概率图模型理论及应用

Theory and Applications of Probabilistic Graphical Models

(Lesson 3)

欧智坚

清华大学电子工程系

Addr: 罗姆楼 6-104

Tel: 62796193

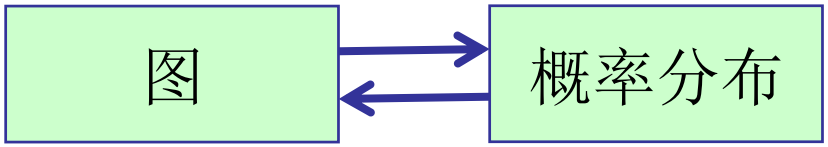
Email: ozj@tsinghua.edu.cn

课程章节

- 第一章 引言（1）
- 第二章 图模型的表示理论（2）
  - Semantics (DGM, UGM)
  - HMM, CRF
- 第三章 图模型的推理理论（6）
  - 精确推理：variable-elimination，cluster-tree，triangulate
  - 连续变量：Kalman
  - 采样近似：sampling
  - 变分近似：variational
- 第四章 图模型的学习理论（3）
  - 参数学习：maxlikelihoodEstimate，RFLearning, BayesEstimate
  - 结构学习：StructureLearning
- 第五章 一个综合例子（1）

DAG的语义

一个图表示了怎样的概率分布



一个概率分布如何表示成一个图

一个DAG表示一个怎样的概率分布（by definition）

- 满足有向分解性
- 满足有向有序Markov性

一个概率分布如何表示成一个DAG

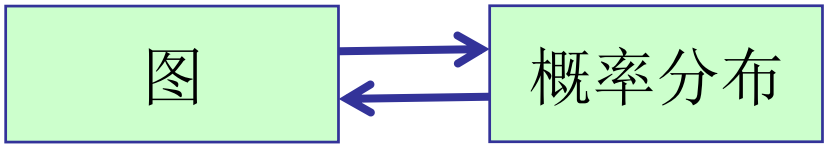
- 从条件独立性的包含关系上，寻找尽可能紧凑的I-map

❖ DAG适用于随机因素间具有明确依赖、影响关系，通过条件分布来表达的建模问题

- 通过画图来建模

UGM的语义

一个图表示了怎样的概率分布



一个概率分布如何表示成一个图

一个UGM表示一个怎样的概率分布（by definition）

- 满足 分解性
- 满足 Markov性

一个概率分布如何表示成一个UGM

- 从条件独立性的包含关系上，寻找尽可能紧凑的I-map

UGM适用于随机因素间彼此影响，通过局部函数的倾向性取值来表达的建模问题

- 局部函数$\phi(x_C)$ 体现 局部变量x_C取值的随机规律

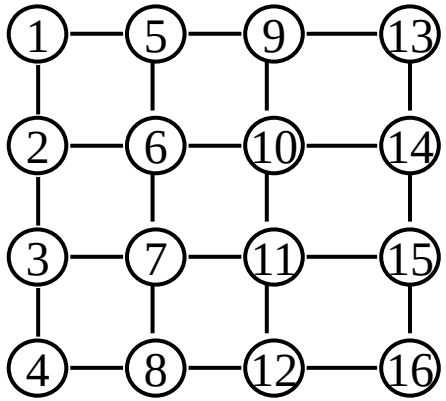
Ising model

x_1 \ x_5	-1	1
-1	e^{β}	$e^{-\beta}$
1	$e^{-\beta}$	e^{β}

$\phi(x_1, x_5)$

- Consider a lattice of binary RV's, $x_i \in \{-1,1\}$

$$p\left(x_{1:N^2}\right) \propto \exp\left\{\sum_{i-j}\phi\left(x_i, x_j\right)\right\} = \exp\left\{\beta\sum_{i-j}x_i x_j\right\} \quad \beta > 0$$



- β 表示 相邻变量取相同值 的可能性大小.
- β 取不同值时Ising model的随机采样结果:

β = ?

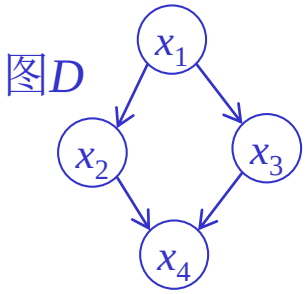
β = 1

β = ?

Definition－DAG的(DG)性质

- 称一个多变量的分布 $p(x_V)$ 服从依图D的有向全局Markov性（directed global Markov property），如果对图D中任意互不相交的三个结点子集 (A, B, S)，其中 S 有向隔离了 A 和 B，成立

$$A \perp B \mid S$$



$p(x_1, x_2, x_3, x_4)$

Definition: d-separation

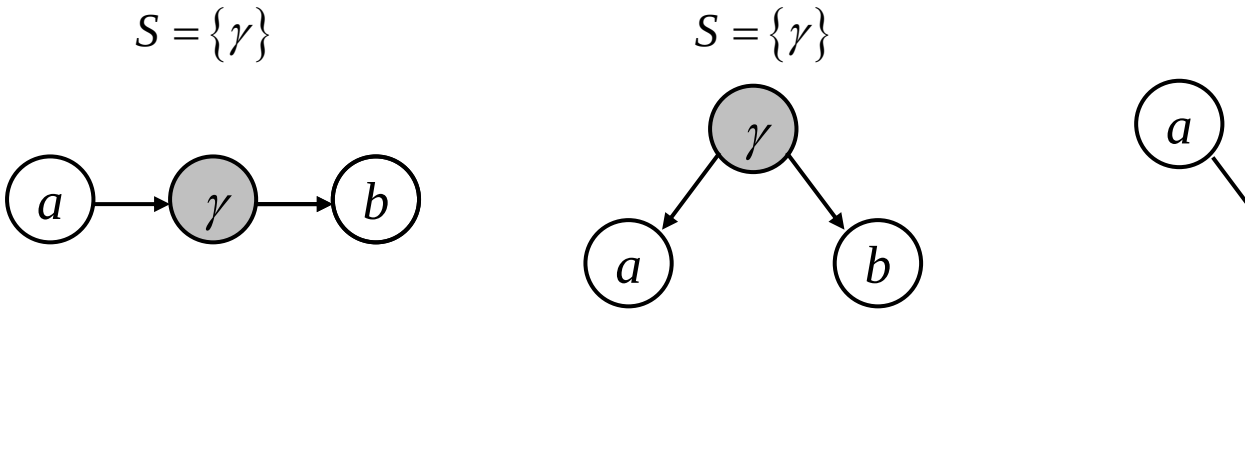
- 称结点集A和结点集B被结点集S有向隔离(d-separated)，
  如果从A到B的所有迹(all trails)都被S有向隔断(d-blocked)

