电子科技大学

University of Electronic Science and Technology of China

Chap. 17
Parameter Estimation

Junhua Chen , PengFei XIao

Data Mining Lab,
Big Data Research Center, UESTC

All the problems of estimating parameters for a Bayesian network in this chapter are assumed to have fixed network structure and that our data set consists of fully observed instances of the network variables.

目录：

1. 最大似然估计MLE
   - 极大似然估计
   - 贝叶斯网上的MLE
2. 贝叶斯参数估计
   - 先验后验
3. 贝叶斯网络参数估计
   - 似然分解
4. 共享参数模型

MLE:

一个图钉引发的血案:

假设有一IID的图钉抛掷结果集$x[1],\ldots x[M]$，且$x[m]$头部向上H的概率是θ（向下T为1-θ）。我们的任务是找到一个最好的θ。考虑上一章，我们要给定:

- 假设空间Θ
- 目标函数

如何找到θ呢?

✓ 考虑一个结果序列H,T,T,H,H。发生的概率为

$$P(\langle H,T,T,H,H\rangle : \theta) = \theta^3(1-\theta)^2 = L(\theta : \langle H,T,T,H,H\rangle)$$

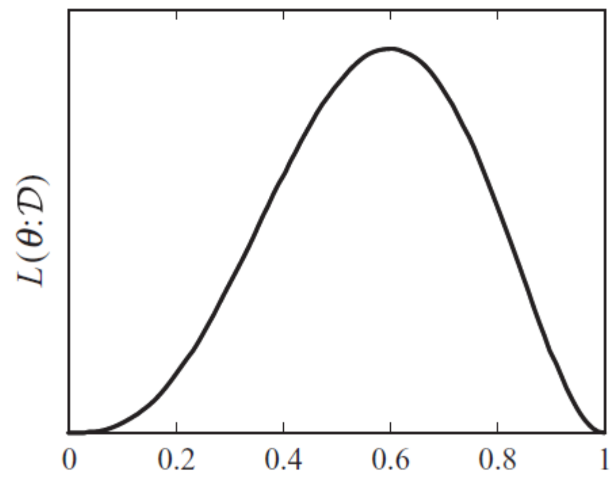
一般的情况，似然可以记为

$$L(\theta : D) = \theta^{M[1]}(1-\theta)^{M[0]}$$

MLE：

数据挖掘实验室
Data Mining Lab

一个图钉引发的血案:



一般的情况，似然可以记为

$$L(\theta:D)=\theta^{M[1]}(1-\theta)^{M[0]}$$

对数似然

$$\ell(\theta:D)=M[1]log(\theta)+M[0]log(1-\theta)$$

最大化$\ell(\theta:D)$

$$\hat{\theta}=\frac{M[1]}{M[1]+M[0]}$$

MLE:

MLE:

- 一颗栗子：假定X是一个可以取值$x^1, \ldots, x^k$的多项式变量。多项式分布为

$$P(x:\boldsymbol{\theta}) = \theta_k, x = x^k$$

参数空间$\Theta = \{\boldsymbol{\theta} \in [0,1]^K : \sum_i \theta_i = 1\}$

似然函数

$$L(\theta:D) = \prod_k \boldsymbol{\theta}_k^{M[k]}$$

似然函数general

$$L(\boldsymbol{\theta}:D) = \prod_m P(\xi[m]:\boldsymbol{\theta})$$

似然估计general

$$L(\hat{\boldsymbol{\theta}}:D) = max_{\theta\in\Theta} L(\boldsymbol{\theta}:D)$$

数据挖掘实验室
Data Mining Lab

最后多说几句:

定义17.1 一个从χ的实例到$\boldsymbol{R}^{\ell}$（对某个ℓ）的函数$\tau(\xi)$是充分统计量，如果任意的两个数据集D, D'，对任意的$\boldsymbol{\theta} \in \Theta$，存在：

$$\sum_{\xi[m]\in D} \tau(\xi[m]) = \sum_{\xi'[m]\in D'} \tau(\xi'[m]) \Rightarrow L(\boldsymbol{\theta}: D) = L(\boldsymbol{\theta}: D')$$

元组$\sum_{\xi[m]\in D} \tau(\xi[m])$称为数据集$D$的充分统计量。

比如在刚才的多项式栗子中

$$\tau(x^k) = (\overbrace{0, \dots, 0}^{k-1}, 1, \overbrace{0, \dots, 0}^{n-k})$$

MLE：

贝叶斯网的MLE:

又奖励一颗栗子：网络结构$X \rightarrow Y$与参数$\boldsymbol{\theta}$。似然函数

$$L(\theta : D) = \prod_m P(x[m], y[m] : \boldsymbol{\theta})$$

$$= \prod_m P(x[m] : \boldsymbol{\theta}) P(y[m] | x[m] : \boldsymbol{\theta})$$

$$= \left(\prod_m P(x[m] : \boldsymbol{\theta})\right)\left(\prod_m P(y[m] | x[m] : \boldsymbol{\theta})\right)$$

