Local Probabilistic Models: Independence of Causal Influence

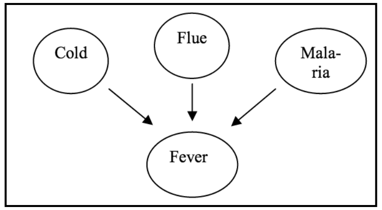
Sargur Srihari
srihari@cedar.buffalo.edu

Topics

- Local Probabilistic Models
  - Independence of Causal Influence
    - Noisy-OR
    - Generalized Linear Models

Independence of Causal Influence

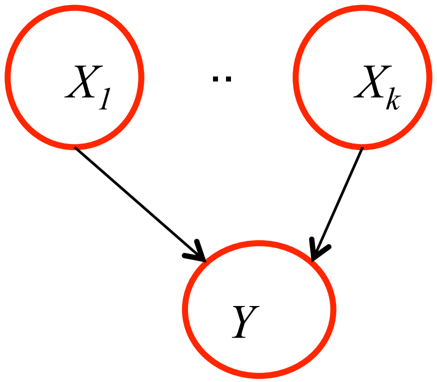
- Very different type of local probability model
- Consider variable Y whose distribution depends on some set of causes $X_1,..X_k$
  - Y can depend on its parents in arbitrary ways



  - If we don't assume independence, we have 2^k possible values for parents
- Assume each parent has an independent influence and their influence is combined in some way

Combining Causal Influence

- Distribution of variable Y depends on several causes $X_1,..,X_k$
- Each parent has an independent influence and their influence is combined



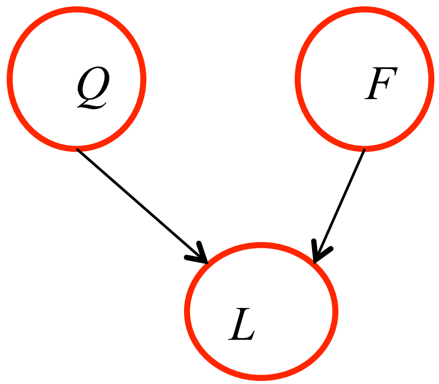
- Two types: *Noisy-or* and *Generalized-Linear*

From Or to Noisy-Or

- Small seminar course where Professor gets to know each student
- Good Letter ($L=l^1$) depends on two things:
  - class participation (asking good questions, $Q = q^1$)
  - good final paper ($F = f^1$)
  - Each event is enough to write good letter

Deterministic CPD
(Or Without Noise)

Q	F	$l0$	$l1$	
$q0$	$f0$	1	0	Bad Letter
$q0$	$f1$	0	1	Good Letter
$q1$	$f0$	0	1	Good Letter
$q1$	$f1$	0	1	Good Letter



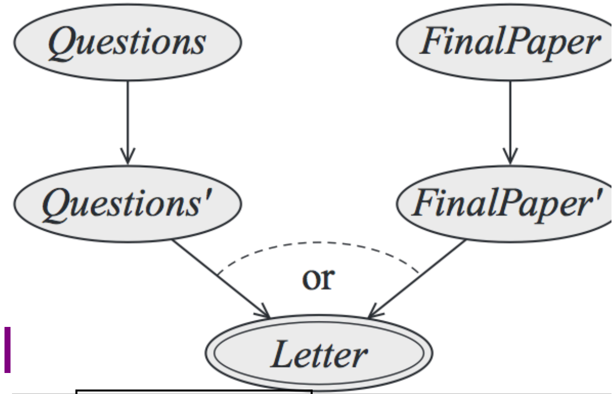
Noisy Or Example

- Professor fails to remember student’s participation
- Professor may not be able to read student’s handwriting and may not appreciate the quality of the final paper
- So there is noise in the process

Noise Parameters

- *Q: Good Questions, But Teacher is Forgetful*
- $P(l^1 \mid q^1, f^0) = 0.8$ Prob good Q *in isolation* causes good L is 0.8
- *F: Good Final Paper, But Poor Handwriting*

