中国科学技术大学
University of Science and Technology of China

Weakly Dependent and Graphical Models

Huiyang Peng

School of management, USTC

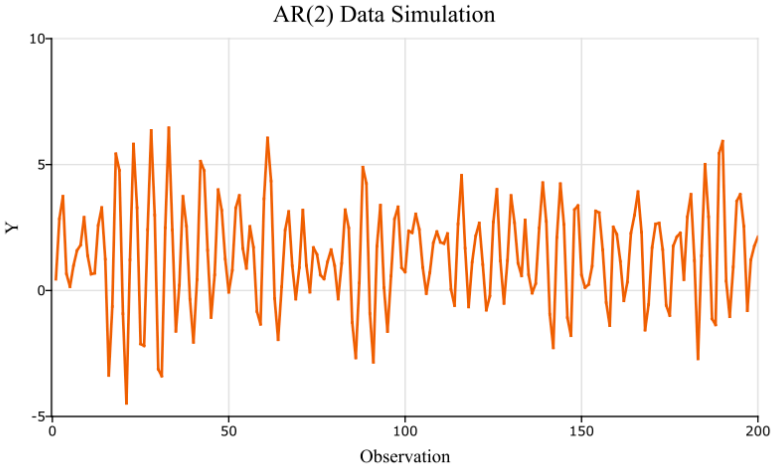
2025 年 3 月 26 日

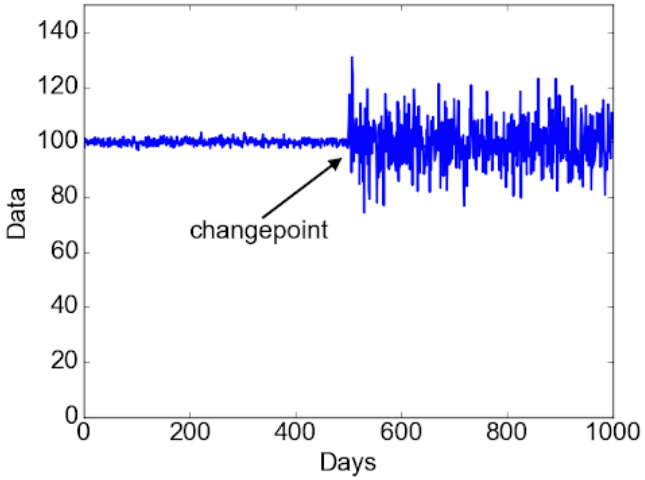
We note that in basic statistics, most statistical theories are built on the assumption that variables are independently and identically distributed (I.I.D.):

- Confidence intervals, hypothesis testing
- Law of large numbers, central limit theorem
- U-statistics
- ...

Observations in time series data are usually correlated, unlike conventional independent data.

- Stock prices
- Sensor data
- Economic growth rates


AR(2) Data Simulation
10
5
0
-5
Y
0
50
100
150
200
Observation


140
120
100
80
60
40
20
0
Data
changepoint
0
200
400
600
800
1000
Days

定义 (Mixing coefficient)

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. A map $c : \mathcal{A} \times \mathcal{A} \to [0, +\infty)$ is called mixing coefficient if for every independent σ-algebra $\mathcal{U}, \mathcal{V} \subset \mathcal{A}$, $c(\mathcal{U}, \mathcal{V}) = 0$.

- Mixing coefficient is some kind of metric.
- The mixing coefficient measures the degree of dependence between two sigma algebras. If the mixing coefficient is 0, it indicates that the two sigma algebras are completely independent.

- Strong mixing coefficient:

$$\alpha(\mathcal{U},\mathcal{V}) = \sup_{U\in\mathcal{U},V\in\mathcal{V}} |\mathbb{P}(U\cap V) - \mathbb{P}(U)\mathbb{P}(V)|$$

- Absolute regularity coefficient:

$$\beta(\mathcal{U},\mathcal{V}) = \frac{1}{2}\sup_{I,J\in\mathbb{N}} \sup_{\substack{(U_i)_{1\le i\le I}\in\mathcal{U}^I \\ (V_j)_{1\le j\le J}\in\mathcal{V}^J}} \sum_{i=1}^{I}\sum_{j=1}^{J} |\mathbb{P}(U_i\cap V_j) - \mathbb{P}(U_i)\mathbb{P}(V_j)|,$$

where $(U_i)_{1\le i\le I}\in\mathcal{U}^I$ and $(V_j)_{1\le j\le J}\in\mathcal{V}^J$ are partitions of Ω.
Or in a more compact way:

$$\beta(\mathcal{U},\mathcal{V}) = ||\mathbb{P}_{\mathcal{U}\otimes\mathcal{V}} - \mathbb{P}_{\mathcal{U}}\otimes\mathbb{P}_{\mathcal{V}}||_{TV}.$$

- Maximal correlation coefficient:

$$\rho(\mathcal{U},\mathcal{V}) = \sup\{|corr(X,Y)| : X \in \mathbb{L}^2(\mathcal{U}), Y \in \mathbb{L}^2(\mathcal{V})\}.$$

- Uniform mixing coefficient:

$$\phi(\mathcal{U},\mathcal{V}) = \sup_{(U,V)\in\mathcal{S}} |\frac{\mathbb{P}(U\cap V)}{\mathbb{P}(U)} - \mathbb{P}(V)|,$$

where $\mathcal{S} = \{(U,V) \in \mathcal{U}\times\mathcal{V} : \mathbb{P}(U) > 0\}$.

