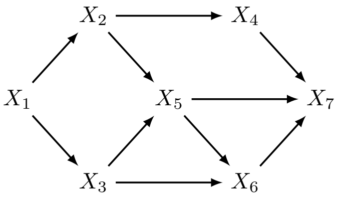


- **Parents**: X is a parent of Y if $X \to Y$
- **Descendants and Ancestors**: Y is a descendant of X and X is an ancestor of Y if $X \to \cdots \to Y$
- **Directed local Markov property**: any variable is conditional independent of its non-descendants, given its parents

$$X_i \perp\!\!\!\perp X_{nd(i)} | X_{pa(i)}$$

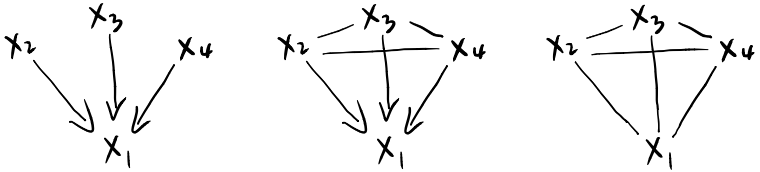
  where $nd(i)$ is the set of non-descendants of node i and $pa(i)$ is the set of parents.
- E.g. $X_5 \perp\!\!\!\perp (X_1, X_4) | X_2, X_3$ and $X_6 \perp\!\!\!\perp (X_1, X_2, X_4) | X_3, X_5$.

Anterior set

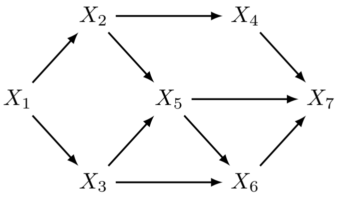


- A subset $A \subseteq V$ is **anterior** if there is not edges pointing towards A.
- E.g. $A = \{2, 5\}$ is not anterior and $A = \{1, 2, 3, 5\}$ is anterior.
- The **minimal anterior set** containing A is denoted by $an(A)$.
- E.g. $A = \{2, 5\} \implies an(A) = \{1, 2, 3, 5\}$
  $A = \{6\} \implies an(A) = \{1, 2, 3, 5, 6\}$

The moral graph G^m of a DAG G is obtained by adding undirected edges between unmarried parents and subsequently dropping directions, as in the example below:



Global Markov property



- **Directed global Markov property (DG)**: a set of variables A is conditionally independent of set B, given the values of a set of variables C if C separates A from B in graph $G^m_{an(A \cup B \cup C)}$.

$$X_A \perp\!\!\!\perp X_B | X_C$$

- E.g. $X_4 \perp\!\!\!\perp X_5 | X_2$. However, X_2 and X_3 are not conditionally independent given X_5.

Factorization

- Product rule of density:

$$f(X_1, \dots, X_p) = f(X_1)f(X_2|X_1)\cdots f(X_p|X_1, \dots, X_{p-1})$$

- A density $f()$ is said to **factorize** with respect to a DAG if

$$f(X_1, \dots, X_p) = \prod_{j=1}^{p} f(X_j|X_{pa(j)})$$

- Example:

$$\begin{tikzcd}
X_3 \arrow[r] & X_4 \\
X_2 \arrow[u] & X_1 \arrow[l] \arrow[u]
\end{tikzcd}$$

$$\begin{aligned}
f(X_1, X_2, X_3, X_4) &= f(X_1)f(X_2|X_1)f(X_3|X_1, X_2)f(X_4|X_1, X_2, X_3) \quad && \textbf{Product rule} \\
&= f(X_1)f(X_2|X_1)f(X_3|X_2)f(X_4|X_1, X_3) && \textbf{Factorization}
\end{aligned}$$

- From last slide,

$$f(X_3|X_1, X_2) = f(X_3|X_2) \iff X_1 \perp\!\!\!\perp X_3|X_2$$
$$f(X_4|X_1, X_2, X_3) = f(X_4|X_1, X_3) \iff X_2 \perp\!\!\!\perp X_4|X_1, X_3$$

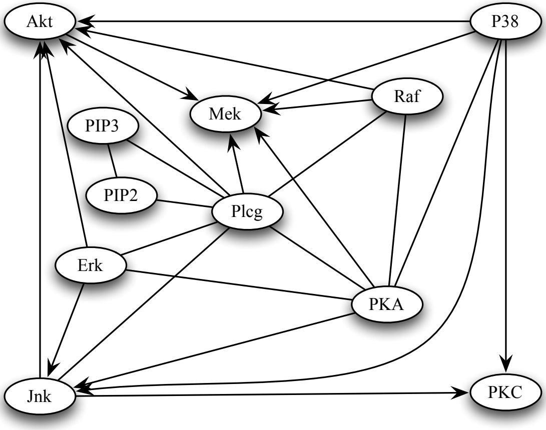
- If $X = (X_1, \ldots, X_p)^T$ is multivariate Gaussian, then for $A \subseteq X \cap \{X_j, X_k\}^c$

$$X_j \perp\!\!\!\perp X_k|A \ \text{ if and only if } \ \rho_{jk|A} = 0$$

  where $\rho_{jk|A}$ is the **partial correlation** between X_j and X_k given A.
- Partial correlation is the correlation of two residuals, the residual of regressing X_j on A and the residual of X_k on A.
- Learning graph structure $\iff$ **testing** partial correlations

- **PC-algorithm** (Spirtes, Glymour and Scheines, 2000) is one of the popular algorithms that allow us to perform the tests efficiently.
- Like any statistical testing procedure, PC-algorithm needs to specify the **significance level** α.
- The output of PC-algorithm may not be a DAG. It may also contain undirected edges. The directionality of those undirected edges are undetermined because both directions imply exactly the same statistical model (also known as **Markov equivalence**).

PC algorithm applied to the flow cytometry dataset with $\alpha = 0.01$.



Compare with the undirected graph: the connections to Mek.