$P(l^1 \mid q^0, f^1) = 0.9$ Good F causes good L



- What if both *good Q, good F*
  - Independent causal mechanisms
    - Letter Weak only if neither successful
    - Both q^1 and f^1 occur with prob $0.2 \times 0.1 = 0.02$
    - Noise parameters
      - $\lambda_Q = P(q'^1 \mid q^1) = 0.8$
      - $\lambda_F = P(f'^1 \mid f^1) = 0.9$

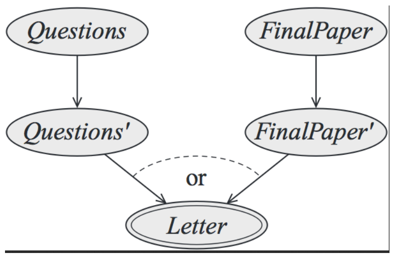
$P(L|Q,F)$

Q	F	l^0	l^1	
q^0	f^0	1	0	Bad Letter
q^0	f^1	0.1	0.9	Good Letter
q^1	f^0	0.2	0.8	Good Letter
q^1	f^1	0.02	0.98	Good Letter

- If both are bad, q^0, f^0, then we still get bad $L = l^0$

Leak Probability

- Professor writes a good recommendation letter for no good reason with probability *0.0001*
  - Because Professor is having a good day
- Introduce another parent of Letter variable to represent this event
  - This variable has no parents and is True with probability $\lambda_0 = 0.0001$
  - It is also a parent of the $Letter$ variable which remains a deterministic Or

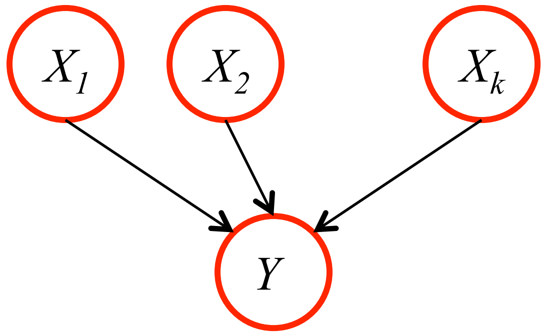


General Definition of Noisy-Or

- Let Y be a binary-valued r.v. with parents $X_1,..X_k$
- The CPD $P(Y|X_1,..X_k)$ is a *noisy-or* if there are $k+1$ parameters $\lambda_0, \lambda_1,..\lambda_k$ such that

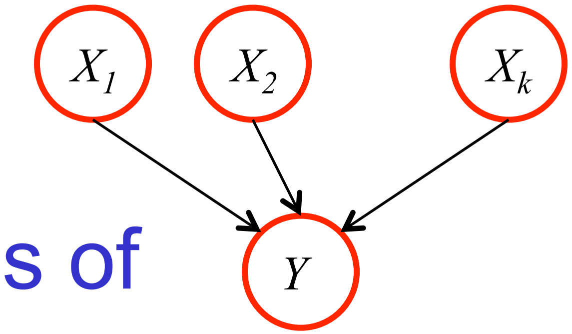
$$P\left(y^0 \mid X1,..Xk\right) = (1-\lambda_0)\prod_i (1-\lambda_i)$$

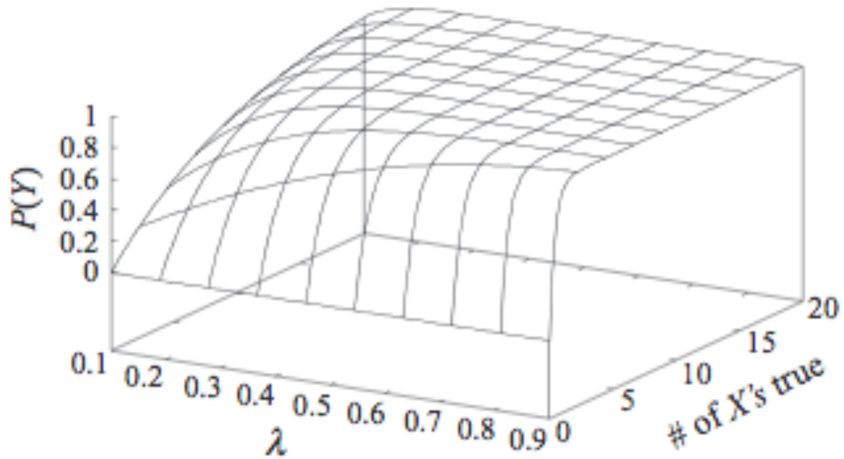
$$P\left(y^1 \mid X1,..Xk\right) = 1-[(1-\lambda_0)\prod_i (1-\lambda_i)\ ]$$