- ψ-mixing coefficient:

$$\psi(\mathcal{U},\mathcal{V}) = \sup_{(U,V)\in\mathcal{H}} |\frac{\mathbb{P}(U\cap V)}{\mathbb{P}(U)\mathbb{P}(V)} - 1|,$$

where $\mathcal{H} = \{(U,V) \in \mathcal{U}\times\mathcal{V} : \mathbb{P}(U) > 0, \mathbb{P}(V) > 0\}$.

The strong mixing coefficient for a random process $X = \{X_t\}_{t\in\mathbb{Z}}$ is defined by, for $r > 0$,

$$\alpha(r) := \alpha_X(r) = \sup_{i\in\mathbb{Z}} \alpha(\sigma(X_t, t \le i), \sigma(X_t, t \ge i + r)).$$

The random process $\{X_t\}_{t\in\mathbb{Z}}$ is called α-mixing if $\alpha(r) \to 0$ as $r \to \infty$. Accordingly, we shall use below the notations $\beta(r), \rho(r), \phi(r)$ and $\psi(r)$, which mean sequences defined with α replaced by β, ρ, ϕ and ψ, respectively.
Those conditions are related to the following diagram

$$\psi\text{-mixing} \Rightarrow \phi\text{-mixing} \Rightarrow \begin{matrix}\rho\text{-mixing} \\ \beta\text{-mixing}\end{matrix} \Rightarrow \alpha\text{-mixing}.$$

定义 (Stationarity and strongly stationary)

The time series $\{X_t, t \in \mathbb{Z}\}$ is called stationary time series when

1. $E|X_t|^2 < \infty, \forall t \in \mathbb{Z}$
2. $EX_t = m, \forall t \in \mathbb{Z}$
3. $\gamma_X(r, s) = \gamma_X(r + t, s + t), \forall r, s, t \in \mathbb{Z}$

Moreover, it's called strongly stationary when $\forall k \in \mathbb{N}_+$, $\forall t_1, \cdots, t_k, h \in \mathbb{Z}$,

$$(X_{t_1}, \cdots, X_{t_n}) \stackrel{d}{=} (X_{t_1+h}, \cdots, X_{t_n+h}).$$

Then for the sake of convenience, we define the autocovariance function as univariate:

$$\gamma_X(h) = \gamma_X(h, 0) = Cov(X_{t+h}, X_t)$$

定义 (White Noise)

If the process $\{Z_t\}$ has a mean of 0 and an autocovariance function of

$$\gamma_Z(h) = \begin{cases} \sigma^2, & h = 0 \\ 0, & h \neq 0 \end{cases}$$

then the process $\{Z_t\}$ is called a zero-mean white noise with variance σ^2, denoted as $\{Z_t\} \sim WN(0, \sigma^2)$.

White noise is not necessarily independently and identically distributed (I.I.D.).

If $\{Z_t\}$ is a zero-mean I.I.D. sequence with variance σ^2, it is denoted as $\{Z_t\} \sim IID(0, \sigma^2)$.

定义 (ARMA)

An AutoRegressive Moving Average (ARMA) sequence $\{X_t\}$ is a time series that satisfies the following equation:

$$X_t - \sum_{i=1}^{p} \phi_i X_{t-i} = Z_t + \sum_{j=1}^{p} \theta_i Z_{t-i},$$

where $\{Z_t\} \sim WN(0, \sigma^2)$.
It can be concisely written as $\phi(B)X_t = \theta(B)Z_t$, where

$$\phi(z) = 1 - \sum_{i=1}^{p} \phi_i z^i$$
$$\theta(z) = 1 + \sum_{i=1}^{p} \theta_i z^i.$$

定义 (Causality)

An $ARMA(p,q)$ process defined by the equations $\phi(B)X_t = \theta(B)Z_t$ is said to be *causal* if there exists a sequence of constants $\{\psi_j\}$ such that $\sum_{j=0}^{\infty} |\psi_j| < \infty$ and

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$$

定义 (Invertibility)

An $ARMA(p,q)$ process defined by the equations $\phi(B)X_t = \theta(B)Z_t$ is said to be *invertible* if there exists a sequence of constants $\{\pi_j\}$ such that $\sum_{j=0}^{\infty} |\pi_j| < \infty$ and

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}$$

Note that causality and invertibility are not properties of $\{X_t\}$ alone, but of the relationship between the two processes $\{X_t\}$ and $\{Z_t\}$.

定理 (Causality condition of ARMA)

Consider an ARMA(p,q) process: $\phi(B)X_t = \theta(B)Z_t$, for which the polynomials $\phi(\cdot)$ and $\theta(\cdot)$ have no common zeroes. Then $\{X_t\}$ is causal ***if and only if***

$$\phi(z) \neq 0, \forall z \in \mathbb{C}, |z| = 1.$$

定理 (Invertibility condition of ARMA)

Consider an ARMA(p,q) process: $\phi(B)X_t = \theta(B)Z_t$, for which the polynomials $\phi(\cdot)$ and $\theta(\cdot)$ have no common zeroes. Then $\{X_t\}$ is invertible ***if and only if***

$$\theta(z) \neq 0, \forall z \in \mathbb{C}, |z| = 1.$$

证明.

Theorem 3.1.1 and Theorem 3.1.2 in Reference [1]. □

ARMA is not necessarily a stationary time series.

