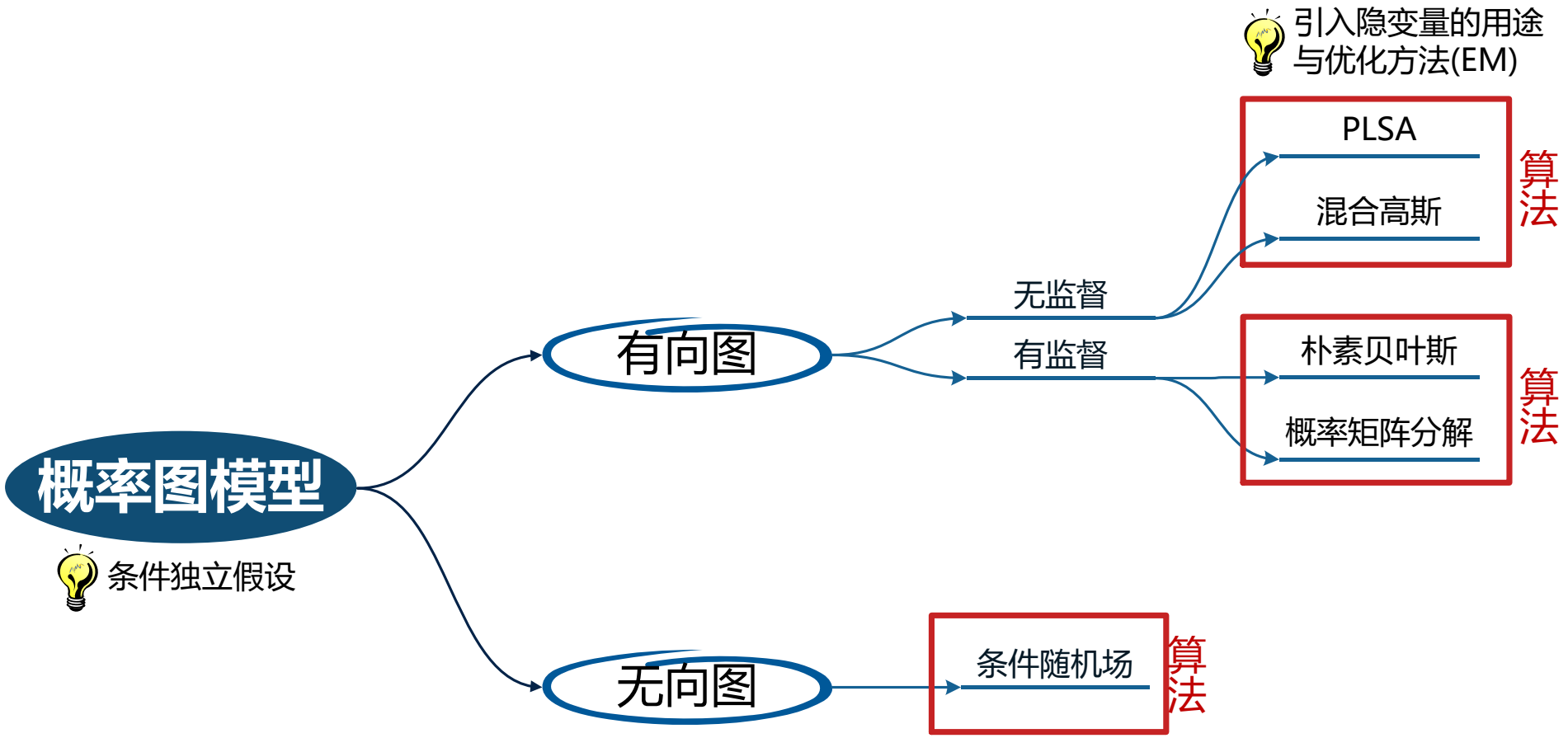
课程专题二：概率图模型方法

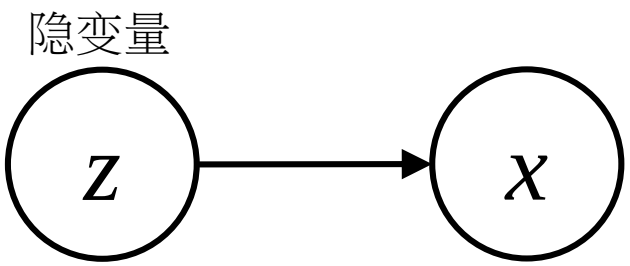


EM应用于混合高斯模型
Mixture of Gaussian

- 数据$X = \{(x_i, c_i)\}_{\mathrm{i=1}}^{\mathrm{N}}$，$x_i$出现了$c_i$次

$$\sum_{i=1}^{N} c_i \sum_{z} P(z|x_i, \theta') \ln P(x_i, z|\theta)$$

- 应用于混合高斯模型，数据$X = \{(x_i, 1)\}_{\mathrm{i=1}}^{\mathrm{N}} (x_i \in \mathbb{R}^m)$



