概率图模型课程辅导

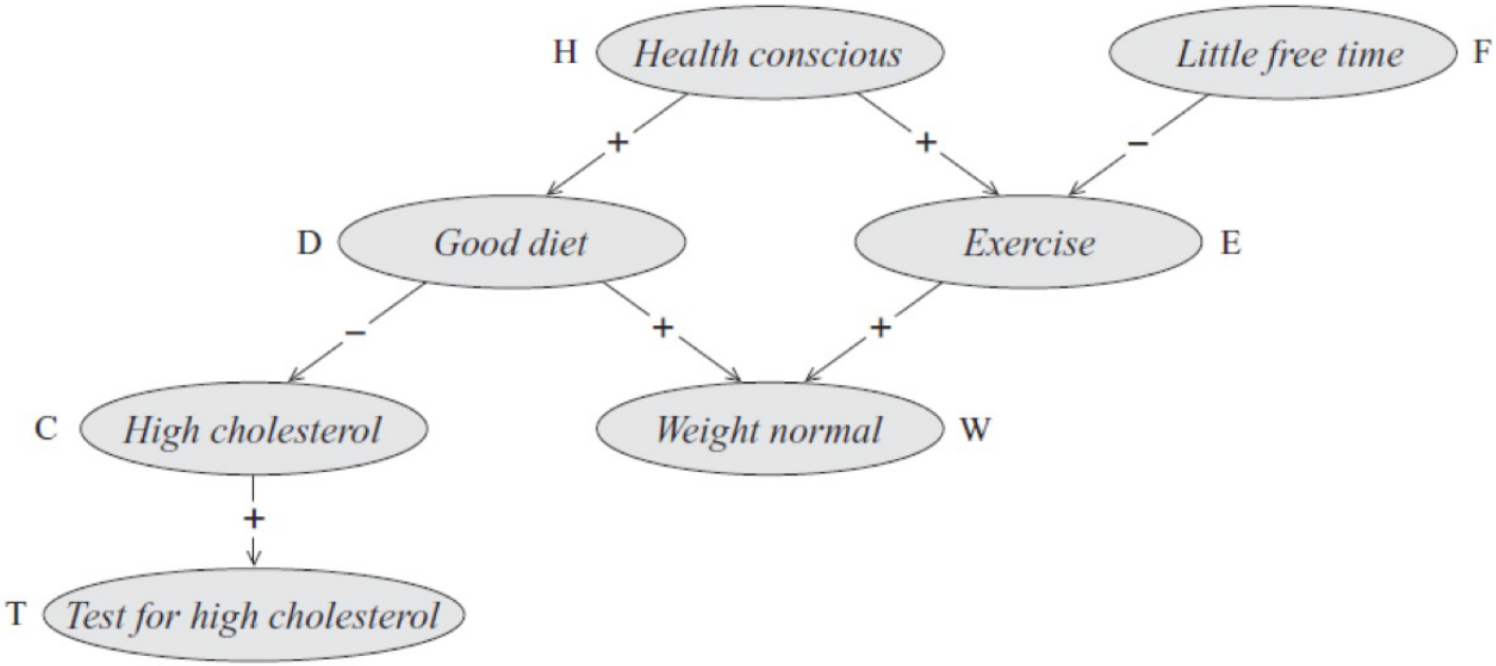
2016 年秋季学期

Bayesian Network - Representation

Independence from the graph: active trail/d-seperation

The trail $X_1 \leftrightarrow \cdots \leftrightarrow X_n$ is active given evidence E if:

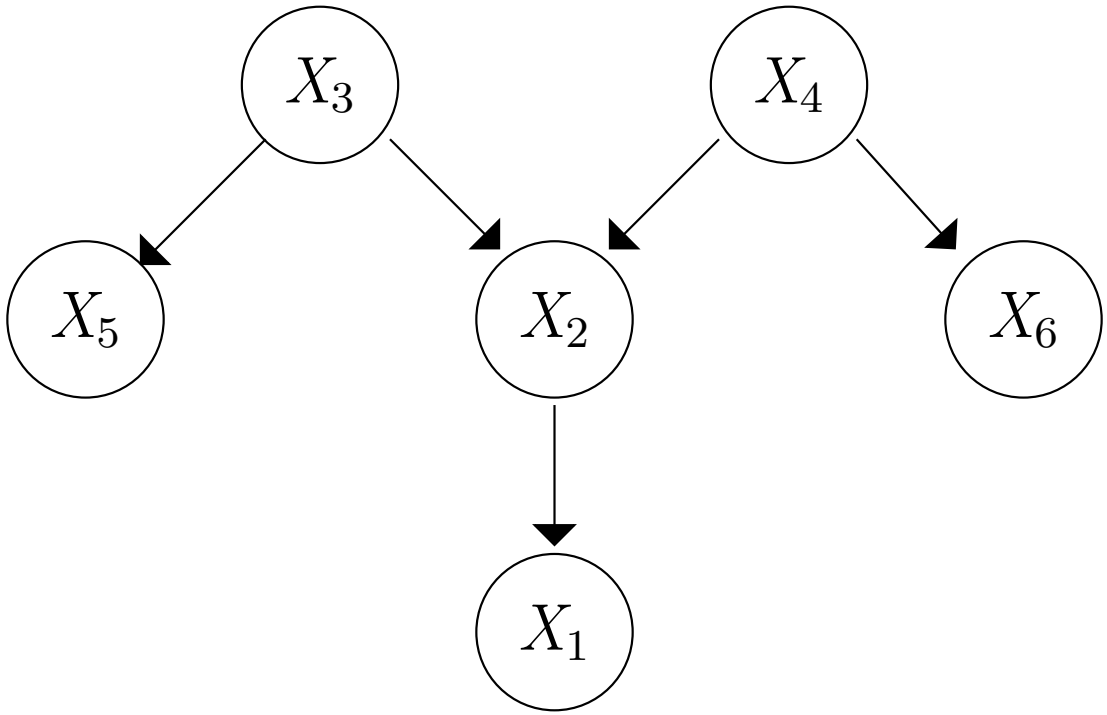
- For every $V-$structure $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, X_i **or one of its descendants** is observed.
- No other nodes along the trail are in E.



Active trails between F and T given W?

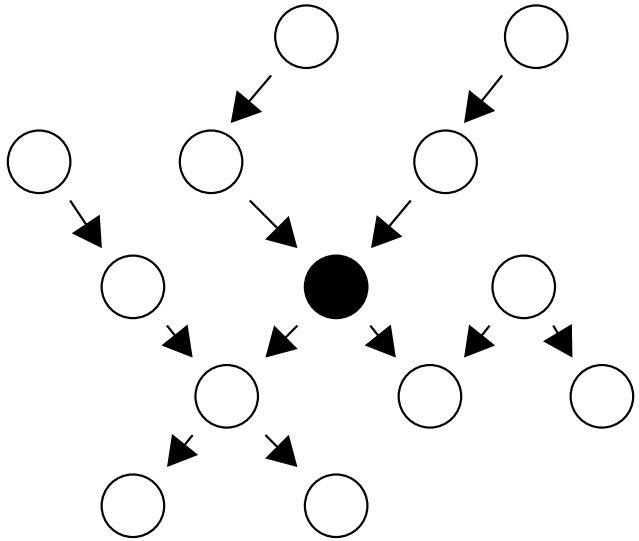
Markov Blanket in BNs

X is independent with other variables if the corresponding MB(X) is given.



$$P(X_2 = x_2 | X_1 = x_1, X_3 = x_3, X_4 = x_4, X_5 = x_5, X_6 = x_6)$$

How about the Markov Blanket of the filled node?