- 称从结点a到结点b的一条迹π 被 结点集 S 有向隔断

  如果 下列条件之一成立，

  ① 迹π 上存在一个结点 $\gamma \in \pi$， $\gamma \in S$，并且迹 π 在结点γ 处的箭头不是V连接

  ② 迹π 上存在一个结点 $\gamma \in \pi$， γ 与它的所有后代结点均 $\notin S$，并且迹 π 在结点γ 处的箭头是V连接



课程章节

- 第一章 引言（1）
- 第二章 图模型的表示理论（2）
  - **Semantics (DGM, UGM)**
  - **HMM, CRF**
- 第三章 图模型的推理理论（6）
  - 精确推理：variable-elimination，cluster-tree，triangulate
  - 连续变量：Kalman
  - 采样近似：sampling
  - 变分近似：variational
- 第四章 图模型的学习理论（3）
  - 参数学习：maxlikelihoodEstimate，UGMLearning, BayesEstimate
  - 结构学习：StructureLearning
- 第五章 一个综合例子（1）

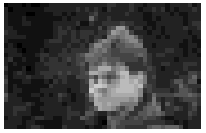
第二章 图模型的表示理论

Hidden Markov Model (HMM)
隐含马尔可夫模型

Rabiner. “A tutorial on hidden Markov models and selected applications in speech recognition”, Proceedings of IEEE, 1989.

Introduction

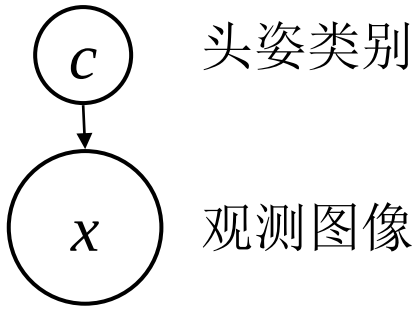
- 400幅图像（44*28=1232）
  - 摘要：找出头姿的代表图像

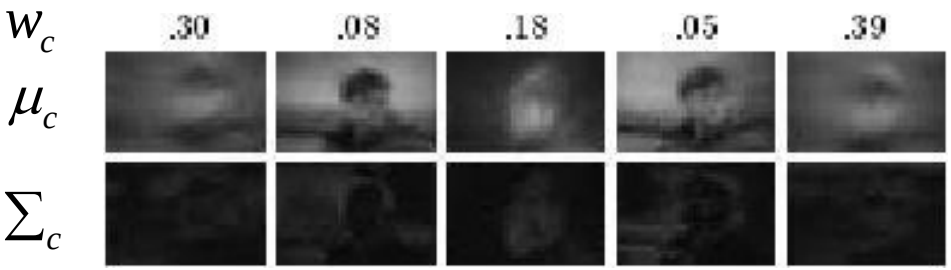


高斯混合模型

$$p(c,x)=p(c)\,p(x|c)$$

$$p(x)=\sum_{c=1}^{K} w_c\, N(x|\mu_c,\Sigma_c)$$

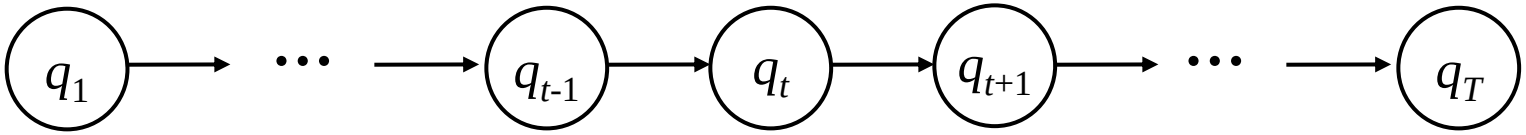




- 对时序数据的建模

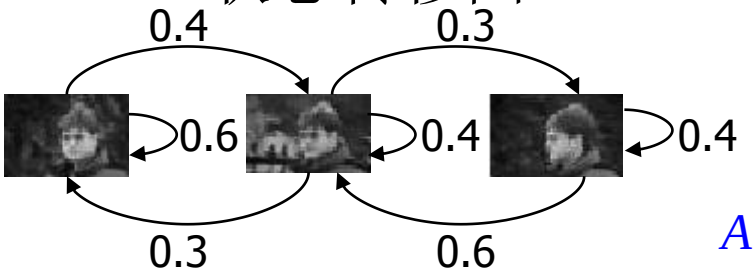
马氏链

- 时间离散, 状态离散的Markov过程$\{q_t, t = 1,2,\dots\}$
  - 如果 $q_t = i$, 称马氏链在时刻 t 驻留在状态 i， i=1,...,N