  - 其中z是离散型随机变量，有k个取值$z_1, z_2 \ldots, z_k$

$$x|z_k \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$P(x|z_k) = N(x|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

EM应用于混合高斯模型

- 参数：
  - $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$(k=1,2,...,K)
  - $P(z_k)$ (k=1,2,...,K)
- $P(x_i, z_k|\theta) = P(z_k)P(x_i|z_k) = P(z_k)N(x_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- $P(z_k|x_i, \theta') = \frac{P(x_i, z_k|\theta')}{P(x_i|\theta')} = \frac{P(z_k|\theta')N\left(x_i\middle|\boldsymbol{\mu}'_k, \boldsymbol{\Sigma}'_k\right)}{\sum_l P(z_l|\theta')N\left(x_i\middle|\boldsymbol{\mu}'_l, \boldsymbol{\Sigma}'_l\right)}$

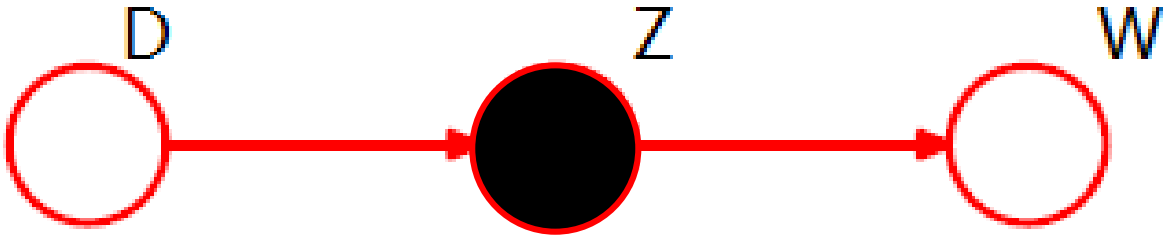
EM应用于混合高斯模型

- 模型的输出结果：$P(z_k|x_i)$，达到了聚类的效果
- 引入隐变量的目的
  - **获得可观察变量与隐变量的关系**

方法论：概率图模型建模

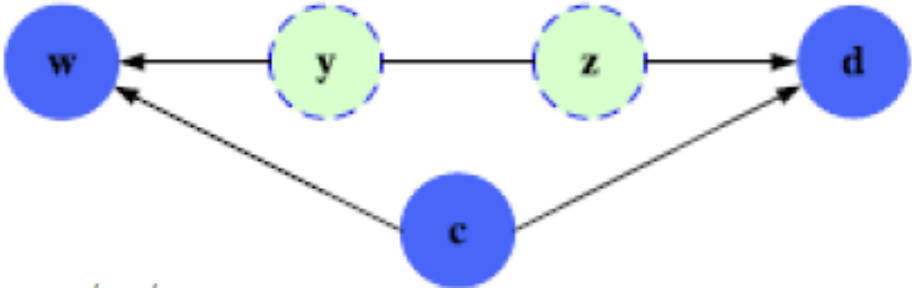
方法论：概率图模型

- 根据实际问题，总结问题的随机变量
  - **可观察到的**
  - **随机变量类型：连续型、离散型**
- 引入隐变量
- 根据实际问题，提出合理的假设，建立概率图模型