分解—从参数$\boldsymbol{\theta}$开始

$$\left(\prod_m P(x[m] : \theta_X)\right)\left(\prod_m P\big(y[m] | x[m] : \theta_{Y|X}\big)\right)$$

MLE：

数据挖掘实验室
Data Mining Lab

贝叶斯网的MLE:

第二项

$$\prod_{m} P\big(y[m]|x[m]:\theta_{Y|X}\big)$$

$$= \prod_{m:x[m]=x^0} P\big(y[m]|x[m]:\theta_{Y|X}\big) \cdot \prod_{m:x[m]=x^1} P\big(y[m]|x[m]:\theta_{Y|X}\big)$$

$$= \prod_{m:x[m]=x^0} P\big(y[m]|x[m]:\theta_{Y|x^0}\big) \cdot \prod_{m:x[m]=x^1} P\big(y[m]|x[m]:\theta_{Y|x^1}\big)$$

进一步

$$\prod_{m:x[m]=x^0} P\big(y[m]|x[m]:\theta_{Y|x^0}\big) = \theta_{y^1|x^0}^{M[x^0,y^1]} \cdot \theta_{y^0|x^0}^{M[x^0,y^0]}$$

最大化

$$\theta_{y^1|x^0} = \frac{M[x^0,y^1]}{M[x^0]}$$

MLE：

数据挖掘实验室
Data Mining Lab

全局似然分解:

一个general的似然可以分解为

$$L(\theta:D)=\prod_i L_i\left(\theta_{X_i|Pa_{X_i}}:D\right)$$

其中X_i的局部似然函数是

$$L\left(\theta_{X_i|Pa_{X_i}}:D\right)=\prod_m P\left(x_i[m]|pa_{X_i}[m]:\theta_{X_i|Pa_{X_i}}\right)$$

命题17.1 令D为$X_1,\ldots,X_n$的一个完备数据集，$\mathcal{G}$为这些变量上的一个网络结构，并且假定对于所有的$j\neq i$，参数$\theta_{X_i|Pa_{X_i}}$与$\theta_{X_j|Pa_{X_j}}$不相交。令$\hat{\theta}_{X_i|Pa_{X_i}}$是最大化$L\left(\theta_{X_i|Pa_{X_i}}:D\right)$的参数，那么，$\hat{\theta}=\left\langle\hat{\theta}_{X_1|Pa_{X_1}},\ldots,\hat{\theta}_{X_n|Pa_{X_n}}\right\rangle$最大化$L(\theta:D)$。

MLE：

在结束这个话题之前

甜点1： 非参数模型

假定我们希望从数据中学习分布$P(X|U)$，一个合理的假设是这个CPD是光滑的。于是，如果在训练集中观测到x, u，那么对U的类似值，观测到X的类似值得概率将会增大。正式的，对于ϵ以及δ的较小的值，我们增加了$P(X = x + \epsilon | U = u + \delta)$的密度。

刻画这种直觉的方法是核密度估计（也称Parzen窗口）：给定数据集D，通过在每一个样本$x[m], u[m]$处展开密度来估计一个局部联合密度

$$\tilde{P}_X(x, u) = \frac{1}{M} \sum_m K(x, u; x[m], u[m], \alpha)$$

-  ：估计灵活，可以根据观测自我调整；
-  ：对原数据无“压缩”使得分布有分歧，样本充足时更严重。
-  ：参数的近似或者依赖采样。

MLE：

数据挖掘实验室 Data Mining Lab

在结束这个话题之前

甜点2：M-投影

简单的MLE很简单，能否有一个泛化的求θ方法呢？

命题17.2 设D是一个数据集，那么

$$\log L(\boldsymbol{\theta}: D) = M \cdot \boldsymbol{E}_{\hat{P}_D}[\log P(\chi: \boldsymbol{\theta})]$$

定理17.1 参数族中相对于数据集D的MLE$\boldsymbol{\theta}$是$\hat{P}_D$到这个参数族的M-投影：

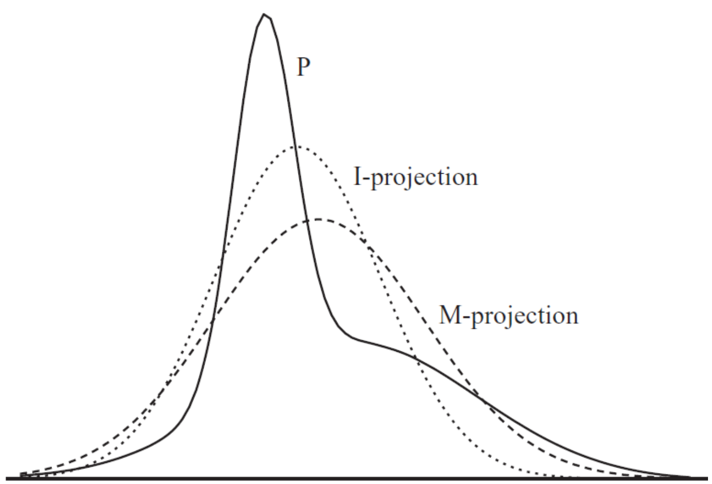
$$\hat{\boldsymbol{\theta}} = argmin_{\theta\in\Theta} D\left(\hat{P}_D || P_\theta\right)$$