$$p(q_{t+1} = j \mid q_t = i, q_{t-1} \cdots) = p(q_{t+1} = j \mid q_t = i)$$

状态转移图



状态初始分布

$$\pi = \{p(q_1 = i)\}_{i=1:N}$$

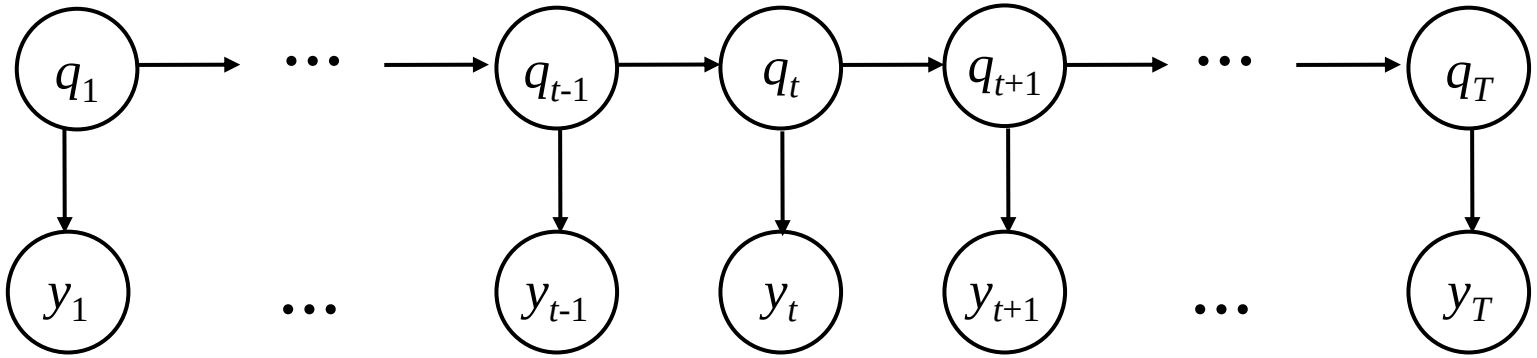
正脸	斜脸	侧脸
1.0	0.0	0.0

状态转移矩阵

$$A = \{p(q_{t+1} = j \mid q_t = i)\}_{i,j=1:N}$$

	正脸	斜脸	侧脸
正脸	0.6	0.4	0
斜脸	0.3	0.4	0.3
侧脸	0	0.6	0.4

HMM viewed as DGM



- The joint probability distribution

$$p(q_{1:T}, y_{1:T}) = p(q_1) \cdot \prod_{t=1}^{T-1} p(q_{t+1} \mid q_t) \cdot \prod_{t=1}^{T} p(y_t \mid q_t)$$

$\lambda = (\pi, A, B)$

状态输出分布

$$B = \{p(y_t = k \mid q_t = i)\}_{i=1:N, k=1:V}$$

- 400帧图像视为HMM的一个样本

Three basic problems for HMM

① Learning/training (parameter estimation)——EM算法

- 给定一个观测值序列 $y_{1:T}$ ，如何估计模型参数 $\lambda=(\pi, A, B)$ ？

$$\max_{\lambda} p(y_{1:T} \mid \lambda)$$

概率分布函数　似然函数

http://en.wikipedia.org/wiki/Likelihood_function

给定参数λ下，随机变量Y特定取值y的概率（密度）值 $p(y \mid \lambda)$ 视为

给定随机变量Y特定取值y下，参数λ的似然值

② Likelihood calculation——Forward-backward算法

- 给定一个模型$\lambda=(\pi, A, B)$ ，如何能有效计算一个观测值序列 $y_{1:T}$ 的概率值 ？

$$p(y_{1:T} \mid \lambda)$$

Inference problem

- 考虑朴素计算

$$p(y_{1:T}) = \sum_{q_1}\sum_{q_2}\cdots\sum_{q_t}\cdots\sum_{q_T} p(q_{1:T}, y_{1:T})$$

  - 需要 $O(N^T)$ 求和

- 定义

$$\alpha(q_t) = p(y_{1:t}, q_t)$$
$$\beta(q_t) = p(y_{t+1:T} \mid q_t)$$

for $q_t = 1, \cdots, N$ (L.E. Baum, et al, 1966)

$$p(y_{1:T}) = \sum_{q_t=1}^{N} p(y_{1:T}, q_t) = \sum_{q_t=1}^{N} \alpha(q_t)\beta(q_t)$$

- 试证明：

$$\alpha(q_{t+1}) = p(y_{t+1} \mid q_{t+1}) \sum_{q_t=1}^{N} p(q_{t+1} \mid q_t)\alpha(q_t)$$

需要 $O(TN^2)$ 求和

$$\beta(q_t) = \sum_{q_{t+1}=1}^{N} p(y_{t+1} \mid q_{t+1})\beta(q_{t+1})p(q_{t+1} \mid q_t)$$

需要 $O(TN^2)$ 求和

Three basic problems for HMM

① Learning/training (parameter estimation)——EM算法

Lesson?_mlEstimate

② Likelihood calculation——Forward-backward算法

Lesson?_ve

③ Decoding (recognition)——Viterbi算法

- 给定一个模型 $\lambda=(\pi, A, B)$ 和一个观测值序列 $y_{1:T}$，如何寻找'最优'状态序列？

$$\max_{q_{1:T}} p\left(q_{1:T} \mid \lambda, y_{1:T}\right)$$

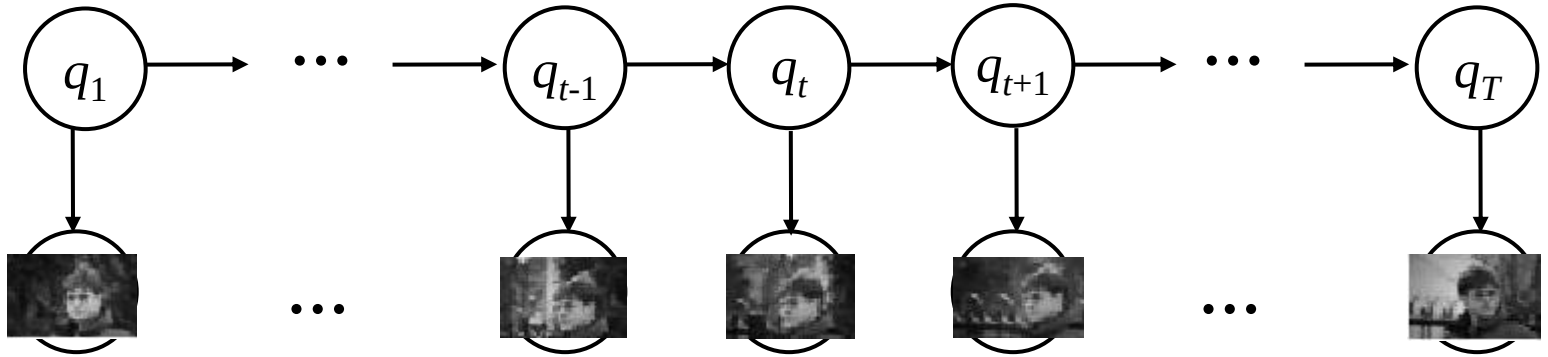


Image Processing / Computer Vision

Removal of sensor noise

Image stabilization

Object tracking

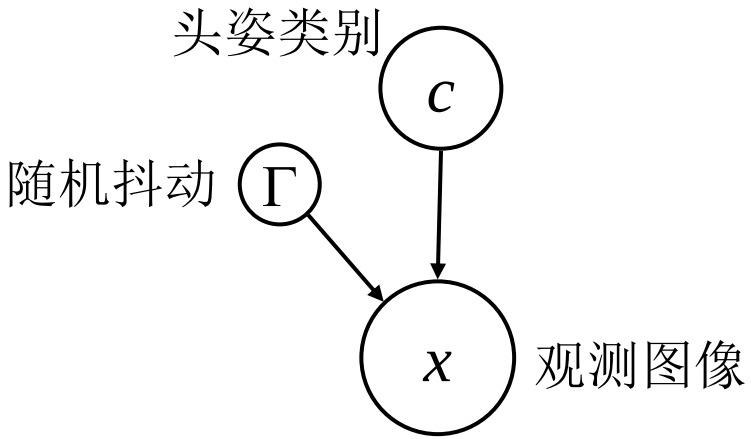
Video summary

Transformed mixture of Gaussians (TMG)