Markov Networks - HC theorem

$$P(X_1, \cdots, X_n) = \frac{1}{Z} \prod_{i=1}^{m} \pi_i(\boldsymbol{D}_i)$$

- where each $\boldsymbol{D}_i$ is a complete subgraph (or **clique**) in the network.
- Usually we define $\pi_i(\boldsymbol{D}_i)$ by an exponential form,

$$\pi_i(\boldsymbol{D}_i) = e^{-\epsilon(\boldsymbol{D}_i)}$$

and $\epsilon(\boldsymbol{D}_i)$ is the energy function of $\boldsymbol{D}_i$.

$$P(\boldsymbol{X}) = \frac{1}{Z} \exp\left(-\sum_{i=1}^{m} \epsilon_i(\boldsymbol{D}_i)\right)$$

For a positive distribution P, we can expand the energy function:

$$P(\boldsymbol{X}) = e^{Q(x)} P(\mathbf{0})$$

$$\begin{aligned} Q(\boldsymbol{X}) = & \sum_i X_i \psi_i(X_i) + \sum_{i,j} X_i X_j \psi_{i,j}(X_i, X_j) + \cdots \\ & + X_1 X_2 \cdots X_n \psi_{1,2,\cdots,n}(X_1, \cdots, X_n) \end{aligned}$$

- $\psi_S > 0$ only if $v \in S$ form a clique.
- If all variables are binary variables $\{0, 1\}$, all ψs are constants.