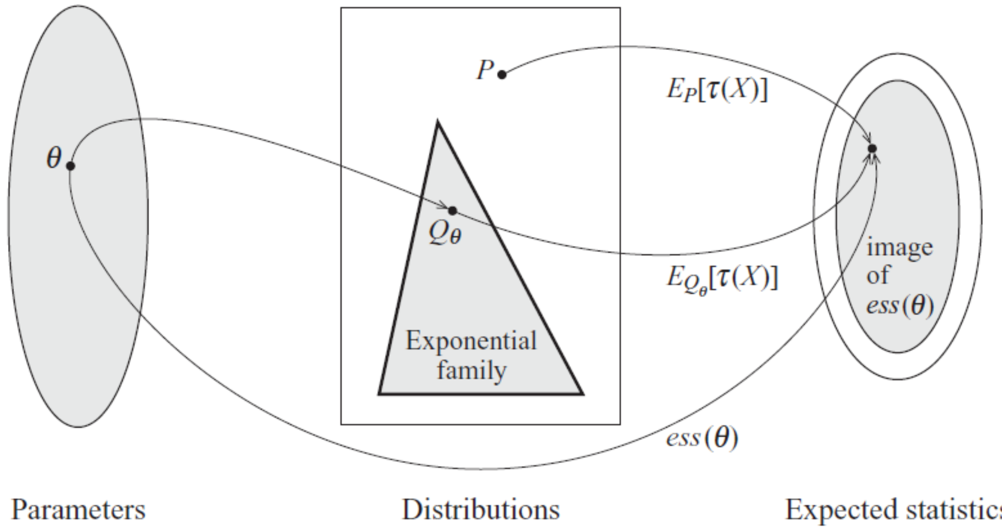
注意到M-投影满足$E_{Q_\theta}[\tau(\chi)] = E_P[\tau(\chi)]$。如果我们的CPD属于一个指数族，并且从参数到充分统计量的映射ess可逆，那么我们可以简单地根据$\hat{P}_D$选择充分统计量，然后求逆映射来生成MLE。

MLE：

在结束这个话题之前

甜点2：M-投影





注意到M-投影满足$E_{Q_\theta}[\tau(\chi)] = E_P[\tau(\chi)]$。如果我们的CPD属于一个指数族，并且从参数到充分统计量的映射ess可逆，那么我们可以简单地根据$\hat{P}_D$选择充分统计量，然后求逆映射来生成MLE。

贝叶斯参数估计：

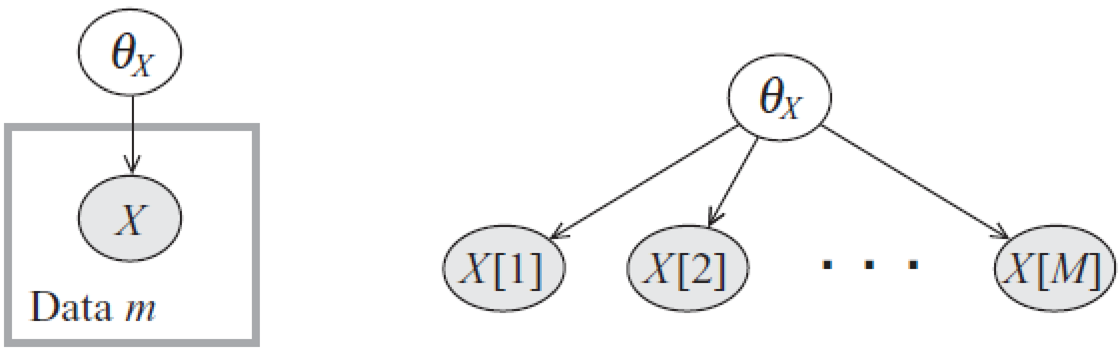
图钉流血案，no，抛硬币

嗯，图钉3+in 10→θ0.3似fu很有道理，但假如硬币3+in 10呢？不好说了吧，因为老子抛硬币经验丰富得很！不过，你要是抛出300k+in 1m，我就信那是个有偏的硬币。

Donald Bayesian 川普

贝叶斯参数估计：

川普抛硬币



(a) (b)

联合分布

$$P(x[1],\ldots,x[M],\theta)=P(x[1],\ldots,x[M]|\theta)P(\theta)$$

$$=P(\theta)\prod_{m=1}^{M}P(x[m]|\theta)=P(\theta)\theta^{M[1]}(1-\theta)^{M[0]}$$

后验

$$P(\theta|x[1],\ldots,x[M])=\frac{P(x[1],\ldots,x[M]|\theta)P(\theta)}{P(x[1],\ldots,x[M])}$$

贝叶斯参数估计：

预测

后验的一个重要作用—预测：

假如我们要预测一个新的实例$x[M+1]$，

$$P(x[M+1]|x[1],\ldots,x[M])$$
$$=\int P(x[M+1]|\theta,x[1],\ldots,x[M])P(\theta|x[1],\ldots,x[M])\mathrm{d}\theta$$
$$=\int P(x[M+1]|\theta)P(\theta|x[1],\ldots,x[M])\mathrm{d}\theta$$

