Probabilistic Graphical Models
What have we learned so far?

Graph $\rightarrow$ Graph Laplacian $\rightarrow$ Structures

https://team.inria.fr
What Are Probabilistic Graphical Models (PGMs)?

Graph

Probabilistic Models

\[ P(Y_1, \ldots, Y_p) \]

https://team.inria.fr
What Are Probabilistic Graphical Models (PGMs)?

Graph

Probabilistic Models

\[ P(Y_1, \ldots, Y_p) \]

Data

\[ D = \{ Y_1^{(i)}, Y_2^{(i)}, \ldots, Y_p^{(i)} \}_{i=1}^n \]

https://team.inria.fr
Where Do We Use PGMs?

Brain imaging genetic studies

Age 7

Age 12

Huitong Qiu et. Al., 2015
Where Do We Use PGMs?


Democratic senators  Republican senators

Banerjee et. al., 2008
Where Do We Use PGMs?

**Breast cancer example** (Hess et al., 2007)

Two types of patients

- Responders to Chemo therapy \( n = 34 \)
- Non responders to Chemo \( n = 99 \).

After collecting gene expression data for those tumors, what can we do?
Where Do We Use PGMs?

Breast cancer example (Hess et al., 2007)
Two types of patients
▶ Responders to Chemo therapy (n = 34)
▶ Non responders to Chemo (n = 99).

After collecting gene expression data for those tumors, what can we do?
Supervised classification
▶ Find genes with significant different expression levels between groups.
Where Do We Use PGMs?

Breast cancer example (Hess et al., 2007)

Two types of patients

- Responders to Chemo therapy (n = 34)
- Non responders to Chemo (n = 99).

After collecting gene expression data for those tumors, what can we do?

**Supervised classification**

- Find genes with significant different expression levels between groups.

**Prediction**
Where Do We Use PGMs?

**Breast cancer example** (Hess et al., 2007)
Two types of patients
- Responders to Chemo therapy \((n = 34)\)
- Non responders to Chemo \((n = 99)\).

After collecting gene expression data for those tumors, what can we do?

**Supervised classification**
- Find genes with significant different expression levels between groups.

**Prediction**
**Intervention?** How can we turn non-responder to Responder?
Where Do We Use PGMs?

Breast cancer example (Hess et al., 2007)
How to Fit PGMs?

Definition (Dependence of events)

Two events $A$ and $B$ are independent iff

$$P(A, B) = P(A)P(B),$$

denoted as $A \perp \perp B$. Equivalently,

- $A \perp \perp B \iff P(A|B) = P(A)$
- $A \perp \perp B \iff P(A|B^c) = P(A)$
How to Fit PGMs?

Definition (Dependence of events)

Two events $A$ and $B$ are independent iff

$$P(A, B) = P(A)P(B),$$

denoted as $A \perp \perp B$. Equivalently,

- $A \perp \perp B \iff P(A|B) = P(A)$
- $A \perp \perp B \iff P(A|B^c) = P(A)$

Example 1: Snow Attitude and Age
How to Fit PGMs?

Definition (Dependence of events)

Two events $A$ and $B$ are independent iff

$$P(A, B) = P(A)P(B),$$

denoted as $A \perp \perp B$. Equivalently,

- $A \perp \perp B \iff P(A|B) = P(A)$
- $A \perp \perp B \iff P(A|B^c) = P(A)$

Example 1: Snow Attitude and Age

Joint distribution

<table>
<thead>
<tr>
<th></th>
<th>Like Snow</th>
<th>Dislike Snow</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teenagers</td>
<td>0.48</td>
<td>0.12</td>
<td>0.6</td>
</tr>
<tr>
<td>Middle Age</td>
<td>0.32</td>
<td>0.08</td>
<td>0.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Like Snow</th>
<th>Dislike Snow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teenagers</td>
<td>0.8</td>
<td>0.2</td>
</tr>
<tr>
<td>Middle Age</td>
<td>0.8</td>
<td>0.2</td>
</tr>
</tbody>
</table>
Introduction: Statistical Dependence