For example, $X_i, i = 1, 2, 3$ are all binary variables $\{0, 1\}$

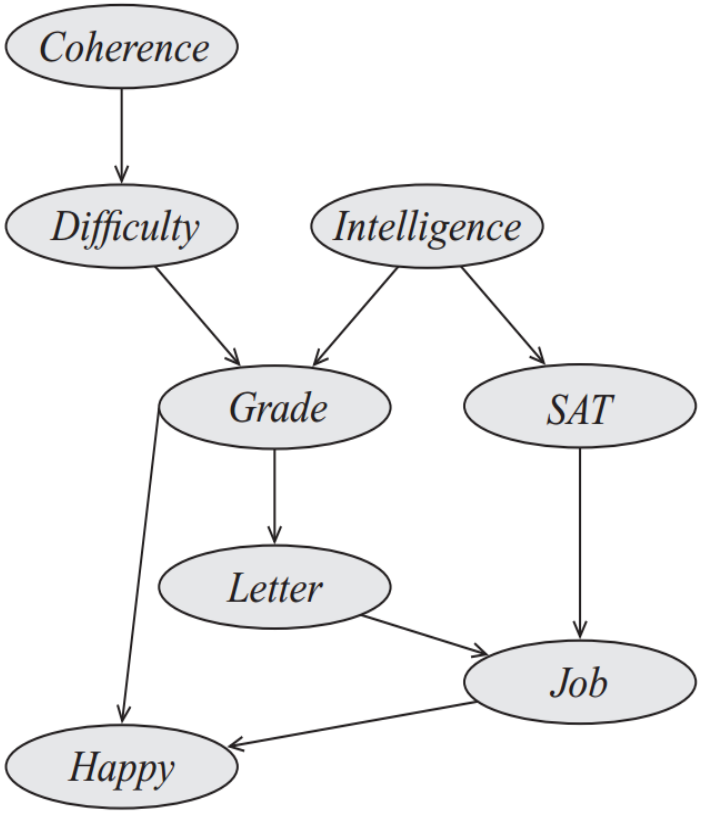
$$X_1 - X_2 - X_3$$

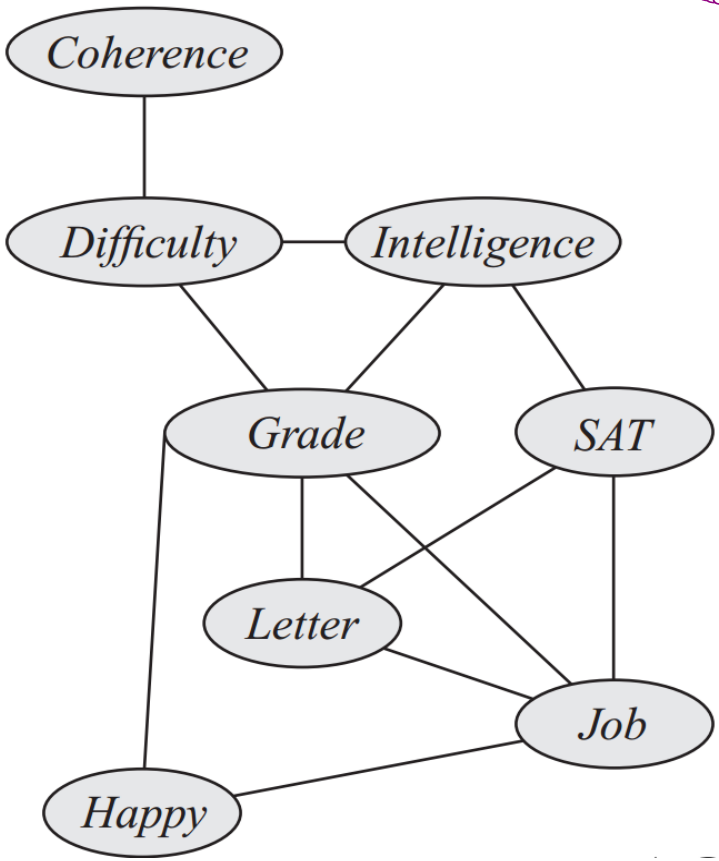
We can write $Q(\boldsymbol{X})$ as:

$$Q(\boldsymbol{X}) = \sum_{i=1}^{3} \alpha_i X_i + \sum_{1 \leq 1 < j \leq 3} \alpha_{ij} X_i X_j + \alpha_{123} X_1 X_2 X_3$$

where:

- $\alpha_{13} = 0$;
- $\alpha_{123} = 0$.


Coherence
Difficulty
Intelligence
Grade
SAT
Letter
Job
Happy


Coherence
Difficulty
Intelligence
Grade
SAT
Letter
Job
Happy

Homework Solutions

H2.2. Consider a set of variables $X_1, X_2, \cdots, X_n$ where each X_i has $|Val(X_i)| = l$. (1). Assume that we have a Bayesian network over $X_1, \cdots, X_n$, such that each nodes has at most k parents. What is a simple upper bound on the number of independent parameters in the Bayesian network? How many independent parameters are in the full joint distribution over $X_1, \cdots, X_n$?

$$\sum_{i=1}^{n}(l-1)l^k = n(l-1)l^k, \qquad l^n - 1$$

(2). Now, assume that each variable X_i has the parents $X_1, \cdots, X_{i-1}$. How many independent parameters are there in the Bayesian network? What can you conclude about the expressive power of this type of network?

$$P(X_1, X_2, \cdots, X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\cdots P(X_n|X_1, \cdots, X_{n-1})$$