考虑$X[M+1]=x^1$的情况

$$P(x[M+1]=x^1|x[1],\ldots,x[M])=\frac{1}{P(x)}\int\theta\theta^{M[1]}(1-\theta)^{M[0]}\,\mathrm{d}\theta$$
$$=\frac{M[1]+1}{M[1]+M[0]+2}$$

拉普拉斯矫正。

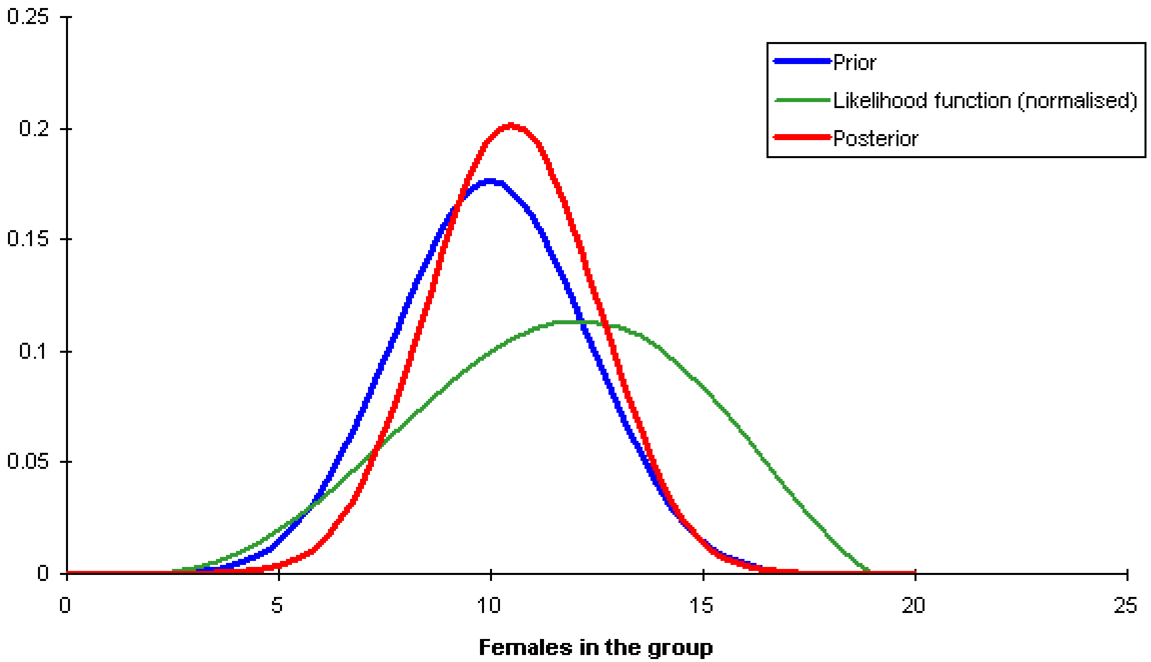
贝叶斯参数估计：

先验：

假如给川普抛硬币加一个beta先验：

$$P(x[M+1]=x^1|x[1],\ldots,x[M])=\frac{M[1]+\alpha_1}{M+\alpha}$$

先验是我们的经验或者是D数据之外的数据D'。MLE是最大化似然vsB派是最大化后验；MLE是考虑D，B派是考虑D,D'。



贝叶斯参数估计：

先验分布&后验分布：

这个已经讲得灰常灰常多了！

略

贝叶斯参数估计：

先验分布&后验分布：

贝叶斯参数估计：

先验分布&后验分布：

后验估计

$$P(\xi[M+1]|D) = \int P(\xi[M+1]|\theta)P(\theta|D)d\theta$$