Example 2: Intelligence and Height

Consider the events $A = \text{’having high IQ’}$, $B = \text{’Being tall’}$. Are $A$ and $B$ independent?
Introduction: Statistical Dependence

Example 2: Intelligence and Height

Consider the events $A = \text{’having high IQ’}, \ B = \text{’Being tall’}. Are A and B independent?
Introduction: Statistical Dependence

Example 2: Intelligence and Height

Consider the events $A = \text{’having high IQ’}$, $B = \text{’Being tall’}$. Are $A$ and $B$ independent?
Introduction: Statistical Dependence

Example 2: Intelligence and Height

Consider the events $A = 'having high IQ'$, $B = 'Being tall'$. Are $A$ and $B$ independent?
Definition (Conditional dependence of events)

Two events $A$ and $B$ are independent given $C$ iff

$$P(A, B|C) = P(A|C)P(B|C),$$

denoted as $A \perp \!\!\!\!\!\!\! \perp B|C$. 
Introduction: Statistical Dependence

Definition (Conditional dependence of events)

Two events $A$ and $B$ are independent given $C$ iff

$$P(A, B|C) = P(A|C)P(B|C),$$

denoted as $A \perp B|C$.

Example 2:

Consider the events $A = 'having high IQ'$, $B = 'being tall'$, and $C = 'at a given age'$, Are $A$ and $B$ independent given $C$?
Introduction: Statistical Dependence

Definition (Dependence of random variables)

Let random variables $X, Y, Z$ have marginal probability density function (pdf) $f_X, f_Y, f_Z$, and joint pdf $f_{XY}, f_{XYZ}$. Then

- $X$ and $Y$ are independent iff $f_{XY}(x, y) = f_X(x)f_Y(y)$;
- $X$ and $Y$ are conditionally independent given $Z$ iff $f_{XY|Z}(x, y; z) = f_{X|Z}(x; z)f_{Y|Z}(y; z)$. 

Introduction: Covariance and Correlation

Let $X$ and $Y$ be two real random variables.

Definitions

$$\text{cov}(X, Y) = \mathbb{E}\left[ (X - \mathbb{E}(X))(Y - \mathbb{E}(Y)) \right] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$
Introduction: Covariance and Correlation

Let $X$ and $Y$ be two real random variables.

**Definitions**

\[
\text{cov}(X, Y) = \mathbb{E} \left[ (X - \mathbb{E}(X))(Y - \mathbb{E}(Y)) \right] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).
\]

\[
\rho_{XY} = \text{cor}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}
\]

- $\text{cov}(X, X) = \text{Var}(X)$.
- $X \perp Y \Rightarrow \text{cov}(X, Y) = 0.$
Introduction: Covariance and Correlation

Let $X$ and $Y$ be two real random variables.

**Definitions**

\[
\text{cov}(X, Y) = \mathbb{E} \left[ (X - \mathbb{E}(X))(Y - \mathbb{E}(Y)) \right] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).
\]

\[
\rho_{XY} = \text{cor}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}.
\]

**Proposition**

- $\text{cov}(X, X) = \text{Var}(X)$.
- $X \perp \!\!\!\!\!\perp Y \Rightarrow \text{cov}(X, Y) = 0$.
- $X \perp \!\!\!\!\!\perp Y \iff \text{cov}(X, Y) = 0$ only when $(X, Y)$ follow multivariate Gaussian (Normal) distribution.
Introduction: Covariance and Correlation

Let $X$ and $Y$ be two real random variables.

**Definitions**

\[
\text{cov}(X, Y) = \mathbb{E} \left[ (X - \mathbb{E}(X))(Y - \mathbb{E}(Y)) \right] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).
\]

\[
\rho_{XY} = \text{cor}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}.
\]

**Proposition**
Partial Covariance and Partial Correlation

Let $X$, $Y$ and $Z$ be real random variables.

**Definitions**

$$\text{cov}(X, Y | Z) = \text{cov}(X, Y) - \frac{\text{cov}(X, Z)\text{cov}(Y, Z)}{\text{var}(Z)}.$$  

$$\rho_{XY|Z} = \text{cor}(X, Y | Z) = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{(1 - \rho_{XZ}^2)(1 - \rho_{YZ}^2)}}.$$  

Measure the interaction between $X$ and $Y$ after removing the effect of $Z$.