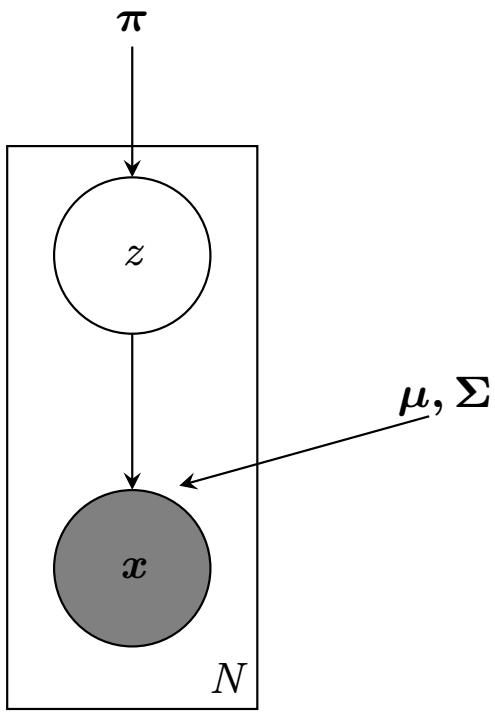
(3). Now, consider a naïve Bayes model where $X_1, \cdots, X_n$ are evidence variables, and we have an additional class variable C, which has k possible values $c_1, c_2, \cdots, c_k$. How many independent parameters are required to specify the naïve Bayes model? How many independent parameters are required for an explicit representation of the joint distribution?

$$k - 1 + nk(l - 1)$$

H3.1.Mixture Model.

(1). Gaussian Mixture model.

$$P(\boldsymbol{x}[n]|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}[n]|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) \qquad \pi_k = P(z[n] = k), \qquad n = 1, \cdots, N$$

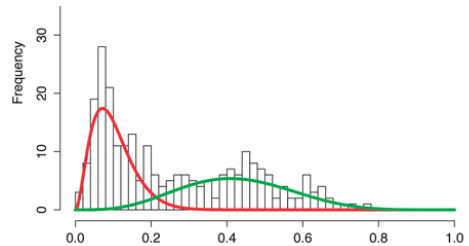


(2). Now suppose we've known all parameters: $\boldsymbol{\pi} = \{\pi_k, k = 1, \cdots, K\}$ and $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \{(\mu_k, \Sigma_k), k = 1, \cdots, K\}$. For an observation x, please calculate $P(z = k|x), k = 1, \cdots, K$.

$$P(z = k|\boldsymbol{x}) = \frac{\pi_k \mathcal{N}(\boldsymbol{x}|\mu_k, \Sigma_k)}{\sum_{i=1}^{K} \pi_i \mathcal{N}(\boldsymbol{x}|\mu_i, \Sigma_i)}$$

(3). **Base distribution.** Different local base distributions could help us model different kinds of data.

Now we have a batch of observations $x[n] \in R,\ n = 1, \cdots, N$, which are real numbers located at the interval $(0, 1)$. We plot the histogram of these observations:



Beta

(3). **Number of mixture components.** All of the above problems consider a **finite** mixture model. So how to determine the K? This problem has no general standards, but there exists some solutions. We hope you give $2 \sim 3$ possible solutions.

BIC score, explained variance, Nonparametrics Bayesian, etc.

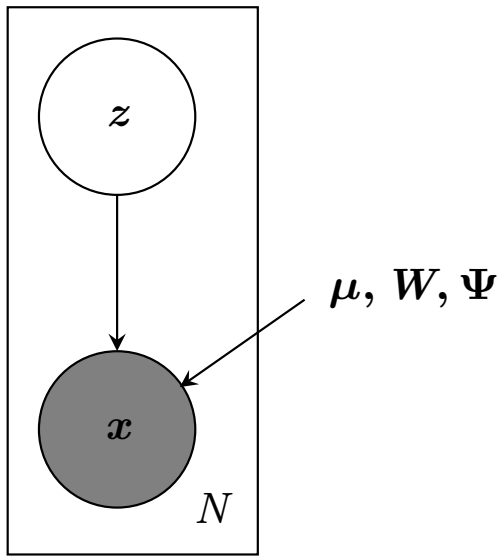
H3.2 .Latent variable models (LVMs).

(1). **Factor analysis(FA).**

$$p(\boldsymbol{z}[n]) = \mathcal{N}(\boldsymbol{z}[n]|\mathbf{0}, \boldsymbol{I})$$

$$p(\boldsymbol{x}[n]|\boldsymbol{z}[n]) = \mathcal{N}(\boldsymbol{x}[n]|\,\boldsymbol{W}\boldsymbol{z}[n] + \boldsymbol{\mu}, \Psi)$$

$\boldsymbol{z}[n] \in R^L, \boldsymbol{x}[n] \in R^D (D \gg L), \Psi$ is forced to be diagonal.