$$= E_{P(\theta|D)}[P(\xi[M+1]|\theta)]$$

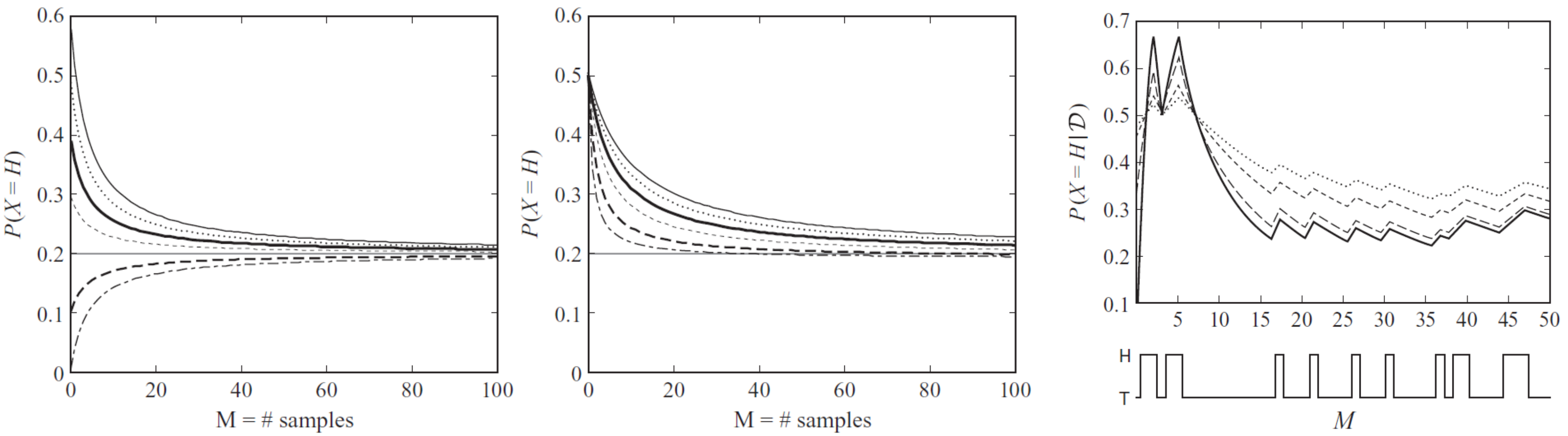
对于那些有封闭解的问题，此积分不算什么，但没有封闭解呢？观察上面的期望项，它正是后验分布上的期望啊，对于*Dirichlet*分布，$E[\theta_k] = \alpha_k/\alpha$，容易求出$P\left(x[M+1] = x^k|D\right) = \frac{M[k]+\alpha_k}{M+\alpha}$。

贝叶斯参数估计：

数据挖掘实验室
Data Mining Lab

先验分布&后验分布：

先验超参数的影响：对于*Dirichlet*的栗子，可以将其超参数写成$\alpha_k = \alpha\theta'_k$。

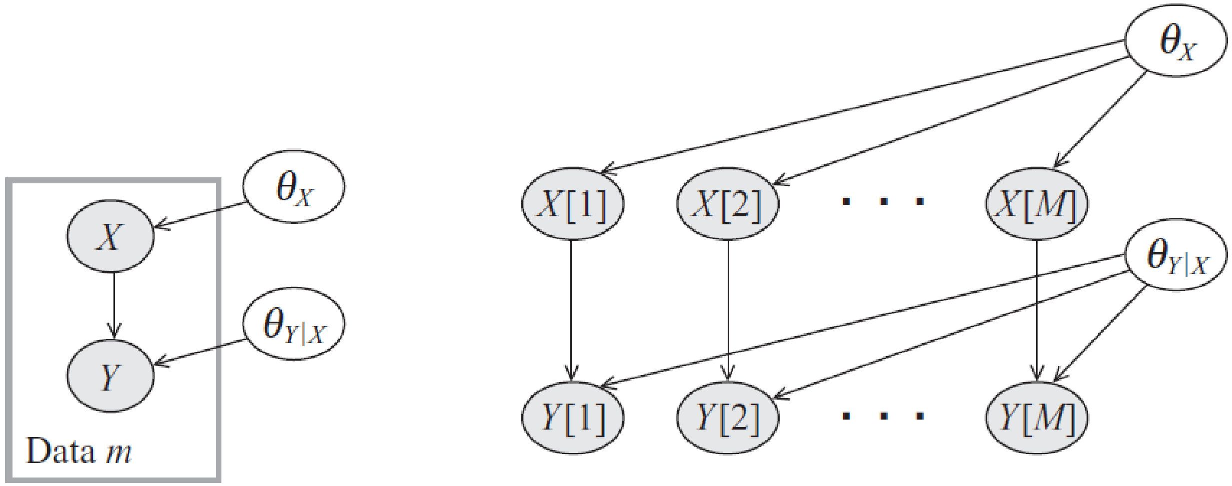


贝叶斯网络中参数估计：

参数独立性与分解：

其实我们第3章就学过了这类分解！

➢ 全局参数独立性：



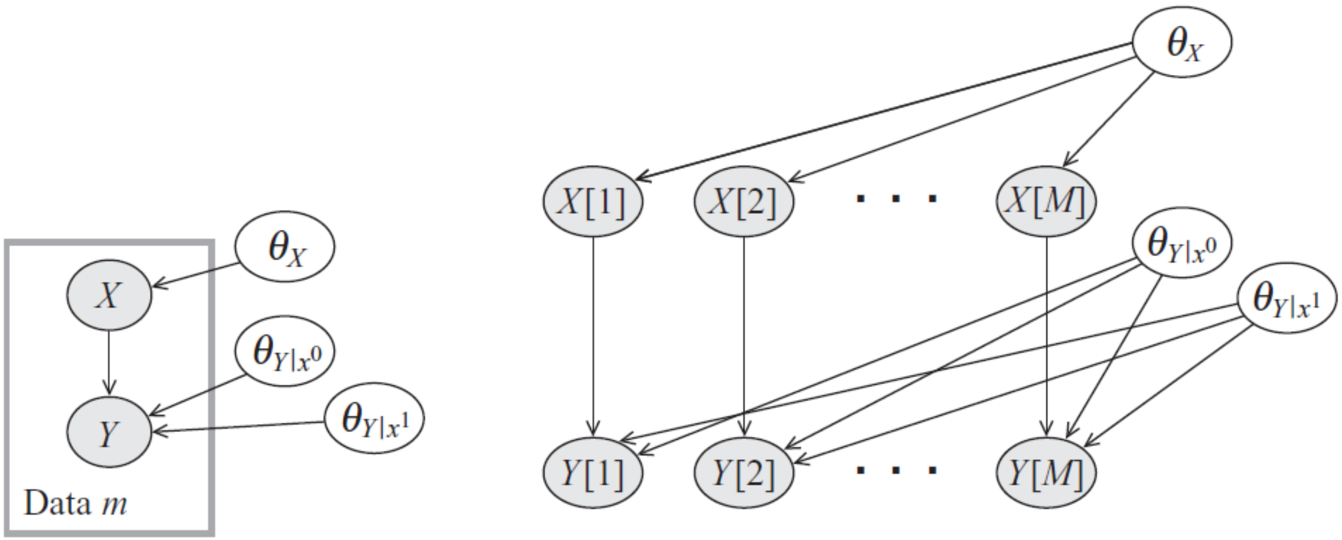
(a) (b)