定理 (Stationarity condition of ARMA)

Consider an ARMA(p,q) process: $\phi(B)X_t = \theta(B)Z_t$. If

$$\phi(z) = 1 - \sum_{i=1}^{p} \phi_i z^i \neq 0, \forall z \in \mathbb{C}, |z| = 1,$$

then the ARMA(p,q) equations have the unique stationary solution

证明.

Theorem 3.1.3 in Reference [1].

In fact it's a sufficient and necessary condition, refer to (3.1.4) in Reference [2]. □

定理 (Augmented Dickey–Fuller test)

AutoRegressive Model:

$$\Delta X_t = \alpha + \beta t + \gamma X_{t-1} + \delta_1 \Delta X_{t-1} + \cdots + \delta_{p-1} \Delta X_{t-p+1} + \epsilon_t$$

$H_0 : \gamma = 0, H_1 : \gamma < 0.$

A value for the test statistic

$$DF_\tau = \frac{\hat{\gamma}}{SE(\hat{\gamma})}.$$

定理 (Kwiatkowski-Phillips-Schmidt-Shin test)

Model:

$$X_t = \beta t + r_t + \epsilon_t,$$

where r_t *is a random walk, displayed as*

$$r_t = r_{t-1} + e_t, e_t \sim IID(0, \sigma_e^2).$$

$H_0 : \sigma_e^2 = 0, H_1 : \sigma_e^2 > 0.$

A value for the test statistic

$$LM = \frac{\sum_{t=1}^{T} S_t^2}{T^2 \hat{\sigma_\epsilon^2}}, S_t = \sum_{i=1}^{t} e_t.$$

As an example, consider the linear process

$$X_t = \sum_{j=0}^{+\infty} a_j Z_{t-j}, t \in \mathbb{Z}$$

where $a_j = O(e^{-rj}), r > 0$, and Z_t are independent zero-mean real random variables with a common density and finite second moment.
Then $\{X_t\}$ is ρ-mixing and therefore α-mixing with coefficients which decrease to zero at an exponential rate.

And we note that the condition Z_t **has a common density** is crucial, next we provide an example for that.

Consider the process:

$$X_t = \sum_{j=0}^{+\infty} 2^{-j-1} Z_{t-j}, t \in \mathbb{Z},$$

where the Z_ts are independent with common distribution $\mathcal{B}(1, \frac{1}{2})$.
So $X_t \sim U(0,1)$ and $2X_{t+1} = X_t + Z_{t+1}$, so X_t is the fractional part of $2X_{t+1}$, hence $\sigma(X_t) \subset \sigma(X_{t+1})$, and we get

$$\sigma(X_t) \subset \sigma(X_s, s \geq t+k),$$

thus

$$\alpha_k \geq \alpha(\sigma(X_t), \sigma(X_s)) \geq \alpha(\sigma(X_t), \sigma(X_t)) \geq \frac{1}{4},$$

which proves that $\{X_t\}_{t\in\mathbb{Z}}$ is not α-mixing.

定理 (Ibragimov, 1962)

Let $r \geq p \geq 1$ *and let* $\mathcal{F}_{-\infty}^{t} = \sigma(\cdots, X_{t-2}, X_{t-1}, X_t)$, *then*

$$||E[X_{t+n}|\mathcal{F}_{-\infty}^{t}] - E[X_{t+n}]||_p \leq 2(2^{1/p}+1)\alpha_n^{1/p-1/r}||X_{t+n}||_r$$

A corollary of the theorem is:

推论

For $s > 1$ *and* $p \geq s/(s-1)$,

$$|Cov(X_t, X_{t+n})| \leq 2(2^{1-1/s}+1)\alpha_n^{1-1/s-1/p}||X_t||_s||X_{t+n}||_p$$

Note that, by the assumptions of the corollary:

$$1/s < 1,\ 1/p + 1/s \leq 1$$

In the case that we take $p > s/(s-1)$, we would have $1 - 1/s - 1/p > 0$.
It follows that in the case of α-mixing, where $\alpha_n \to 0$ as $n \to \infty$, we would have

$$|Cov(X_t, X_{t+n})| \to 0 \text{ as } n \to \infty$$

assuming of course that the moment condition $||X_t||_\gamma < \infty$ for some $\gamma \geq \max\{p, s\}$.

定理 (McLeish, 1975)

Let $\{X_t\}$ *be a scalar stochastic sequence with finite means* $\mu_t = E[X_t]$ *and with* $\phi(n) = O(n^{-r/(2r-2)})$ *for* $r \geq 2$ *or* $\alpha(n) = O(n^{-r/(r-2)})$ *for* $r > 2$. *In addition, suppose that*

$$\sum_{t=1}^{\infty} \left(\frac{E|X_t - \mu_t|^p}{t^p}\right)^{2/r} < \infty$$

for some p *such that* $r/2 < p \leq r < \infty$. *Under these conditions,*

$$\frac{1}{n}\sum_{t=1}^{n}(X_t - \mu_t) \xrightarrow{a.s.} 0 \text{ as } n \to \infty.$$

Observation 1: the requirements for α-mixing and ϕ-mixing are not the same. Since $r > 2$, $-\frac{r}{2r-2} > -\frac{r}{r-2}$, the theorem requires a stronger constraint on $\alpha(n)$ than $\phi(n)$(since ϕ-mixing is stronger than α-mixing).