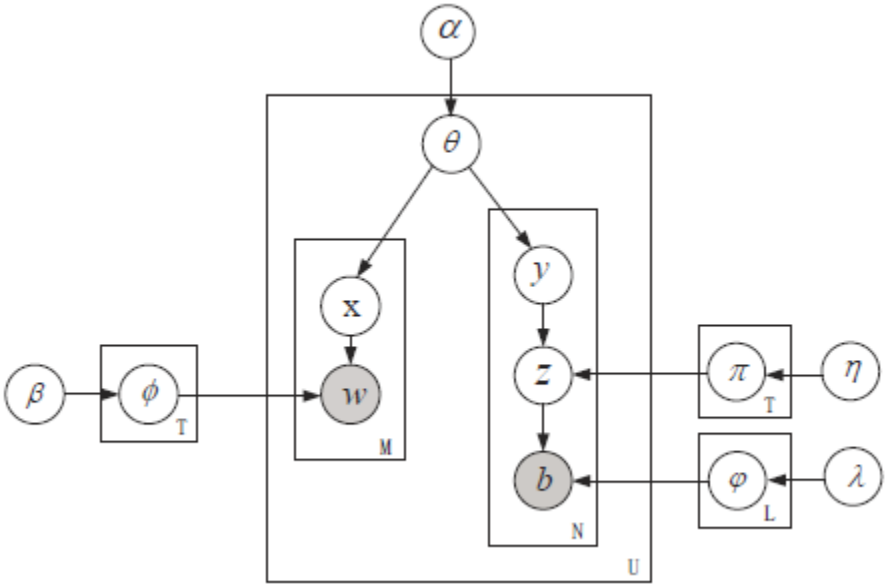


- 根据最大化数据似然，求解模型参数
  - **EM算法**

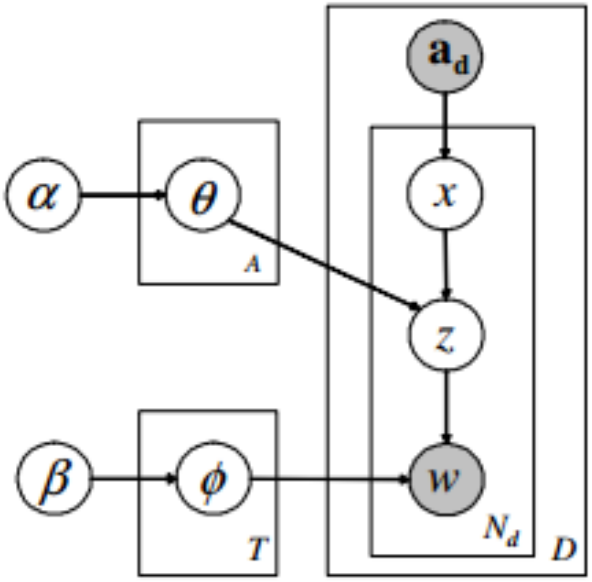
解决实际问题：概率图模型



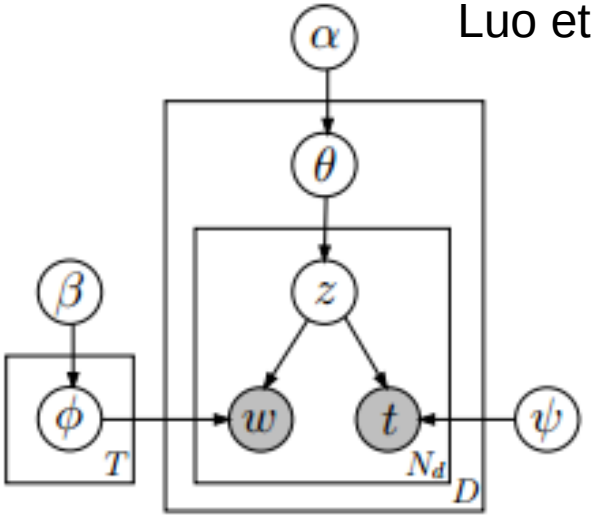
Luo et al, TKDE, 2012



Luo et al, KDD, 2016



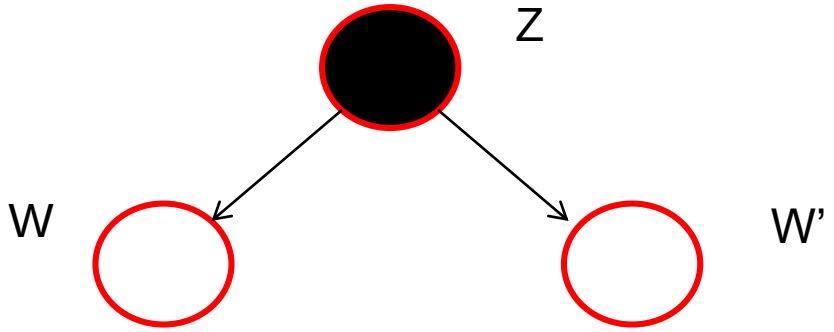
Rosen-Zvi et al, UAI, 2004



Wang et al, KDD, 2006

作业：短文本的Topic Modeling

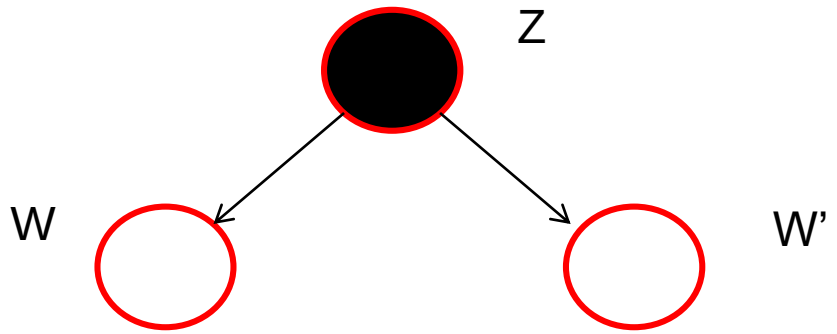
- 短文本上的Topic Modeling
  - **短文本：一条微信朋友圈、一条微博**
  - **“文档-词”共现矩阵：非常稀疏**
- 数据转化
  - **从“文档-词”共现矩阵**
  - **到“词-词”共现矩阵**



Xueqi Cheng et al. BTM: Topic Modeling over Short Texts. TKDE, 2014.

作业：短文本的Topic Modeling

- 使用EM算法，求解此概率图模型的参数



要求：

1）写出目标函数，目标函数的下界

2）EM算法推导出的迭代式子