Behavior of Noisy-Or

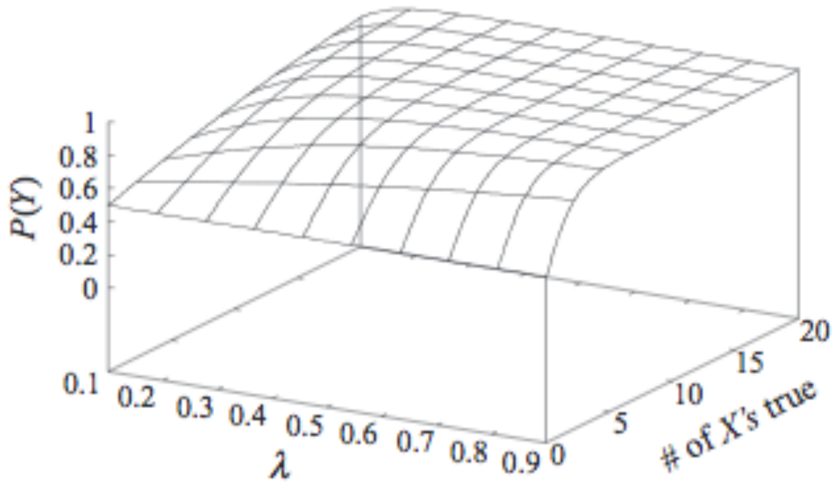
- All variables have same noise parameter λ
- Probability of child $Y=y^1$ in terms of
  - λ and number of X_i that have value true





(a)

Leak Probability of 0



(b) Leak Probability of 0.5
Higher $P(Y)$

Applicability of Noisy-Or

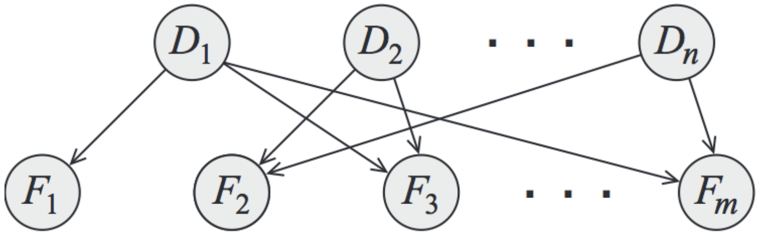
- Applicable in a wide variety of settings
- Most obvious is the medical domain
- A symptom variable such as Fever has a very large number of parents (Diseases)
- It is reasonable to assume that different diseases have different causal mechanisms
- If any disease succeeds in activating its mechanism, the symptom is present

BN2O Network

- A class of networks that has received attention in medical diagnosis is the class of BN2O networks
- It is a two-layer network where the top layer corresponds to a set of causes, such as diseases, and the second to findings that might indicate the causes, such as symptoms or test results

BN2O

- 2-Layer Noisy-Or BN for Medical Diagnosis



- BN2O Top layer: causes
  - diseases: flu, pneumonia, etc
- BN2O Bottom layer: findings
  - symptoms (caughing, sneezing), test results
  - All variables in lower layer are Noisy-Or
  - CPD of F_i is given by

$$P\left(f_i^0 \mid \mathrm{Pa}_{F_i}\right) = (1-\lambda_{i,0}) \prod_{Dj \in \mathrm{Pa}_{F_i}} (1-\lambda_{i,j})^{d_j}$$

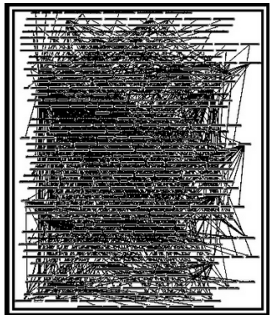
    - Where $\lambda_{i,j}$ is probability that d_i in isolation causes f_j