$$P(\boldsymbol{\theta}|D) = \prod_i P\left(\boldsymbol{\theta}_{X_i|Pa_{X_i}}|D\right)$$

贝叶斯网络中参数估计：

参数独立性与分解：

➢ 局部参数独立性：



(a) (b)

$$P(\boldsymbol{\theta}|D) = \prod_{i} \prod_{Pa_{X_i}} P\left(\boldsymbol{\theta}_{X_i|Pa_{X_i}}|D\right)$$

贝叶斯网络中参数估计：

如何评价贝叶斯先验：

➢ 超参数对结果的影响：

?定义*Dirichlet*的参数α，一种简单的方法是令$\alpha_{x_i|pa_{X_i}}=\alpha[x_i, pa_{X_i}]$

假定有一个假想的“先验”实例数据集D'。等价于MLE$\langle D, D'\rangle$。这种方法的问题，他需要存储一个可能很大的数据集D'。作为替代，可以存储数据集的规模α和在这个先验数据集中时间的频率的一个表示$P'(X_1,\ldots,X_n)$，于是$\alpha_{x_i|pa_{X_i}}=\alpha\cdot P'(x_i, pa_{X_i})$。不明请听于伯伯大哥解释 >_<。

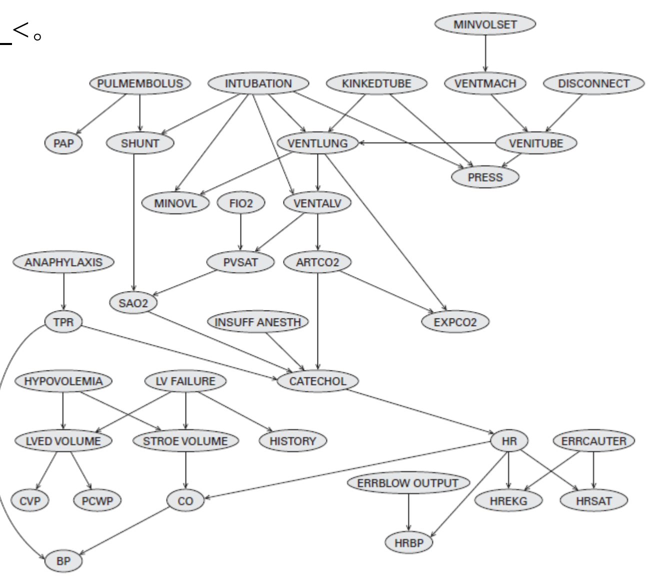
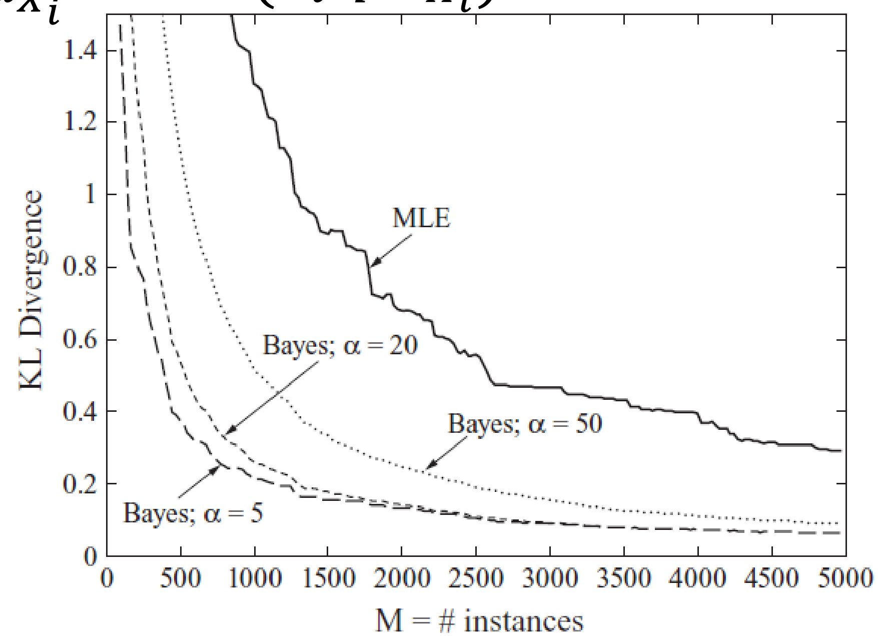