**Proposition**
Partial Covariance and Partial Correlation

Let $X$, $Y$ and $Z$ be real random variables.

**Definitions**

\[
\text{cov}(X, Y|Z) = \text{cov}(X, Y) - \frac{\text{cov}(X, Z)\text{cov}(Y, Z)}{\text{var}(Z)}
\]

\[
\rho_{XY|Z} = \text{cor}(X, Y|Z) = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{(1 - \rho_{XZ}^2)(1 - \rho_{YZ}^2)}}.
\]

Measure the interaction between $X$ and $Y$ after removing the effect of $Z$.

**Proposition**

$\Downarrow \quad X \perp \!
\!
\!
\!
\!
\!
\!\perp Y|Z \Rightarrow \text{cov}(X, Y|Z) = 0$
Partial Covariance and Partial Correlation

Let $X$, $Y$ and $Z$ be real random variables.

**Definitions**

\[
\text{cov}(X, Y|Z) = \text{cov}(X, Y) - \frac{\text{cov}(X, Z)\text{cov}(Y, Z)}{\text{var}(Z)}
\]

\[
\rho_{XY|Z} = \text{cor}(X, Y|Z) = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{(1 - \rho_{XZ}^2)(1 - \rho_{YZ}^2)}}.
\]

Measure the interaction between $X$ and $Y$ after removing the effect of $Z$.

**Proposition**

- $X \perp \perp Y|Z \Rightarrow \text{cov}(X, Y|Z) = 0$

- $X \perp \perp Y|Z \Leftarrow \text{cov}(X, Y|Z) = 0$ only when $(X, Y, Z)$ follow multivariate Gaussian distribution.
Multivariate Gaussian Distribution

Let \( Y = (Y_1, \ldots, Y_p)^T \sim \mathcal{N}(\mu, \Sigma) \) with

\[
\mu = (\mathbb{E}(Y_1), \ldots, \mathbb{E}(Y_p))^T
\]

\[
\Sigma = [\text{cov}(Y_i, Y_j)], \quad i = 1, \ldots, p; \quad j = 1, \ldots, p,
\]

is the covariance matrix we have

\[
f(Y) = \frac{1}{\sqrt{\det(\Sigma)(2\pi)^p}} \exp\left\{ -\frac{1}{2} (Y - \mu)^T \Sigma^{-1} (Y - \mu) \right\}.
\]
What Are the edges in Undirected PGMs?

- An edge between two nodes iff variables are (conditionally) dependent.
- Undirected edges since this relationship is symmetric.
- The graph $G(\mathcal{P}, \mathcal{E})$ summarizes the (conditional) dependency structure among variables.

Graph $G(\mathcal{P}, \mathcal{E})$
Dependence or Conditional Dependence?

**Example:** Given a network as below

![Network Diagram]

- A → B → C → D → E
Dependence or Conditional Dependence?

Example: Given a network as below

![Network Diagram]

What is the graph representing the dependency structure among random variables?
Dependence or Conditional Dependence?

Example: Given a network as below

What is the graph representing the dependency structure among random variables?
Dependence or Conditional Dependence?

**Example:** Given a network as below

![Network Diagram]

What is the graph representing the conditional dependency structure given all other nodes?
Dependence or Conditional Dependence?

**Example:** Given a network as below

![Diagram of a network with nodes A, B, C, D, E and arrows between them]

What is the graph representing the conditional dependency structure given all other nodes?

![Diagram of a conditional dependence graph with nodes A, B, C, D, E and bidirectional arrows between some nodes]
Dependence or Conditional Dependence?

Example: Given a network as below

What is the graph representing the conditional dependency structure given all other nodes?
Markov Random Field

Markov Random Field (MRF) with graph $\mathcal{G}(\mathcal{P}, \mathcal{E})$

- Summarizes conditional independence among variables
- $(i,j) \notin \mathcal{E} \iff X_i \perp\!\!\!\perp X_j | \mathcal{P} \setminus \{i,j\}$

Gene expression network models
Gaussian Graphical Models (GGMs)

Assuming $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$.