The number of independent parameters: $2D + L \times D$, compared with multivariate Gaussian: $D + D^2$.

(2). **Modeling of cancer gene expression.**

Suppose for every gene of a patient, we can acquire its **gene expression information** which is the **count** of every gene's measured reads (more reads means this gene is highly expressed).

Now have N patients, and test their G genes' information. We use e_{ij} to represent j-th gene's gene expression state of i-th patient, $e_{ij} \in \{0, 1, 2, 3, \cdots\}$.

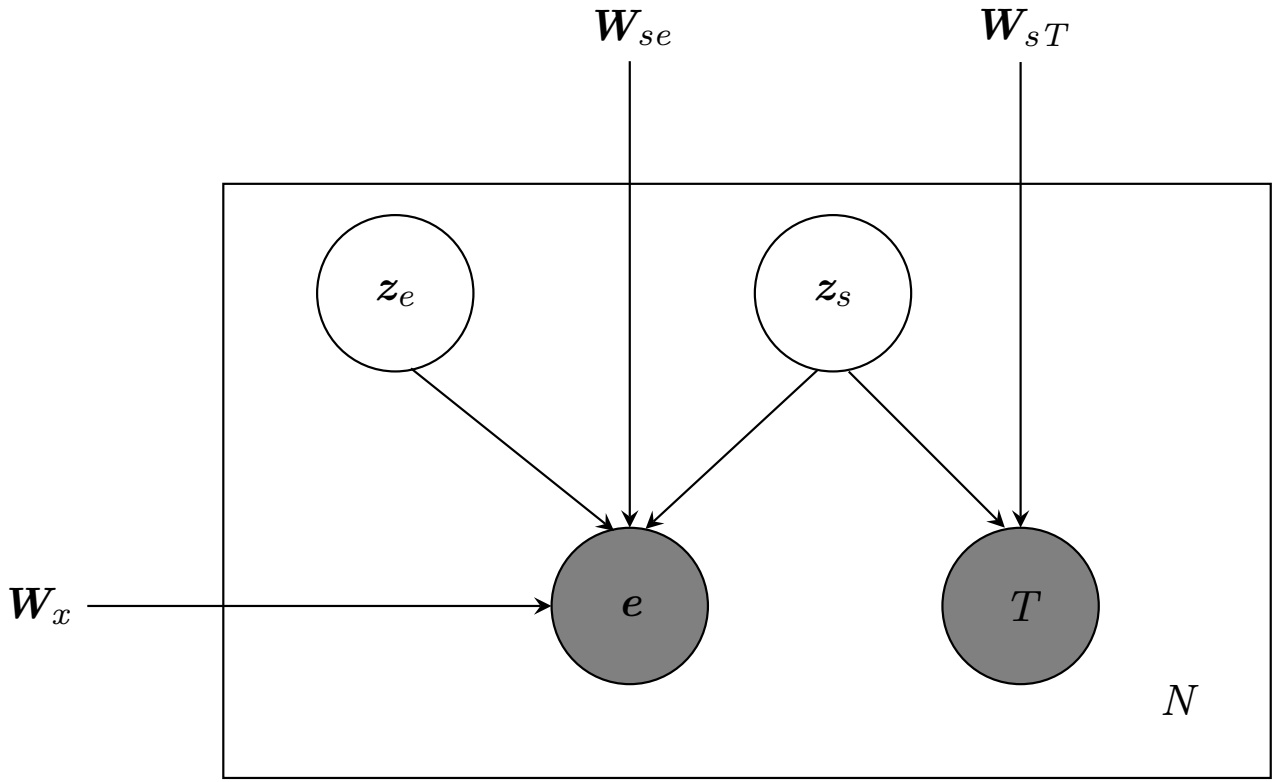
(a). One hypothesis could be that there exists a small number of **driver factors**, and these factors leads to the patients' gene expression state. We hope you to model this problem based on this hypothesis. Please design and give a detailed description of an appropriate model and its local probability form.