Observation 2: A sufficient condition for $\sum_{t=1}^{\infty}(\frac{E|X_t-\mu_t|^p}{t^p})^{2/r} < \infty$ is simply that

$$\sup_t E|X_t - \mu_t|^p < \infty$$

in which case

$$\sum_{t=1}^{\infty}(\frac{E|X_t-\mu_t|^p}{t^p})^{2/r} \leq (\sup_t E|X_t - \mu_t|^p)^{2/r} \sum_{t=1}^{\infty} \frac{1}{t^{2p/r}} < \infty$$

given that $\sum_{t=1}^{\infty} \frac{1}{t^{2p/r}}$ is a convergent series.

定理 (Denker, 1986)

Assume $\{X_k\}_{k\in\mathbb{Z}}$ *is a strictly stationary and strongly mixing sequence such that*

$$EX_0 = 0, EX_0^2 < \infty \text{ and } Var(S_n) \to \infty.$$

Then

$$\frac{S_n}{\sqrt{S_n}} \xrightarrow{d} N(0,1)$$

is equivalent to

$$\{\frac{S_n^2}{Var(S_n)}\}_{n\geq 1} \text{ is an uniformly integrable family.}$$

Another classic result is(without stationary):

定理 (White, 2001)

Let $\{X_{tn}\}$ *be a triangular array of scalars with* $\mu_{tn} = E[X_{tn}]$ *and* $\sigma_{tn}^2 = Var(X_{tn})$ *such that* $||X_{tn}||_r \leq C < \infty$ *for some* $r \geq 2$ *and for all* t *and* n. *In addition, suppose that* $\{X_{tn}\}$ *has mixing coefficient* $\phi(n) = O(n^{-r/2(r-1)})$ *or* $\alpha(n) = O(n^{-r/(r-2)})$ *for* $r > 2$. *Furthermore, suppose that*

$$\bar{\sigma}_n^2 = Var(\frac{1}{\sqrt{n}} \sum_{t=1}^{n} X_{tn}) \geq \delta > 0$$

for all n *sufficiently large. Under these conditions, as* $n \to \infty$,

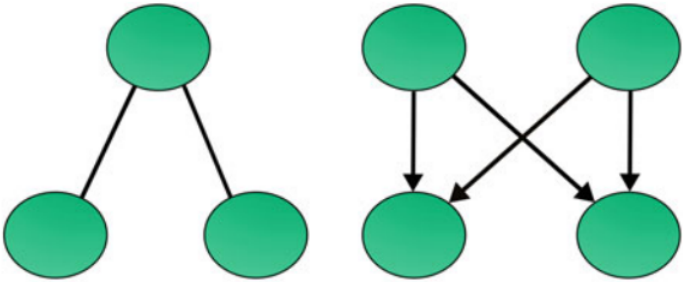
$$\frac{1}{\bar{\sigma}_n \sqrt{n}} \sum_{t=1}^{n} (X_{tn} - \mu_{tn}) = \frac{\sqrt{n}(\bar{X}_n - \bar{\mu_n})}{\bar{\sigma}_n} \xrightarrow{d} N(0,1)$$

In statistics, due to different starting points on **uncertainty**, schools of thought are divided into **Frequentist** and **Bayesian**.
Consider a probability model: $X \sim P(x|\theta)$.

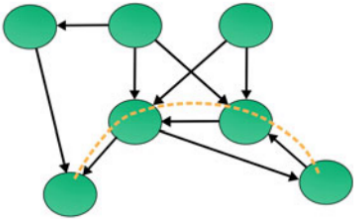
- Frequentist school: θ is an unknown constant and X is a random variable→Estimating parameter θ using the Maximum Likelihood Estimation (MLE) method → Machine Learning → Optimization.
- Bayesian school: $\theta \sim P(\theta)$, called the prior → Bayes' theorem → Bayesian estimation and Maximum Posterior estimation (MAP) → Probabilistic graphical models.

Conclusion: Probabilistic graphical models are rooted in Bayesian school theory.

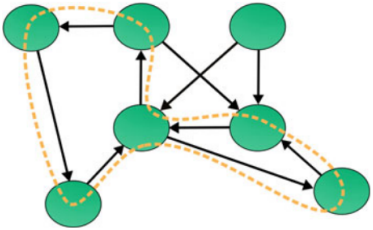
Graph: V is a non empty set, a binary relation $E \subset V \times V$, $G = (V, E)$.
Directed graph: a binary relation.
Undirected graph: symmetric binary relations.



Trajectories

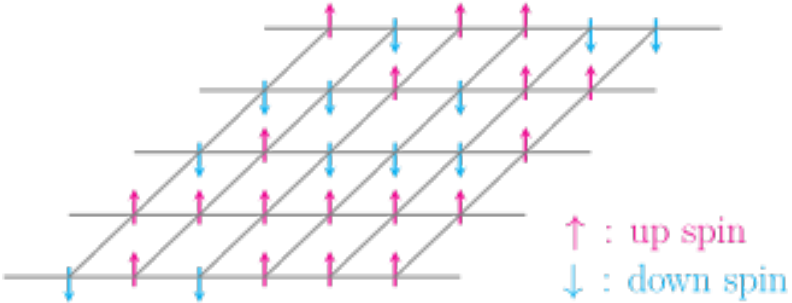


Circuits



- **DAG(Directed Acyclic Graph):** a directed graph that has no directed circuits.
- **Parent / Child:** if there is a directed arc from A to B, A is parent of B and B is a child of A.
- **Ascendants / Descendants:** if there is a directed trajectory from A to B, A is an ascendant of B and B is a descendant of A.