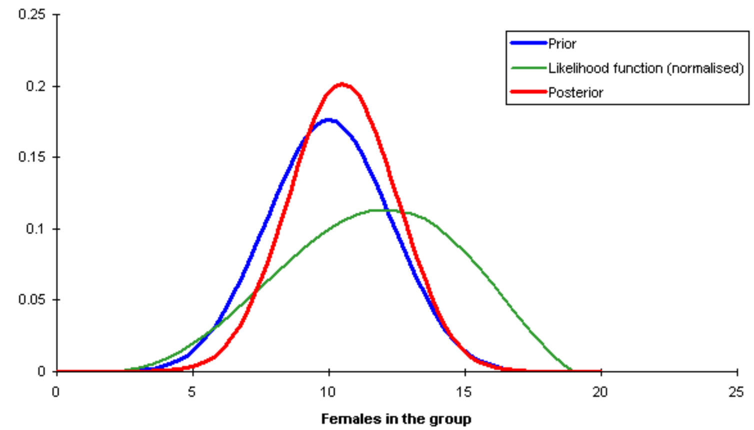
Figure 17.C.1 — The ICU-Alarm Bayesian network.



贝叶斯网络中参数估计：

MAP估计：

目前我们考虑的贝叶斯推断问题都是有封闭解的，当没有的时候，就要用MAP了。



表示独立性representation independence：继续回到图钉栗子，假如我们选用参数η，所以$P'(X=H|\eta)=\frac{1}{1+e^{-\eta}}$。于是有$\eta=\log\frac{\theta}{1-\theta}$，$\eta$与$\theta$之间存在一一对应的关系。这两种参数表示的是同一分布。

这个性质表示是否参数敏感。

贝叶斯网络中参数估计：

数据挖掘实验室
Data Mining Lab

MAP估计：

- MLE对重新参数化不敏感；
- B推断（当使用谨慎）参数不敏感；
- MAP参数敏感。

一盘栗子说明：

✓ 对于MLE，如果$\hat{\eta}$是MLE，那么对应的$\theta(\hat{\eta})$也是MLE；

✓ 对于B推断，考虑贝努力分布-beta先验：

$$P(\theta:\alpha_0,\alpha_1) = c\theta^{\alpha_1-1}\theta^{\alpha_0-1}$$

令

$$\theta = 1/1+e^{-\eta}; \eta = log\,\theta/1-\theta$$

一系列积分后...

$$P(\eta) = c\left(\frac{1}{1+e^{-\eta}}\right)^{\alpha_1}\left(\frac{1}{1+e^{\eta}}\right)^{\alpha_0}$$

样子都一样，yeah！，可惜θ的定义域变成整个实数了，WUWU

贝叶斯网络中参数估计：

MAP估计：

- MLE对重新参数化不敏感；
- B推断（当使用谨慎）参数不敏感；
- MAP参数敏感。

一盘栗子说明：

✓ 对于MAP，继续考虑贝努力分布-beta先验：

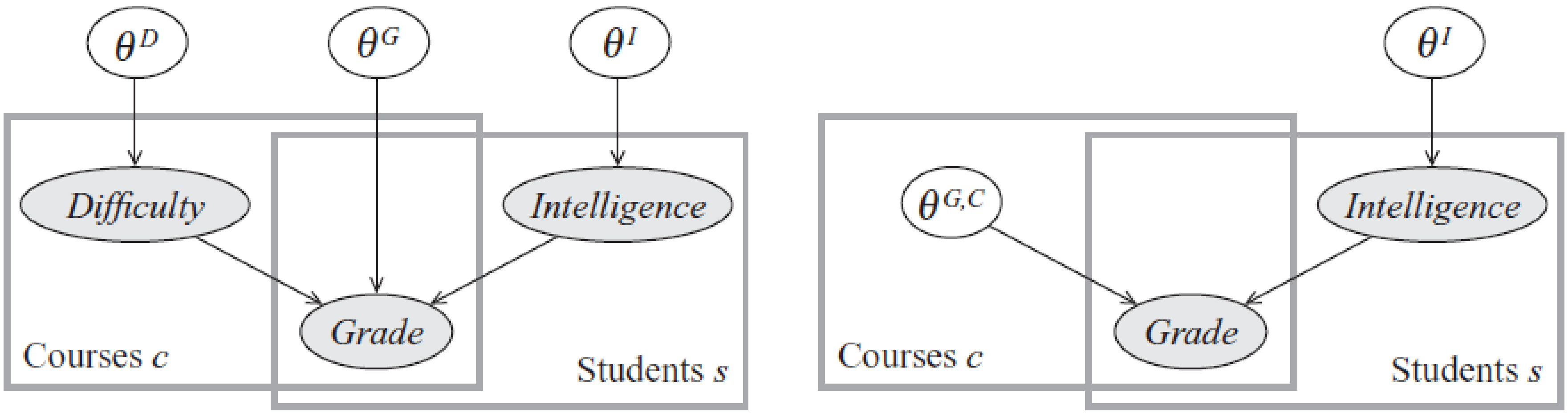
$$\widetilde{\boldsymbol{\theta}} = \operatorname{argmax} \log P(\theta) = \frac{\alpha_1 - 1}{\alpha_0 + \alpha_1 - 2}$$