Properties of BN2O

- Conceptually very simple
  - Need a small no. of easy-to-understand parameters
  - Each edge is causal: cause d_i and finding f_i
  - Each has parameter $\boldsymbol{\lambda}_{i,j}$
    - probability that d_i in isolation causes f_i to manifest
- In practice few symptoms present (many false)
  - Parents become independent, reducing cost of inference
- Although simple, BN2O are reasonable first approximations for a medical diagnosis network

BN2O Software

- QMR: Quick Medical Reference
  - Compiled for diagnosis of internal medicine
  - QMR-DT (Decision Theoretic)
    - Contains more than five hundred significant diseases
    - Four thousand associated findings
    - More than forty thousand disease finding associations
- CPCS
  - Smaller: 500 variables, 900 edges
    - Has variables for predisposing factors, etc
    - Take four values
    - Full CPDs would take 134 million parameters

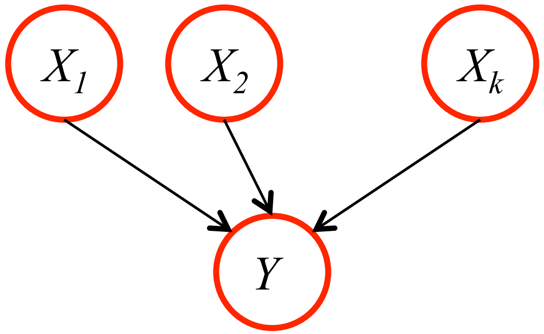


Generalized Linear Models

- A very different class of models that also satisfy independence of causal influence
- We focus on models that define probability distributions $P(Y|X_1,..,X_k)$ where Y takes on values in some discrete finite space
- We first consider the case where Y and all the X_is are binary-valued
- We then extend to the multinomial case

Binary Variables and Linear Threshold

- Consider a CPD where each of several binary variables $X_1,\ldots,X_k$ adds to a total burden.
- Effect on Y is characterized by a linear function $f(X_1,..,X_k) = \Sigma_k w_k X_k$

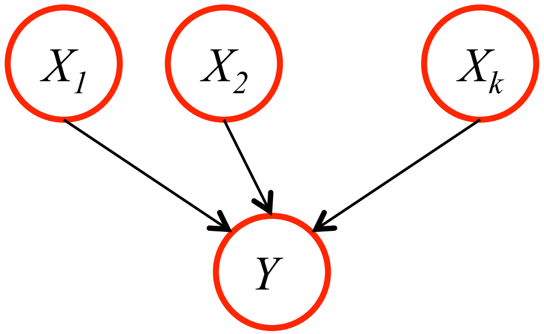


- When the total burden exceeds a threshold τ , the probability transitions from 0 to 1
- Use soft threshold and w_0 to eliminate τ

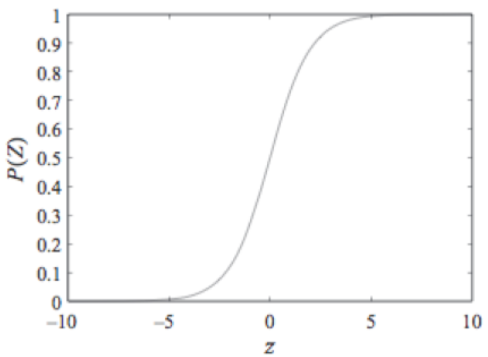
Definition of Logistic CPD

- Child value is a linear function of parents
- Y is binary-valued, parents X_i are numerical
- Effect of the X_i 's on Y is a linear function

$$\boxed{P(y^1 \mid X_1,..X_k) = sigmoid(w_0 + \sum_i w_i X_i)}$$



$$sigmoid(z) = \frac{1}{1+e^{-z}}$$



Interpretation of parameter w_i

- Can be interpreted in terms of its effect on the log-odds of Y
- Log-odds for a binary variable is the ratio of the probability of y^1 and the probability of y^0
- Same concept as when we say odds are 2 to 1

Effect of X_j on Log Odds

- Ratio of the probability of y^1 and the probability of y^0
- We use Z to represent $w_0 + \sum_i w_i X_i$
- Odds for the variable Y

$$O(X) = \frac{P(y^1 \mid X_1, ..X_k)}{P(y^0 \mid X_1, ..X_k)} = \frac{e^Z / (1+e^Z)}{1/(1+e^Z)} = e^Z$$