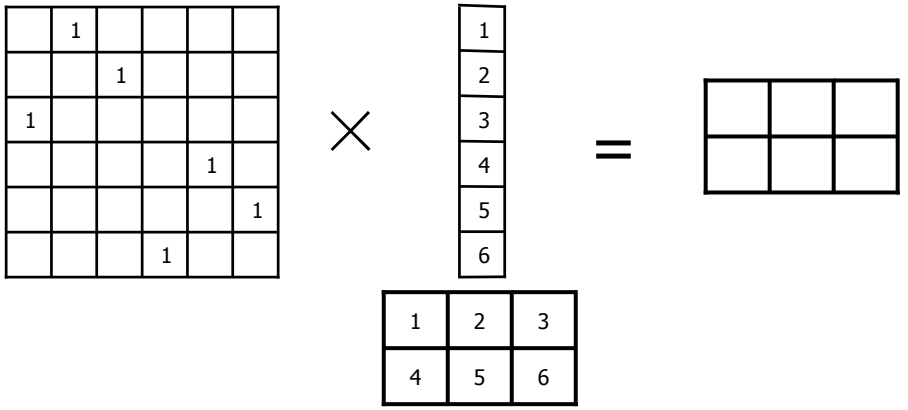
Introduce transformation as hidden variable

$$p(c,\Gamma,x) = p(c)\,p(\Gamma)\,\underline{p(x|c,\Gamma)}$$

$$N\left(x|\Gamma\mu_c,\Gamma\Sigma_c\Gamma^T\right)$$

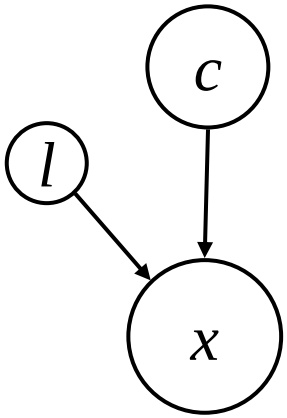


L=121 translation transformations
(11 horizontal shift and 11 vertical shift)



$$p(c,l,x) = p(c)\,p(l)\,\underline{p(x|c,l)}$$

$$N\left(x|\Gamma_l\mu_c,\Gamma_l\Sigma_c\Gamma_l^T\right)$$

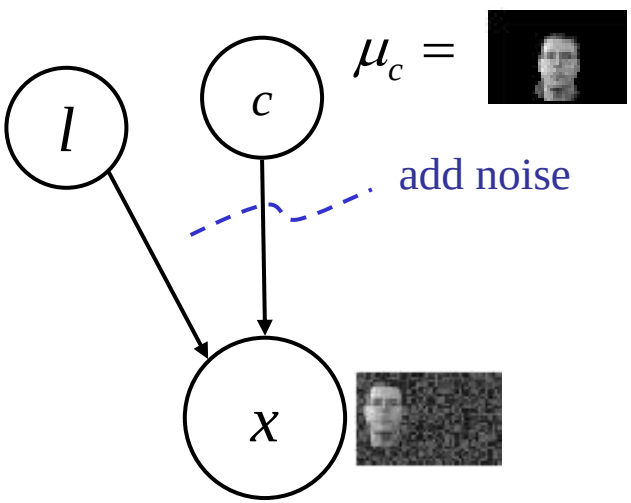


Transformed mixture of Gaussians (TMG)

Introduce transformation as hidden variable

$$p(c,l,x) = p(c)\,p(l)\,p(x|c,l)$$

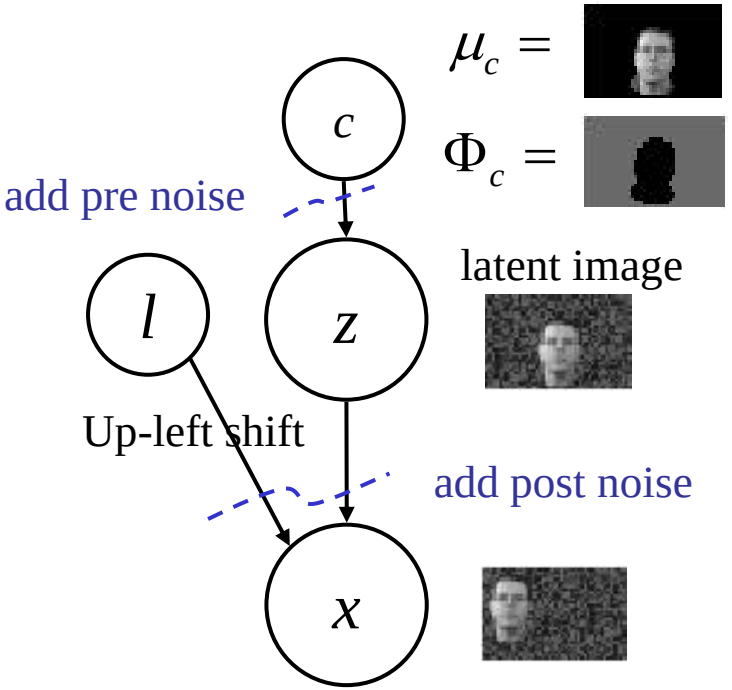
$$N\left(x \mid \Gamma_l \mu_c, \Gamma \Sigma_c \Gamma^T\right)$$



Introduce pre-transformation noise

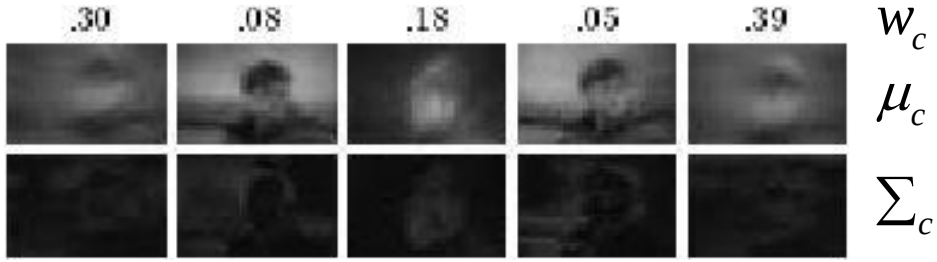
$$p(x,l,z,c) = p(c)\,p(l)\,p(z|c)\,p(x|l,z)$$

$$= \pi_c p_l \mathcal{N}(z \mid \mu_c, \Phi_c)\,\mathcal{N}(x \mid \Gamma_l z, \Psi)$$