另一方面，

$$\tilde{\eta} = \operatorname{argmax} \log P(\eta) = log\frac{\alpha_1}{\alpha_0}; \boldsymbol{\theta}(\tilde{\eta}) = \frac{\alpha_1}{\alpha_0 + \alpha_1}$$

共享参数学习模型：

参数共享：



(a) (b)

- 引入一些符号：

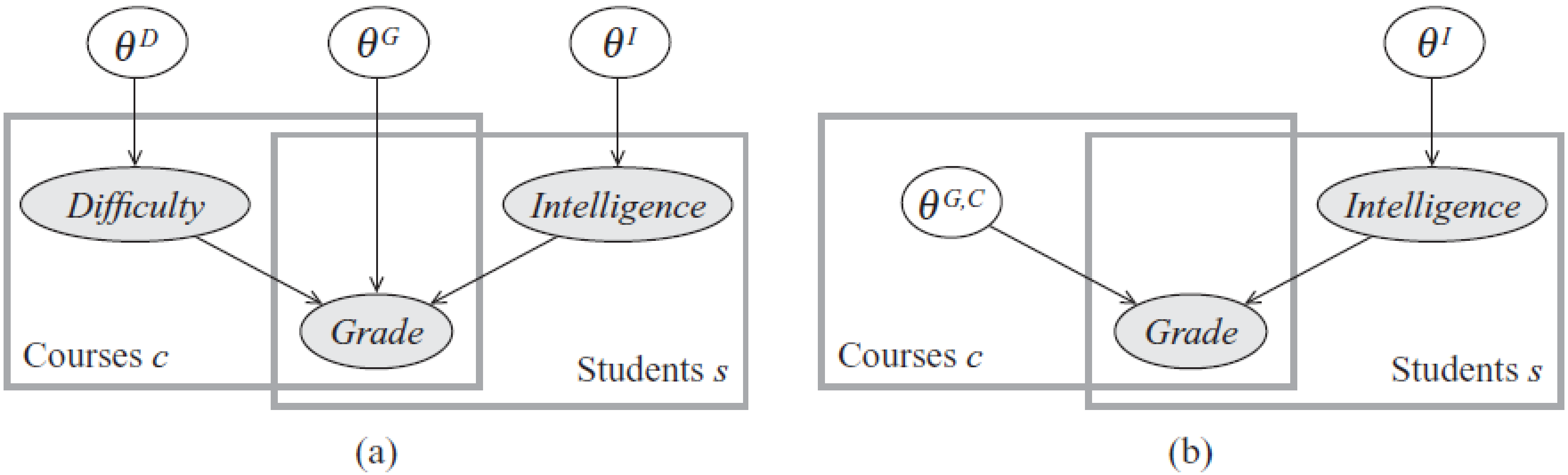
  $\mathcal{V}^k$是变量集χ的划分

  y_k包含了每个变量$X_i \in \mathcal{V}^k$的可能值

  ω_k包含了其父节点的可能值

共享参数学习模型：

参数共享：



$$L(D:\theta) = \prod_{k} \prod_{y_k,\omega_k} \prod_{\substack{X_i \in \mathcal{V}^k: \\ x_i = y_k, u_i = \omega_k}} \theta^k_{y_k|\omega_k} = \prod_{k} \prod_{y_k,\omega_k} \left(\theta^k_{y_k|\omega_k}\right)^{\check{M}_k[y_k,\omega_k]}$$

共享参数学习模型：

共享参数的贝叶斯推断：

考虑具体的预测问题

$$P(\xi[M+1]|D)=\int P(\xi[M+1]|\theta)P(\theta|D)d\theta$$

在我们的栗子中$P(\xi[M+1]|\theta)=\prod_i \theta_{x_i[M+1]|u_i[M+1]}$，这些参数的后验独立，于是

$$P(\xi[M+1]|D)=\prod_i E\left[\theta_{x_i[M+1]|u_i[M+1]}|D\right]$$

每个期望都建立 在关于$\theta_{x_i[M+1]|u_i[M+1]}$的后验上。

总结：

- MLE的思想
  - 似然函数，极大似然估计；
  - 贝叶斯网上的MLE，似然分解；
- 非参数模型，M-投影；
- 贝叶斯推断
  - 先验，后验，共轭；
  - 期望in预测；
- 贝叶斯网络参数估计
  - 参数独立性；
  - 似然分解；
  - MAP的一些事；
- 共享参数模型。

电子科技大学
University of Electronic Science and Technology of China

Thanks!

Data Mining Lab,
Big Data Research Center, UESTC