- Effect of this odds as some variable X_j changes its value from *false* to *true*

$$\frac{O(X_{-j}, x_j^1)}{O(X\_j, x_j^0)} = \frac{\exp(w_0 + \sum_{i \neq j} w_i X_i + w_j)}{\exp(w_0 + \sum_{i \neq j} w_i X_i)} = e^{w_j}$$

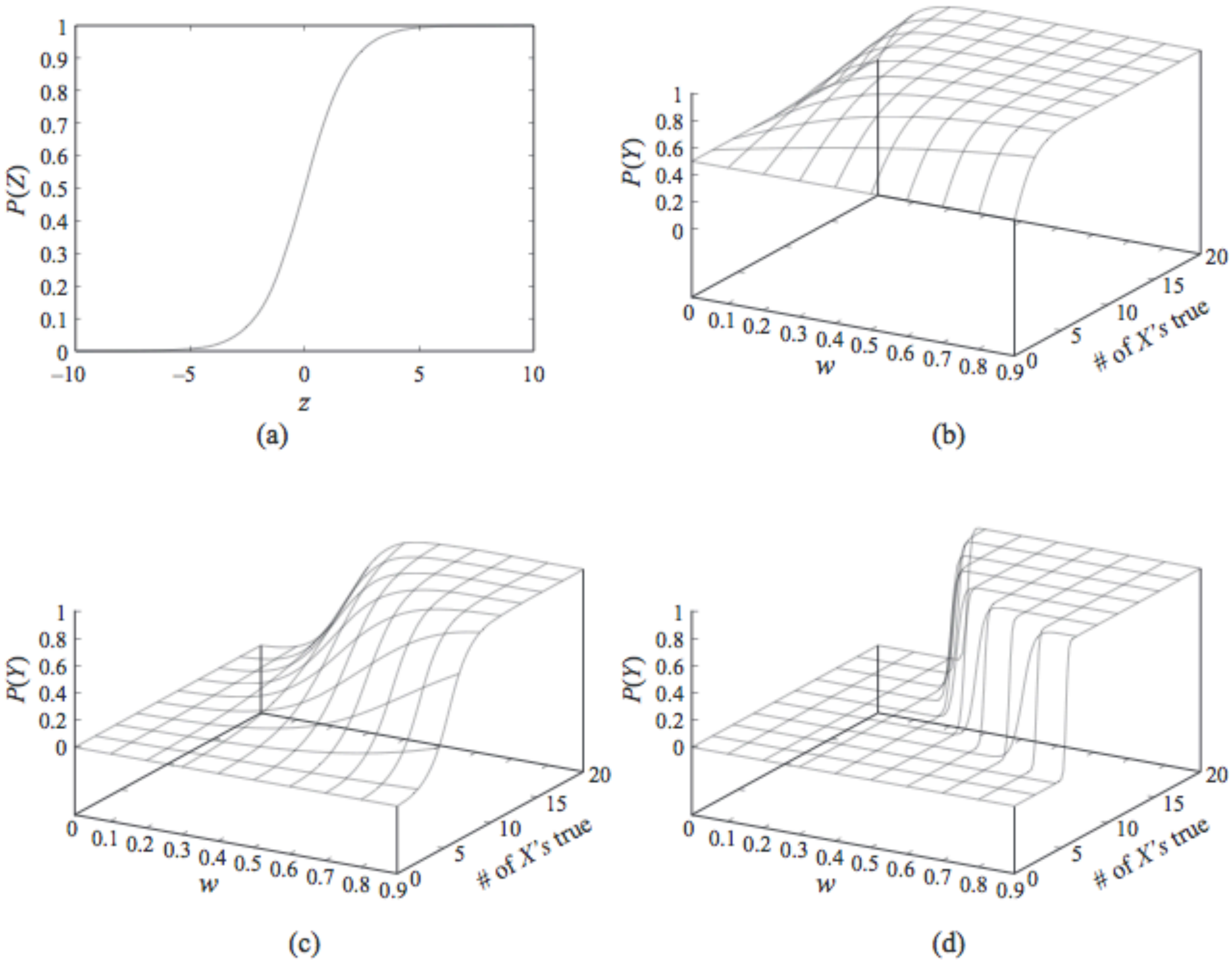
– X_j=*true* changes odds by a multiplicative factor of e^{wj}. If w_j>0 odds increases