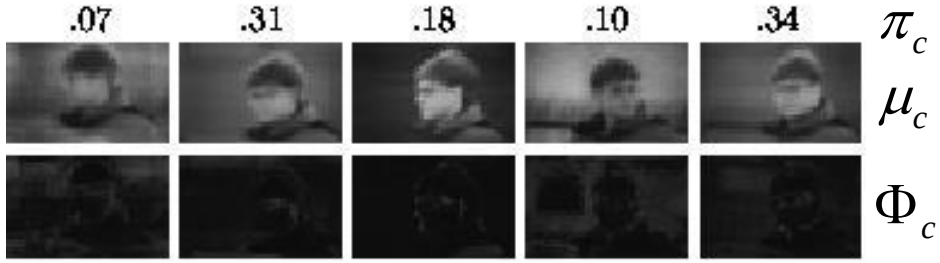


Comparison

- A TMG is trained on 400 images, containing K=5 clusters, and L=121 translation transformations (11 horizontal shift and 11 vertical shift) using 40 EM iterations.



Mixing weights, cluster means and cluster variances for a traditional mixture model.

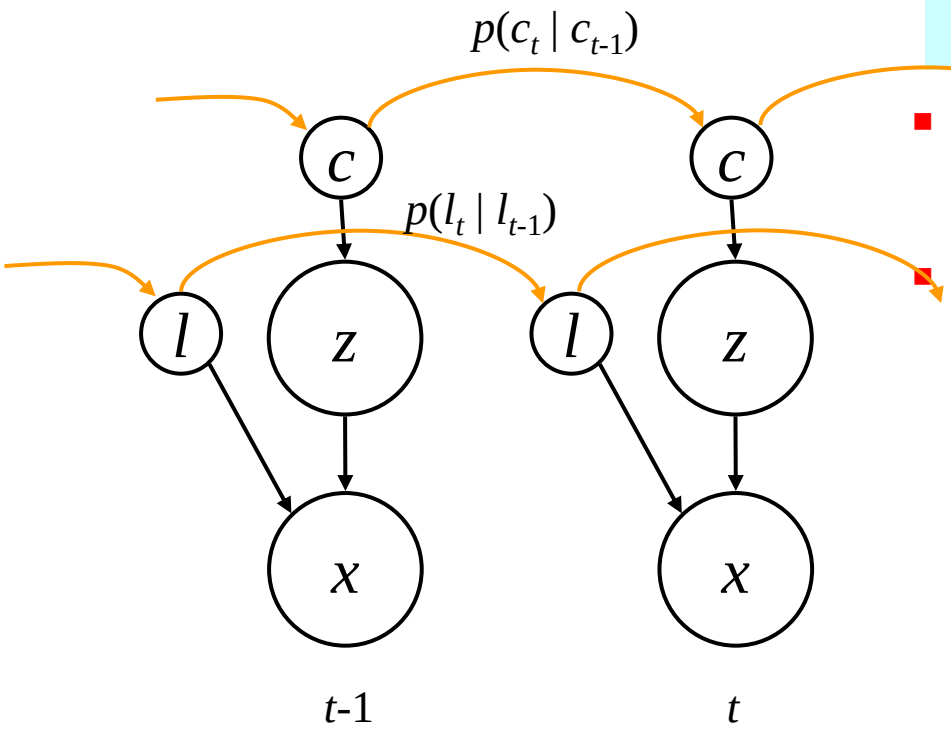


Mixing weights, cluster means and cluster variances for the TMG.

The cluster means are less blurred.

Introducing temporal dependency

- Temporal dependency is obviously valuable for modeling video sequence.



Transformed HMM

- The variables $c_t\ l_t\ z_t\ x_t$ at frame t follows TMG.
- Transition probabilities:

$$p(c_t \mid c_{t-1})$$

$$p(l_t \mid l_{t-1})$$

Image/video analysis as inference in THMM

- Given a THMM, several interesting image/video analysis tasks can be implemented simply as inference of the hidden variables in THMM.

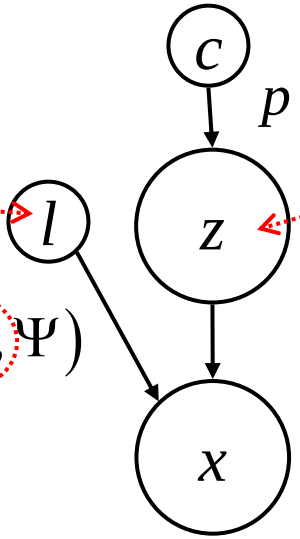
$$\hat{l}_{1:T} = \arg\max_{l_{1:T}} p\left(l_{1:T} \mid x_{1:T}\right)$$

Inferring the transformation index l : object tracking

$$p\left(x \mid l, z\right) = \mathcal{N}\left(x \mid \Gamma_l z, \Psi\right)$$

Inferring the mean of x : removal of sensor noise

$$\left(\hat{l}_{1:T}, \hat{z}_{1:T}\right) = \arg\max_{l_{1:T}, z_{1:T}} p\left(l_{1:T}, z_{1:T} \mid x_{1:T}\right)$$



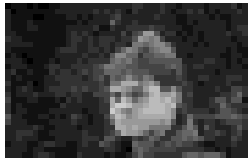
$$p\left(z \mid c\right) = \mathcal{N}\left(z \mid \mu_c, \Phi_c\right)$$

Inferring the latent image z : image stabilization

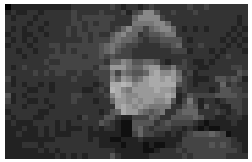
$$\hat{z}_{1:T} = \arg\max_{z_{1:T}} p\left(z_{1:T} \mid x_{1:T}\right)$$

Demo

1) A 400-frame video



2) most likely z

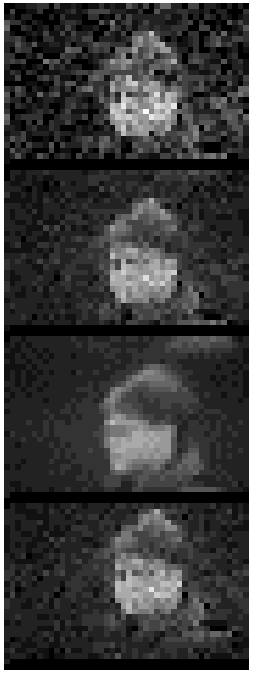


3) 1-adding white noise,

2-removing most likely post-transformation noise

3-further removing most likely pre-transformation noise

4-most likely z



Demo

4) 1-placing a static dark bar

2-removing most likely post-transformation noise

3-most likely z

Open questions

- How can the computer do
  - Segment the video ?
  - Track objects ?
  - Recognize objects ?
  - De-noise the video ?
  - … and do all this completely automatically ?