- Simplest continuous MRFs
- $\Omega = \Sigma^{-1} = [\omega_{i,j}]$ is the precision matrix.
- Partial correlation $= 0 \Leftrightarrow$ conditional independence.
- $\omega_{i,j} = 0$ iff $X_i$ and $X_j$ are conditionally independent given all the other random variables.

$$\text{Corr}(X_i, X_j|X_{\mathcal{P}\setminus i,j}) = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}, i \neq j$$

- Estimating a Gaussian graphical model $\Leftrightarrow$ Estimating $\Omega$.
- Key idea: network inference as parameter estimation
Gaussian Graphical Models (GGMs)

The parameter matrix encodes the graph structure

\[ \Omega = \begin{bmatrix}
3 & 0 & 2 & 0 & 0 \\
2 & 1 & 0 & 0 \\
3 & 2 & 0 \\
4 & 1 \\
5 \\
\end{bmatrix} \]

The undirected graph of \( X = (X_1, X_2, X_3, X_4, X_5) \).
How to Fit GGMs?

- A microarray can be represented as a multivariate Gaussian vector $\mathbf{X} = (X_1, ..., X_p) \sim \mathcal{N}(\mu, \Sigma)$.
- Consider $n$ biological replicate in the same condition, which forms a usual $n$-size sample $\mathbf{D} = (\mathbf{X}^{(1)}, ..., \mathbf{X}^{(n)})^T$.
- $\mathbf{X}^{(1)}, ..., \mathbf{X}^{(n)}$ are independent and identically distributed (i.i.d.) copies of $\mathbf{X}$.
Sparsity

One common assumption to make: **sparsity**: Most of $\omega_{i,j}$ are exactly zero.

- **Makes empirical sense**: Genes are only assumed to interface with small groups of other genes.

- **Makes statistical sense**: Learning is now feasible in high dimensions with small sample size ($p \gg n$).
How to Fit GGMs?

Let $X$ be a random vector with distribution defined by $f_X(x; \Omega)$, where $\Omega$ is the model parameter.
How to Fit GGMs?

Let $X$ be a random vector with distribution defined by $f_X(x; \Omega)$, where $\Omega$ is the model parameter.

**Maximum likelihood estimator (MLE)**

$$\hat{\Omega} = \arg\max_{\Omega} \mathcal{L}(\Omega, X),$$

where $\mathcal{L}(\Omega, X)$ is the log likelihood, a function of $\Omega$:

$$\mathcal{L} = \log \prod_{i=1}^{n} f_X(X^{(i)}; \Omega) \propto \log \det(\Omega) - \text{tr}(\hat{\Sigma}\Omega),$$

where $\hat{\Sigma}$ is sample covariance matrix.
How to Fit GGMs?

Let $X$ be a random vector with distribution defined by $f_X(x; \Omega)$, where $\Omega$ is the model parameter.

**Maximum likelihood estimator (MLE)**

$$\hat{\Omega} = \arg\max_\Omega L(\Omega, X),$$

where $L(\Omega, X)$ is the log likelihood, a function of $\Omega$:

$$L = \log \prod_{i=1}^{n} f_X(X^{(i)}; \Omega) \propto \log \det(\Omega) - \text{tr}(\hat{\Sigma}\Omega),$$

where $\hat{\Sigma}$ is the sample covariance matrix.

**Remark**

- This a concave optimization problem. There is unique solution when $\hat{\Sigma}$ is positive definite.
- The solution $\hat{\Omega}$ typically is not sparse.
- With $n < p$, the empirical covariance matrix cannot be inverted and thus no unique solution.
Sparsity

One common assumption to make: **sparsity**: Most of $\omega_{i,j}$ are exactly zero.

- **Makes empirical sense**: Genes are only assumed to interface with small groups of other genes.
- **Makes statistical sense**: Learning is now feasible in high dimensions with small sample size ($p \gg n$).
How to Fit GGMs?