The Ising model describes a lattice system where each site has a spin that can exist in two states: up (+1) or down (-1). There are interactions between neighboring spins, typically favoring alignment in the same direction or in opposite directions. Additionally, an external magnetic field may influence the orientation of these spins.



Let's abstract the structure of the Ising model:

- Random field(RF): a collection of S random variables, $\mathbf{F} = F_1, \cdots, F_S$.
- Markov random field(MRF): a random field that satisfies the locality property(Markov property).

定义 (Locality/Markov property)

A variable F_i is independent of all other variables in the field given its neighbors, $Nei(F_i)$. That is:

$$P(F_i|\mathbf{F_c}) = P(F_i|Nei(F_i)),$$

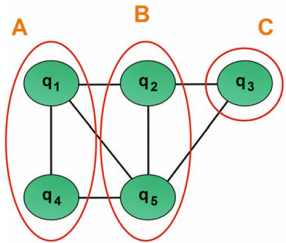
where $\mathbf{F_c}$ is the set of all random variables in the field except F_i.

A MRF is an undirected graphical model which consists of:

- $\mathbf{V}$: a set of random variables
- $\mathbf{E}$: a set of undirected edges

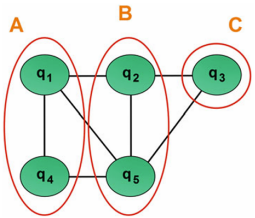
These form an undirected graph that represents the independency relations:

- A subset of variables $\mathbf{A}$ is independent of the subset of variables $\mathbf{C}$ given $\mathbf{B}$, if the variables in $\mathbf{B}$ separate $\mathbf{A}$ and $\mathbf{C}$ in the graph.



The joint probability of a MRF can be expressed as the product of local functions on subsets of variables. These subsets should include, at least, all the *maximal cliques* in the network.

$$P(q_1, q_2, q_3, q_4, q_5) = \frac{1}{k} P(q_1, q_4, q_5) P(q_1, q_2, q_5) P(q_2, q_3, q_5)$$



OR

$$P(q_1, q_2, q_3, q_4, q_5) = \frac{1}{k} P(q_1, q_4, q_5) P(q_1, q_2, q_5) P(q_2, q_3, q_5)$$
$$P(q_1, q_2) P(q_1, q_4) P(q_1, q_5) P(q_2, q_3) P(q_2, q_5) P(q_3, q_5) P(q_4, q_5)$$

Denote $MC(G)$ to be the collection of *maximal cliques* of graph G. The joint probability distribution of a MRF can be factorized as

$$P(X) = \frac{1}{k} \prod_{C \in MC(G)} \phi_C(X_C)$$

where k is a normalizing constant and ϕ_C is a local function over the variables in the corresponding clique C.

Gibbs Random Fields(GRF) is a equivalent expression. For each $C \in MC(G)$, we define $U_C(X_C) = -\log \phi_C(X_C)$, and $\mathbf{U_F} = \sum_{C \in MC(G)} U_C(X_C)$, then

$$P(X) = \frac{1}{k} \exp(-\mathbf{U_F}).$$

- For a physical system, states with lower energy are more likely to occur, as lower energy situations are considered more stable.
- For a statistical model, states with higher likelihood functions are more likely to occur.
- Now we connect these concepts.

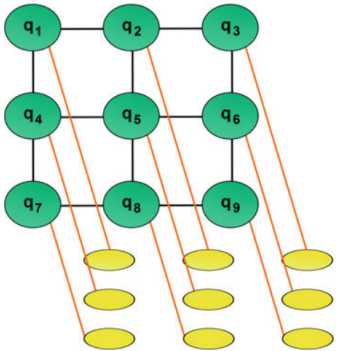
We define $U_C(X_C) = -\log \phi_C(X_C)$ as the local *potentials*, then $\mathbf{U_F} = \sum_{C \in MC(G)} U_C(X_C)$ is the total energy.

Given the Gibbs equivalence, the problem of finding the maximum probability for a MRF is transformed to find the minimum energy.

- Digital images are usually corrupted by high frequency noise.
- MRF can be applied to the image for reducing the noise.

Model: each pixel corresponds to a random variable, and each interior variable is connected to its 4 neighbors.

It can be easily observed that every *maximal cliques* in the graph contains two nodes.



The energy function, in this case, can be expressed as the sum of two types of potentials:

- one associated to pairs of neighbors, $U_c(q_i, q_j)$
- the other for each variable and its corresponding observation, $U_o(q_i, p_i)$

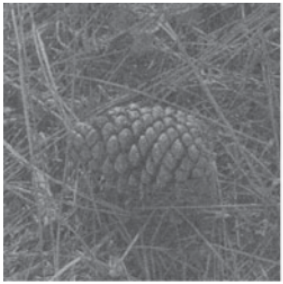
Now we consider how to define the local potentials:

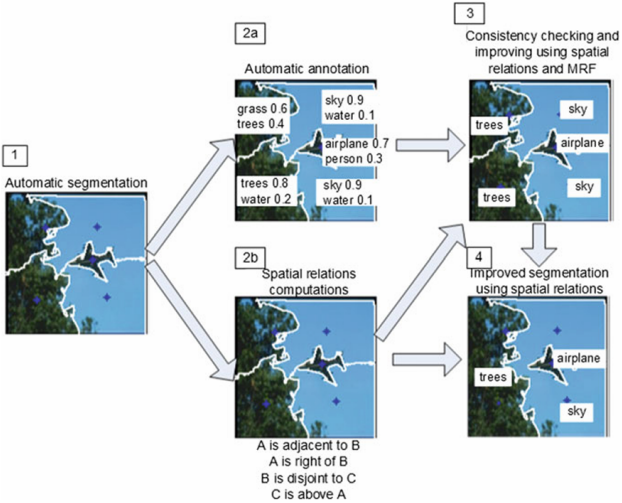
- Smooth: punishing (higher energy) configurations in which neighbors have different values, $U_c(q_i, q_j) = (q_i - q_j)^2$
- Each variable in the MRF should have a value similar to the one in the original image, $U_o(q_i, p_i) = (q_i - p_i)^2$

Then, the energy

$$\mathbf{U_F} = \sum_{i,j} U_c(q_i, q_j) + \lambda \sum_i U_o(q_i, p_i)$$

Using these potentials and applying the stochastic optimization algorithm, a smoothed image is obtained.
Left: original image; center: $\lambda = 1$; right: $\lambda = 0.5$.




1
Automatic segmentation
2a
Automatic annotation
grass 0.6
trees 0.4
sky 0.9
water 0.1
airplane 0.7
person 0.3
trees 0.8
water 0.2
sky 0.9
water 0.1
2b
Spatial relations computations
A is adjacent to B
A is right of B
B is disjoint to C
C is above A
3
Consistency checking and improving using spatial relations and MRF
trees
sky
airplane
trees
sky
4
Improved segmentation using spatial relations
airplane
trees
sky

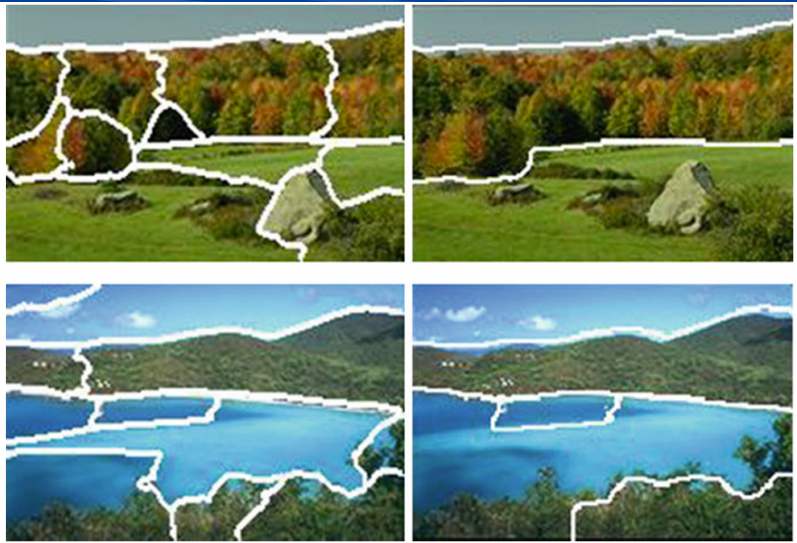
1. An image is automatically segmented (using Normalized cuts).
2. The obtained segments are assigned a list of labels and their corresponding probabilities based on their visual features using a classifier.
3. Concurrently, the spatial relations among the same regions are computed.
4. The MRF is applied, combining the original labels and the spatial relations, resulting in a new labeling for the regions by applying simulated annealing.
5. Adjacent regions with the same label are joined.

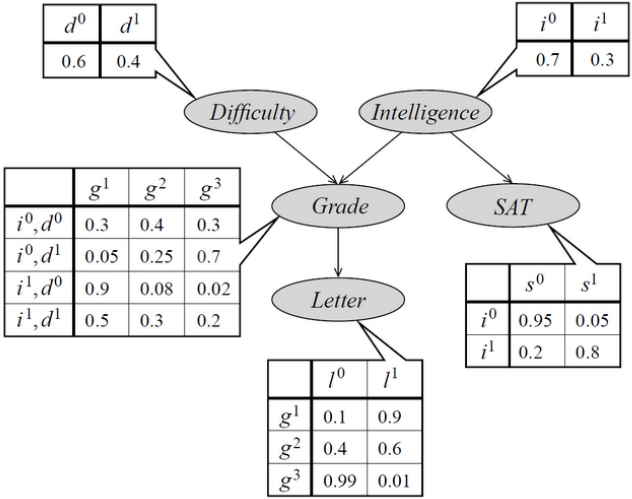
Here, spatial relations are divided in three groups: topological relations, horizontal relations and vertical relations. Thus, the energy function

$$U_p(f) = \alpha_1 V_T(f) + \alpha_2 V_H(f) + \alpha_3 V_V(f) + \lambda \sum_o V_o(f),$$

where V_T is the potential for topological relations, V_H for horizontal relations, and V_V for vertical relations.