Almost the same with (1):

$$P(e_{ij}|\boldsymbol{z}_i) = \textbf{Poisson}(\boldsymbol{w}_i^T \boldsymbol{z}_i + \mu_j)$$

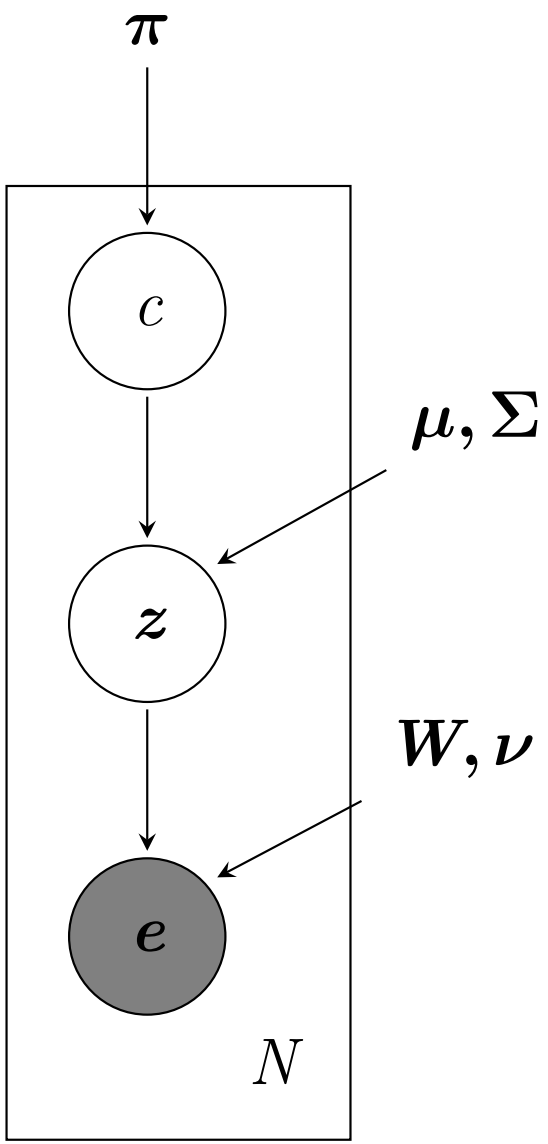
(b). Besides the gene expression data, for every patient, we could also know the days T from his/her diagnosis to his/her death or recurrence. We want to predict T by using omics data. But we cannot ensure variables derived by FA all have association with T. This means we may need **two groups of latent varibales**, one capturing the information shared with T, and the other capturing original featuires' own information. Please use this strategy to design an appropriate model.

This is the partial least square model:



H3.3. In this problem, we will use mixture models and factor analysis models as building blocks to construct more complex models.

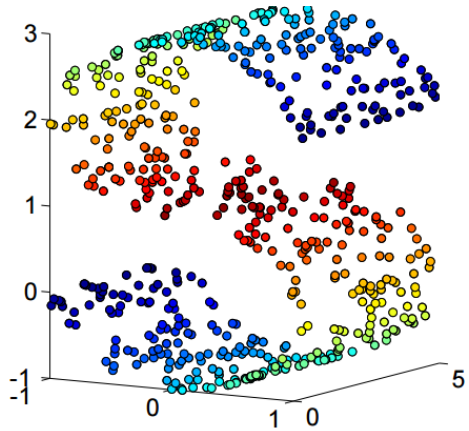
(1). Let's reconsider the problem of *2.(2)(a)*. Even for patients from one kind of cancer, such as breast cancer, we've known there exists very **different groups or subtypes**, who show great heterogeneity in their omics features. These different groups of patients also show great differences in their driver factors. Hypothesis of this problem is that patients from one kind of cancer may have the same driver factors, but those from different groups could have significantly different perturbations of these factors. Please build an appropriate graphical model based on this idea and give the local probability distributions.