Penalized likelihood approach

\[ \hat{\Omega}_\lambda = \arg \max_{\Omega} \{ \mathcal{L}(\Omega, X) - \lambda \text{pen}(\Omega) \} \]

pen is a penalty function tuned by \( \lambda > 0 \)

- Regularization (needed when \( p \gg n \)),
- Encourage sparsity (many \( \omega_{i,j} = 0 \)).
- A commonly used penalty function is the \( \ell_1 \) penalty (LASSO)

\[ \text{pen}(\Omega) = \sum_{i \neq j} |\omega_{i,j}| \]
L1 Regularization (Lasso)

Constrained Form

\[ \hat{\Omega}_\lambda = \arg\max_{\Omega} \mathcal{L}(\Omega, X) \]

subject to: \( \sum_{i \neq j} |\omega_{i,j}| \leq C \)

Lagrangian Form

\[ \hat{\Omega}_\lambda = \arg\max_{\Omega} \{ \mathcal{L}(\Omega, X) - \lambda \|\Omega\|_1 \} \]
Gaussian Graphical Models

Penalized likelihood approach

The $\ell_1$ penalized maximum log-likelihood method for GGMs is then to solve (Yuan and Lin, 2007, Banerjee et al., 2007, d’Aspremont et al., 2008, Friedman et al., 2008, Rothman et al 2008, Lam and Fan 2009, etc.)

$$\hat{\Omega} = \arg\max_{\Omega} \{\log \det(\Omega) - \text{tr}(\hat{\Sigma}\Omega) - \lambda \sum_{i \neq j} |\omega_{i,j}|\},$$

where $\hat{\Sigma}$ is sample covariance matrix.
**Gaussian Graphical Models**

**Penalized likelihood approach**

The $\ell_1$ penalized maximum log-likelihood method for GGMs is then to solve (Yuan and Lin, 2007, Banerjee et al., 2007, d’Aspremont et al., 2008, Friedman et al., 2008, Rothman et al 2008, Lam and Fan 2009, etc.)

$$
\hat{\Omega} = \arg\max_{\Omega} \{ \log \det(\Omega) - \text{tr}(\hat{\Sigma}\Omega) - \lambda \sum_{i \neq j} |\omega_{i,j}| \},
$$

where $\hat{\Sigma}$ is sample covariance matrix.

**Statistical Guarantees**

Under certain regulatory conditions, with high probability,

- $\|\hat{\Omega} - \Omega\|_{\infty} = \mathcal{O}(\sqrt{\log p / n})$

- $\|\hat{\Omega} - \Omega\|_{F}^2 = \mathcal{O}_{P}((p + q) \log pn / n)$
GGMs Extension

- Gaussian assumption $\rightarrow$ non Gaussian (Copula).
GGMs Extension

- Gaussian assumption $\rightarrow$ non Gaussian (Copula).
- Gaussian assumption $\rightarrow$ binary (Ising).
GGMs Extension

- Gaussian assumption $\rightarrow$ non Gaussian (Copula).
- Gaussian assumption $\rightarrow$ binary (Ising).
- I.I.D. samples $\rightarrow$ time varying graphical model.
GGMs Extension

- Gaussian assumption → non Gaussian (Copula).
- Gaussian assumption → binary (Ising).
- I.I.D. samples → time varying graphical model.
- Single GGM → multiple GGMs.
GGMs Extension

- Gaussian assumption → non Gaussian (Copula).
- Gaussian assumption → binary (Ising).
- I.I.D. samples → time varying graphical model.
- Single GGM → multiple GGMs.
- Noisy GGM
Multiple Gaussian Graphs

- Data available for multiple related graphs.
- Gene networks describing different cancer subtypes.

Datasets are Independent.

Assume graphs share common structure.
Multiple Gaussian Graphs

Assume independence among datasets

- The joint log likelihood:

\[
\arg\max_{\Omega^{(k)}} \left\{ \sum_{k=1}^{K} \left[ \log \det(\Omega^{(k)}) \right] - \text{tr}(\hat{\Sigma}^{(k)}\Omega^{(k)}) - \sum_{i\neq j} P_\lambda(\omega^{(k)}_{i,j}) \right\}
\]

- Encourage common structure through joint regularization
  (Varoquaux et al. 2007; Guo et al. 2011; Chiquet et al., 2011; Danaher et al., 2012; etc)
Time Varying GGMs