These potentials can be estimated from a set of labeled training images.
The potential for a certain type of spatial relation between two regions of classes A and B is inversely proportional to the probability(frequency) of that relation occurring in the training set.




d^0 d^1
0.6 0.4
Difficulty
i^0 i^1
0.7 0.3
Intelligence
Grade
SAT
Letter
g^1 g^2 g^3
i^0, d^0 0.3 0.4 0.3
i^0, d^1 0.05 0.25 0.7
i^1, d^0 0.9 0.08 0.02
i^1, d^1 0.5 0.3 0.2
s^0 s^1
i^0 0.95 0.05
i^1 0.2 0.8
l^0 l^1
g^1 0.1 0.9
g^2 0.4 0.6
g^3 0.99 0.01

A Bayesian network (BN) represents the joint distribution of a set of n variables, $X_1, X_2, \cdots, X_n$, as a directed acyclic graph(DAG) and a set of conditional probability tables(CPT).

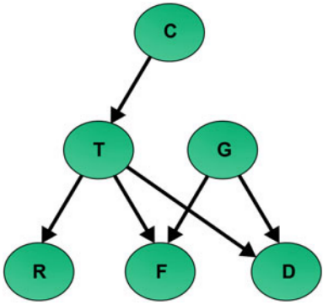
Each node, that corresponds to a variable, has an associated CPT that contains the probability of each state of the variable given its parents in the graph.

The structure of the network implies a set of conditional independence.

- An example of a simple BN.
- The structure of the graph implies a set of conditional independence.

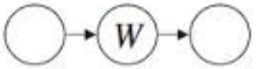
For example, R is conditionally independent of C, G, F, D given T, that is:

$$P(R|C, T, G, F, D) = P(R|T).$$

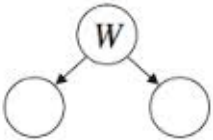


We first consider the 3 basic BN structures for 3 variables and 2 arcs:

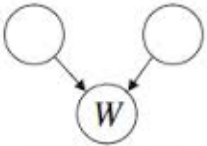
- Sequential: $X \to Y \to Z$.
- Divergent: $X \leftarrow Y \to Z$.
- Convergent: $X \to Y \leftarrow Z$.



Sequential



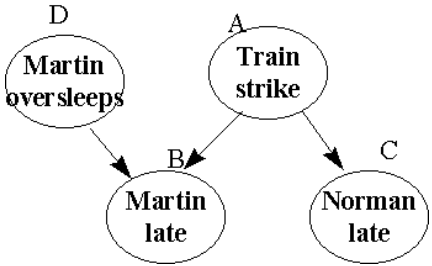
Divergent



Convergent

Now we consider the conditionally independence in the 3 basic BN structures.

- Sequential: $X \to Y \to Z$. X and Z are conditionally independent given Y.
- Divergent: $X \leftarrow Y \to Z$. Same.
- Convergent: $X \to Y \leftarrow Z$. This case is called *explaining away*: knowing the effect and one of the causes, alters our belief in the other cause.



Next we introduce two algorithms to decide the conditional independence in a Bayesian Network.

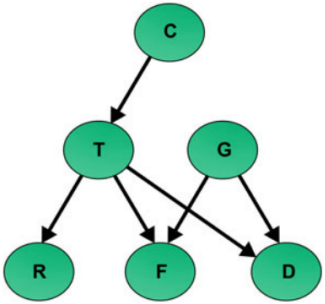
D-Separation:

Given a graph G, a set of variables A is conditionally independent of a set B given a set C, if there is no path in G between A and B such that

1. All convergent nodes are or have descendants in C.
2. All other nodes are outside C.

Consider R, T, C, there's only one trajectory $C \to T \to R$, it has no convergent nodes, so it doesn't satisfly condition 2. So R is independent of C given T.

Consider T, F, G, there exists a trajectory $T \to F \leftarrow G$ such that F is a convergent, and $T, G \notin \{F\}$, so T and G are not independent given F.



Another way to verify D-Separation is by using an algorithm known as the *Bayes ball*.

Bayes ball:

Considering that there is a path from node X to Z with Y in the middle, Y is shaded if it is known, otherwise it is not shaded.
We **throw a ball** from X to Z, if the ball arrives to Z then X and Z are not independent given Y, according to the following rules:

1. If Y is sequential or divergent and is not shaded, the ball goes through.
2. If Y is sequential or divergent and it is shaded, the ball is blocked.
3. If Y is convergent and not shaded, the ball is blocked.
4. If Y is convergent and shaded, the ball goes through.

定理 (Markov assumption)

Any node X is conditionally independent of all nodes in G that are not descendants of X given its parents in the graph, $Pa(X)$.

证明.

We use the **Bayes ball** for proof.

When X pass the ball out, the node can only pass it to its parents or children.

- If X pass the ball to one of its parents, the ball is blocked immediately, since the parent is shaded and can's be convergent(1 arc point to X).
- If X pass the ball to one of its children, if each node passes the ball to its children, then the final receiving node must be a descendant of node X, which leads to a contradiction. Therefore, there must be at least one backward arrow, which results in a convergent node that is not shaded, the ball is blocked here.