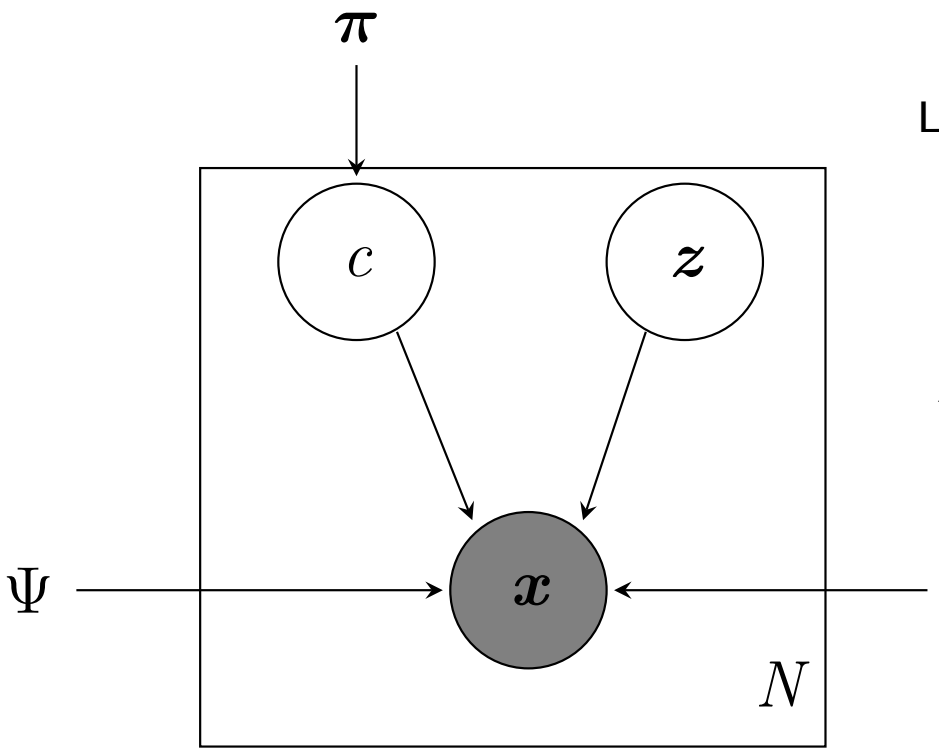


Local probability distribution:

$$
\begin{aligned}
P(c_i = k) &= \pi_k, \qquad \sum_{k=1}^{K} \pi_k = 1 \\
P(\boldsymbol{z}_i | c_i) &= \mathcal{N}(\boldsymbol{z}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\
P(e_{ij} | \boldsymbol{z}_i) &= \textbf{Poisson}(\boldsymbol{w}_i^T \boldsymbol{z}_i + \nu_i)
\end{aligned}
$$

(2). Consider the following data points (see neaxt page),
Maybe you have known we can regard the factor analysis model as a dimenion reduction method, observing it uses a lower-dimension hyerplane to approximate original data points.
But in tha above figure, it's hard to model these data points using FA, becuase no **single** planes can fit them. But we can model them as a finite **mixture** of planes.
Please design and give a detailed description of an appropriate model and its local probability form.





Local CPD:

$$P(c_i = k) = \pi_k, \quad \sum_{k=1}^{K} \pi_k = 1$$

$$P(\boldsymbol{x}_i | c_i, \boldsymbol{z}_i) = \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_k + \boldsymbol{W}_k \boldsymbol{z}_i, \Psi)$$