Behavior of Sigmoid CPD

All variables have same weight w

$w_0=0$

$w_0=-5$



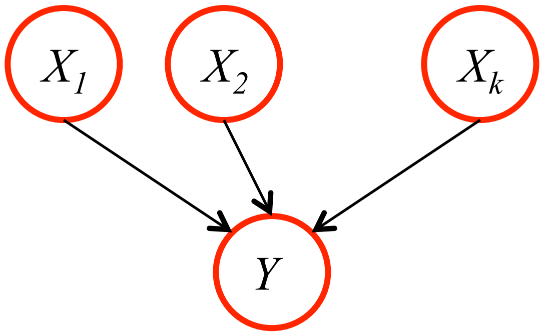
w and w_0 multiplied by 10

Multi-valued Variables

- Y takes on multiple-values, $y^1,..y^m$
- Parents $X_1,.., X_k$ are numerical
- CPD is *multinomial logistic* if for each $j=1,..,m$ there are $k+1$ weights $w_{j,0}, w_{j,1},..w_{j,k}$ such that

$$\ell_j(X_1,..X_k) = w_{j,0} + \sum_{i=1}^{k} w_{j,i} X_i$$

$$P\left(y^j \mid X_1,..X_k\right) = \frac{\exp\left(\ell_j(X_1,..X_k)\right)}{\sum_{j'=1}^{m} \exp\left(\ell_j(X_1,..X_k)\right)}$$



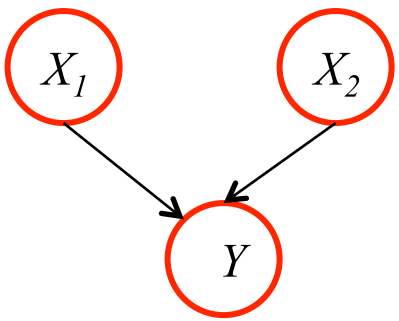
Multinomial Logistic CPD

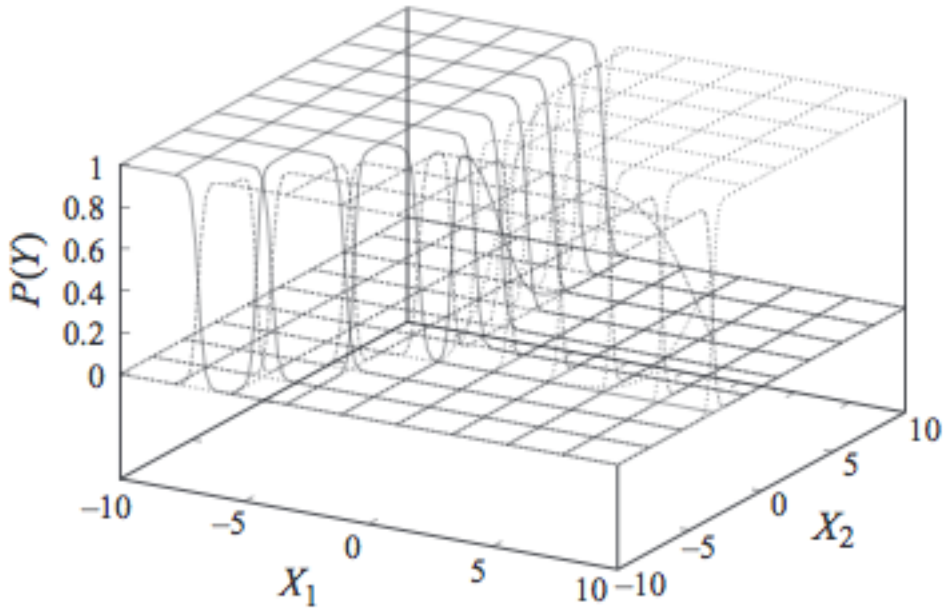
- Y has multiple-values, Parents X_i numerical
- $P(Y|X_1,X_2)$ has the Multinomial logistic model
  - Three-valued child Y