These problems are inter-dependent and have uncertainties, how to perform them jointly using a unified probability model ?

建立起与问题相适应的概率模型，
运用统计学习和推理进行求解。

第二章 图模型的表示理论

Conditional Random Field (CRF)
条件随机场

Lafferty, McCallum, Pereira. "Conditional random fields: probabilistic models for segmenting and labeling sequeance data", ICML 2001.

$q = q_{1:T}$

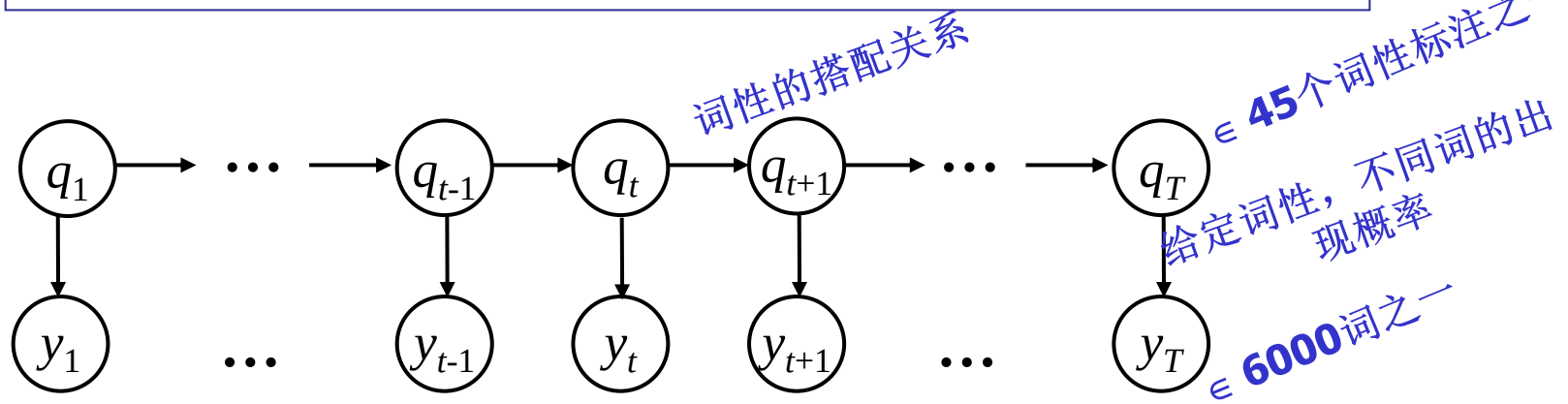
$y = y_{1:T}$

Sequence labeling

- 命名实体识别——标注每个词的属性（人名、地名、机构名、其他）
- 中文分词——标注每个字的词位（词首、词中、词尾、单独成词）

- POS tagging——词性标注 (POS: Part-of-speech)

词序列 $y_{1:T}$:	The	house	is	on	fire
标注序列 $q_{1:T}$:	*def*	*noun*	*verb*	*prep*	*noun*



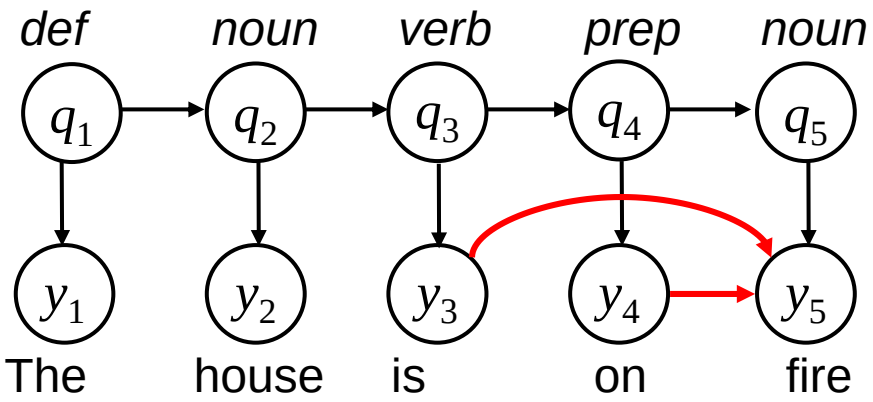
推理 —— p(标注序列q | 观测序列y)

基于 联合分布p(标注序列q , 观测序列y)——隐马模型

Limitations of HMMs

- 对 观测量分布 $p(y)$ 的多余的关注
  - 推理计算只需 $p(q \mid y)$；
  - Generative modeling
    建立联合分布 $p(q, y) = p(q \mid y)\, p(y)$， 如 HMM;