□

The structure of the BN can be specified as:

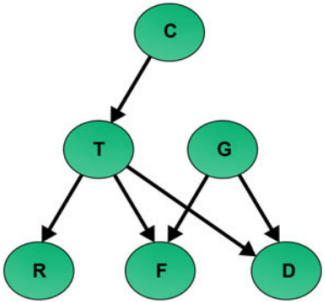
- $Pa(C) = \emptyset, Pa(G) = \emptyset$
- $Pa(T) = \{C\}, Pa(R) = \{T\}$
- $Pa(F) = \{T, G\}, Pa(D) = \{T, G\}$

Given this condition and using the chain rule, similar to MRF:

$$P(X_1, \cdots, X_n) = \prod_{i=1}^{n} P(X_i | Pa(X_i)).$$

For the example:

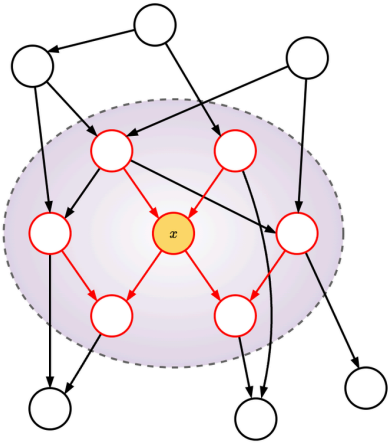
$P(C, T, G, R, F, D) = P(C)P(G)P(T|C)P(R|T)P(F|T, G)P(D|T, G)$



The *Markov Blanket* of a node X, $MB(X)$, is a set of nodes that make it independent of all the other nodes in G, that is $P(X|G-X) = P(X|MB(X))$. For a BN, the Markov blanket of X is:

- the parents of X
- the sons of X
- and other parents of the sons of X

For instance, in the BN mentioned above, the Markov blanket of R is T and the Markov blanket of T is C, R, F, D, G.


x

Consider for example an intensive care unit of a hospital in which sensors monitor the status of an operated patient so that the body temperature is kept beneath certain levels.
Given that the sensors are working constantly, there is potential for them to produce erroneous readings. If this happens two situations may arise:

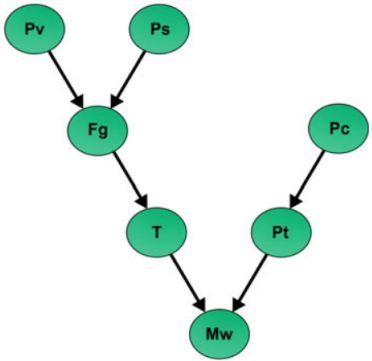
- the temperature sensor indicates no changes in temperature even if it has increased to dangerous levels
- the temperature sensor indicates a dangerous level even if it is normal

Two steps:

- Fault Detection: to find out what's wrong
- Fault Isolation: to find out what's the exact reason

A basic probabilistic model of a gas turbine.

- M_w: the mega watts generated in a gas turbine
- T: temperature
- P_t: pressure in the turbine
- F_g: the flow of gas
- P_v: the valve of gas position
- P_s: the gas fuel pressure supply
- P_c: the pressure at the output of the compressor



We can obtain probability distributions for each value of T.
If the real observed value coincides with a *valid value*—that has a *high* probability, then the sensor is considered correct; otherwise it is considered faulty.
This procedure is repeated for all the sensors in the model.

However, if a validation of a single sensor is made using a faulty sensor, then a faulty validation can be expected.
This is called an *apparent fault*: which is the real faulty sensor?

An isolation stage is needed.

Recall the *Markov basket*: $MB(T) = \{F_g, M_w, P_t\}$.
Additionally, we define the *Extended Markov Blanket* of a node X:
$EMB(X) = \{X\} \cup MB(X)$.

Using this property, if a fault exists in one of the sensors, it will be revealed in all of the sensors in its EMB.

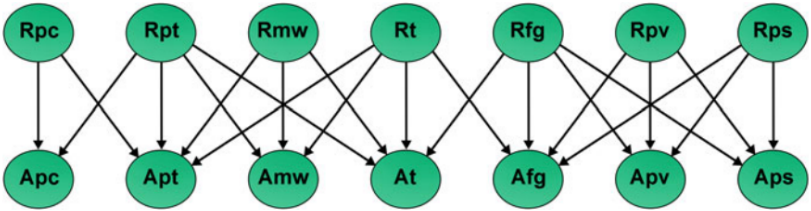
We utilize the EMB to create a fault isolation module that distinguishes the *real faults* from the apparent faults.

After a cycle of basic validations of all sensors is completed, a set S of apparent faulty sensors is obtained.

1. If $S = \emptyset$ there are no faults.
2. If $S = EMB(X)$, and there is no other EMB which is a subset of S, then there is a *single real fault* in X.
3. If $S = EMB(X)$, and there are one or more EMBs which are subsets of S, then there is a real fault in X, and possibly, real faults in the sensors whose EMBs are subsets of S.(*multiple indistinguishable* real faults)
4. If S is equal to the union of several EMBs and the combination is unique, then there are *multiple distinguishable* real faults in all the sensors whose EMB are in S.
5. If none of the above cases is satisfied, then there are multiple faults but they can not be distinguished. All the variables whose EMBs are subsets of S could have a real fault.

Isolation network: Each node X above, connected to $EMB(X)$ below.

- $S = \{T, P_t, M_w\}$, case 2, $\rightarrow M_w$
- $S = \{T, P_c, P_t, M_w\}$, case 3, $\rightarrow P_t$, possibly P_c and M_w
- $S = \{P_v, P_s, F_g\}$, case 4, $\rightarrow P_v$ and P_s



1. Brockwell, Peter J., and Richard A. Davis. Time series: theory and methods. Springer science & business media, 1991.
2. Brockwell, Peter J., and Richard A. Davis, eds. Introduction to time series and forecasting. New York, NY: Springer New York, 2002.
3. John C. Chao. Lecture Notes on Mixing Processes and Martingale Difference Arrays. Lecture note
4. Sucar, Luis Enrique. Probabilistic graphical models. Springer International Publishing, 2021.

Thanks!