$l_1(X_1, X_2)$=-3X_1 - 2X_2+1

$l_2(X_1, X_2)$=5X_1 - 8X_2 - 4

l_3=x-y+10



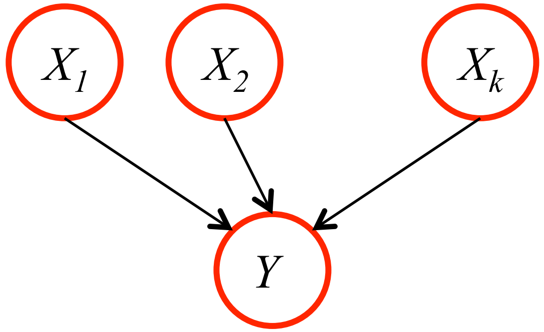


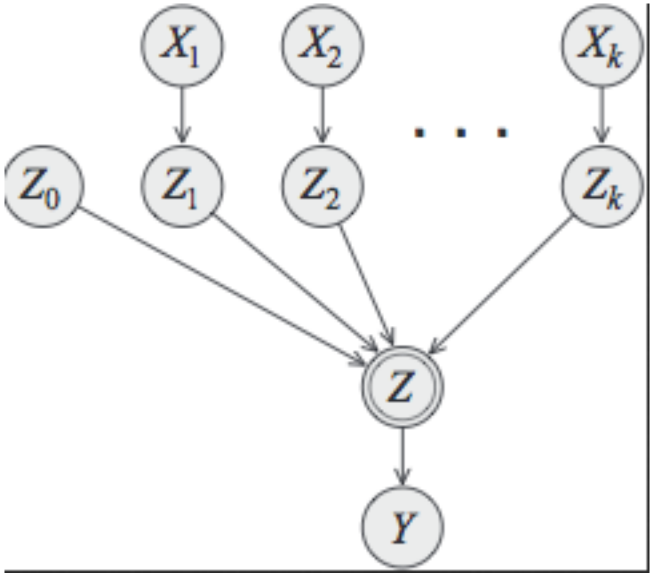
General Formulation

- Noisy-or and Generalized linear models are special cases of Causal Independence or independence of Causal Influence
- Influence of multiple causes can be decomposed into separate influences

Independence of Causal Influence (ICI)

- Let Y be a variable with parents $X_1,..X_k$.
- The CPD $P(Y|X_1,..X_k)$ exhibits ICI if it can be described by:





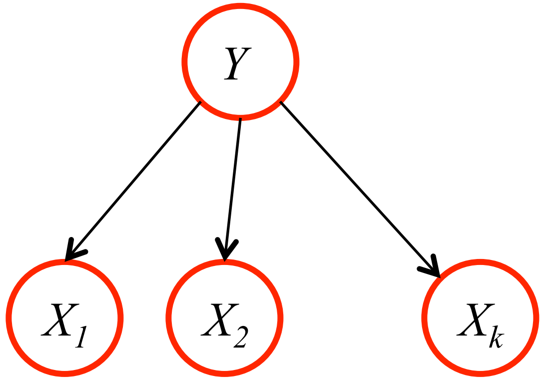
Where Z is a deterministic function f

Each variable can be transformed separately

Comparison with Naiive Bayes

- BN for Naiive Bayes Classifier with joint distribution

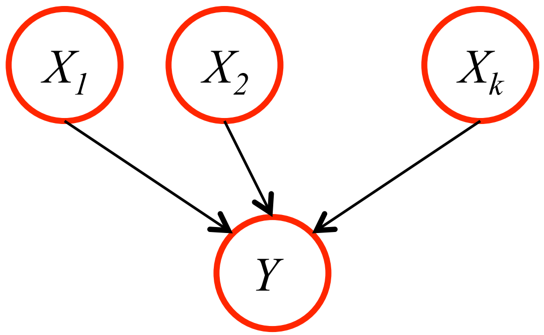
  $P(Y,X_1,..X_k)=P(Y)P(X_1|Y)..P(X_k|Y)$



- We are interested here in learning the local CPD, as in

  $P(Y,X_1,..X_k)=P(X_1)\ldots P(X_k)\ P(Y|X_1,\ldots,X_k)$

  – CPD can be learnt with Naïve Bayes or a neural network!



  – Given the joint we can use it to determine $P(X_i/Y)$ which may not be independent (note V-structure)