    ☹ 建立联合分布 $p(q, y)$ 意味着 $p(q \mid y)$， $p(y)$ 都要贴近真实的 $p^*(q \mid y)$， $p^*(y)$；



$$p^*(y_5 = \text{fire} \mid q_5 = noun)$$

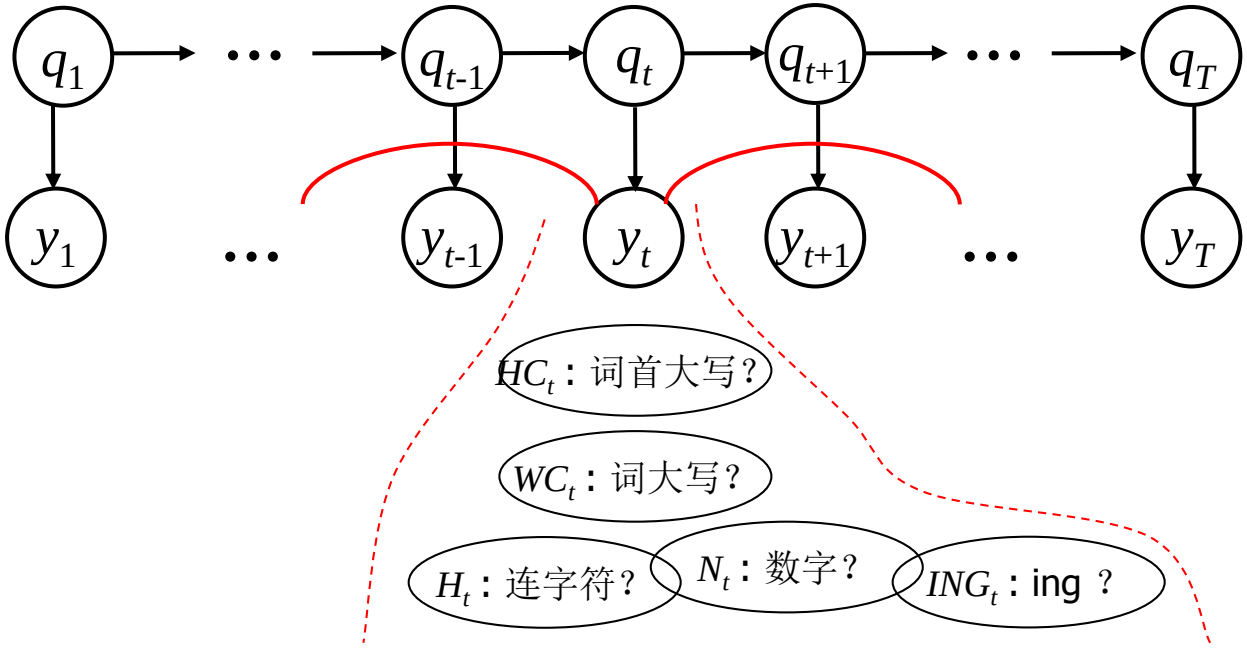
$$p^*(y_5 = \text{fire} \mid q_5 = noun, y_4 = \text{on})$$

Limitations of HMMs

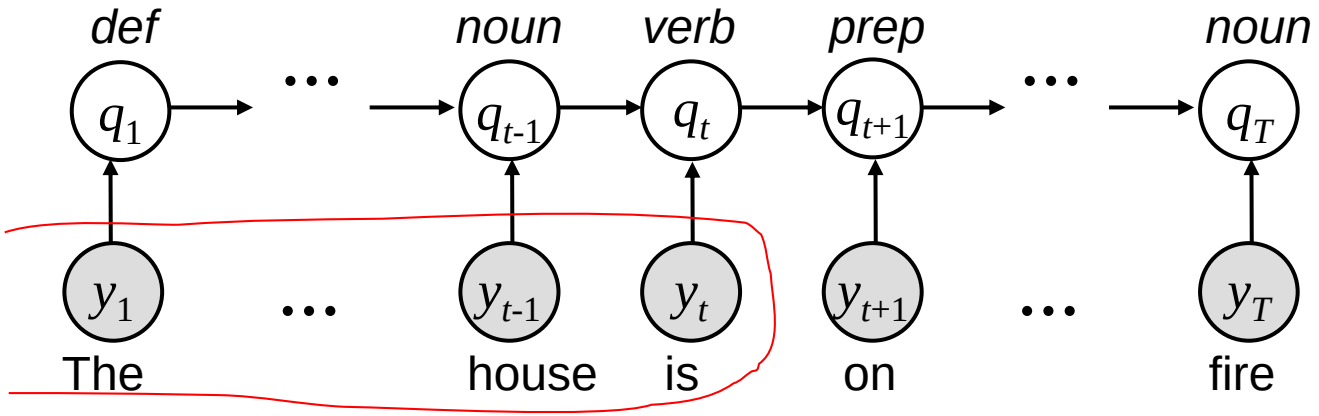
- Conditional modeling/discriminative modeling

  只建立条件分布 $p(q \mid y)$，如 CRF；

  ☺ 允许使用更丰富的特征（e.g. 前后词的搭配关系？词首大写？词大写？含连字符？含数字？含ing？）；



Maximum-Entropy-Markov-Model (MEMM)



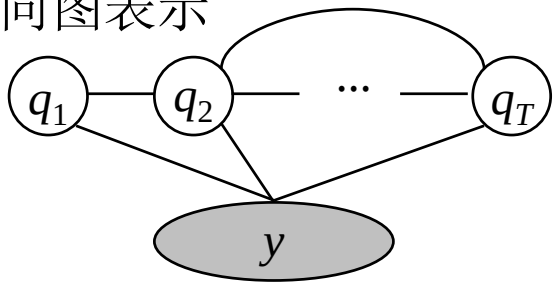
- The conditional distribution

$$p(q_{1:T} \mid y_{1:T}) = p(q_1 \mid y_1) \cdot \prod_{t=2}^{T} p(q_t \mid q_{t-1}, y_t)$$

  - Label bias problem

Conditional Random Field (CRF)

- MRF：无向概率图模型——联合分布的无向图表示
- CRF: 条件分布的无向图表示



目标结点集　观测结点集

- 称条件分布 $p(q \mid y)$ 是依图$g=(V,E)$的条件随机场，如果 q 与V 一一对应；对图g 中所有簇C，存在一个非负函数$\phi_C(q_C, y)$，使得

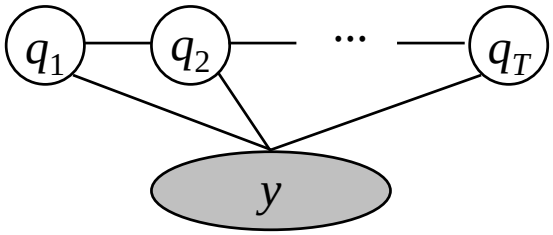
特征函数

$$p(q \mid y) \propto \exp\left\{\sum_{C \in \mathbf{C}} \psi_C(q_C, y)\right\} \qquad p(q \mid y) = \frac{1}{Z(y)} \prod_{C \in \mathbf{C}} \phi_C(q_C, y)$$

其中 $Z(y)$ 是归一化常数 $Z(y) = \sum_q \prod_{C \in \mathbf{C}} \phi_C(q_C, y)$

- $\mathbf{C}$ 是图g中所有簇的集合
- **所有的概率计算与条件独立性分析都是在给定y下讨论的。**

Linear-chain CRF



q_t, q_{t+1}组成的簇上的特征函数 $\psi_t(q_t, q_{t+1}, y)$

q_t 组成的簇上的特征函数 $\psi_t(q_t, y)$

$$p(q_{1:T} \mid y) \propto \exp\left\{\sum_{t=1}^{T-1}\sum_i \lambda_i f_i(q_t, q_{t+1}, y, t) + \sum_{t=1}^{T}\sum_j \mu_j f_j(q_t, y, t)\right\}$$

- 每个特征函数（不管是边特征函数，还是结点特征函数），都实现为一系列示性函数的线性组合。

$$\lambda_i f_i(q_t, q_{t+1}, y, t) = \lambda_i \cdot 1(q_t = prep,\ q_{t+1} = non)$$ i 穷尽所有的tag-tag pair

$$\mu_j f_j(q_t, y, t) = \mu_j \cdot 1(q_t = prep,\ y_t = \text{on})$$ j 穷尽所有的tag-word pair

词本身作为观测特征

$$\mu_j f_j(q_t, y, t) = \mu_j \cdot 1(q_t = adv,\ y_t \text{以ly结尾})$$ j 穷尽所有的tag

新观测特征（可以是任何有意义的观测y的函数）

Linear-chain CRF

$$\mu_j f_j(q_t, y, t) = \mu_j \cdot 1(q_t = non, y_t\text{词首大写})$$ j 穷尽所有的tag

新观测特征（可以是任何有意义的观测y的函数）

Error rates for POS tagging

Model	Error	OOV error
HMM	5.69%	45.99%
CRF	5.55%	48.05%
CRF+	4.27%	23.76%

+ 加入新观测特征

y_t包含连字符

y_t包含后缀-ing

y_t包含后缀-ogy

y_t包含后缀-ed

y_t包含后缀-s

y_t包含后缀-ly

y_t包含后缀-ion

y_t包含后缀-tion

y_t包含后缀-ity

y_t包含后缀-ies

Manning,“统计自然语言处理基础”，p14：统计自然语言处理的困难：对于在语料库中没有出现或者几乎不会出现的词，我们很难预测它们的行为

Three basic problems for CRF

Lesson?_RFLearning

① Learning/training (parameter estimation)——Iterative scaling/拟牛顿法
收敛慢 实用

Lesson?_ve

② Likelihood calculation——Forward-backward算法

③ Decoding (recognition)——Viterbi算法

Open questions

- How can the computer do
  - 进行（中文）分词
  - 词性标注，句法分析
  - 命名实体识别，信息提取
  - 语义消歧与分析
  - 篇章分析
  - … and do all this completely automatically ?

Mannning p8:
语言与认知是随机现象

These problems are inter-dependent and have uncertainties,
how to perform them jointly using a unified probability model ?

建立起与问题相适应的概率模型，
运用统计学习和推理进行求解。

Summary

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 27, NO. 9, SEPTEMBER 2005

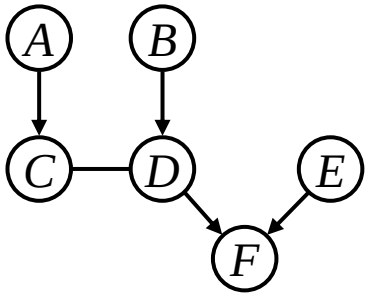
A Comparison of Algorithms for Inference and Learning in Probabilistic Graphical Models

Brendan J. Frey, *Senior Member*, *IEEE*, and Nebojsa Jojic

Abstract—Research into methods for reasoning under uncertainty is currently one of the most exciting areas of artificial intelligence, largely because it has recently become possible to record, store, and process large amounts of data. While impressive achievements have been made in pattern classification problems such as handwritten character recognition, face detection, speaker identification, and prediction of gene function, it is even more exciting that researchers are on the verge of introducing systems that can perform large-scale combinatorial analyses of data, decomposing the data into interacting components. For example, computational methods for automatic scene analysis are now emerging in the computer vision community. These methods decompose an input image into its constituent objects, lighting conditions, motion patterns, etc. Two of the main challenges are finding effective representations and models in specific applications and finding efficient algorithms for inference and learning in these models. In this paper, we advocate the use of graph-based probability models and their associated inference and learning algorithms. We review exact techniques and various approximate, computationally efficient techniques, including iterated conditional modes, the expectation maximization (EM) algorithm, Gibbs sampling, the mean field method, variational techniques, structured variational techniques and the sum-product algorithm, "loopy" belief propagation. We describe how each technique can be applied in a vision model of multiple, occluding objects and contrast the behaviors and performances of the techniques using a unifying cost function, free energy.

图模型的语义——发散思维

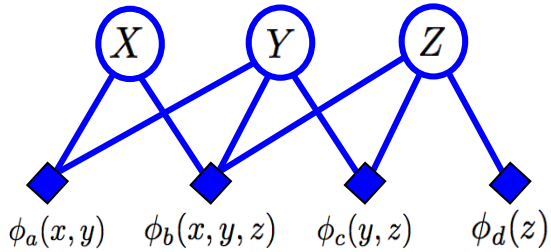
- Chain graph
  - 既有无向边，又有有向边



- Factor graph（因子图） ☺Loopy sum-product ☹CI
  - a bipartite graph（变量结点，函数结点）
  - 连乘积形式的联合分布

$$p(x_{1:5}) = \frac{1}{Z}\phi_A(x_1)\phi_B(x_2)\phi_C(x_1,x_2,x_3)\phi_D(x_3,x_4)\phi_E(x_3,x_5)$$

- Cumulative distribution network
  - 连乘积形式的累积分布
  - Useful for rank data



$$F(x,y,z) = P(X \le x, Y \le y, Z \le z) = \phi_a(x,y)\phi_b(x,y,z)\phi_c(y,z)\phi_d(z)$$

Conf. Uncertainty in Artificial Intelligence (UAI 2008) (Runner-up for Best Student Paper Award)

课程章节

- 第一章 引言（1）
- 第二章 图模型的表示理论（2）
  - Semantics (DGM, UGM)
  - HMM, CRF
- 第三章 图模型的推理理论（6）
  - 精确推理：variable-elimination，cluster-tree，triangulate
  - 连续变量：Kalman
  - 采样近似：sampling
  - 变分近似：variational
- 第四章 图模型的学习理论（3）
  - 参数学习：maxlikelihoodEstimate，RFLearning, BayesEstimate
  - 结构学习：StructureLearning
- 第五章 一个综合例子（1）