Tutorial on Graphical Models

David Heckerman
Microsoft Research

Valencia 7
June 1, 2002

Graphical Model

A graphical representation of a (typically highly multivariate) set of joint distributions

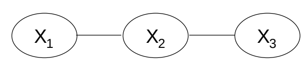
- **Intuitive interface for modeling**
- **Modular: Useful tool for managing complexity**
- **Useful data structure for applying Bayes rule efficiently**
- **Common formalism for many models**
  - **Facilitates transfer of ideas between communities**
  - **Facilitates design of new systems**

Overview

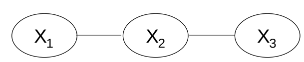
- **Introduction to graphical models**
- **Applications without data: Expert systems**
- **Learning from data**
- **Applications of learning**
- **Influence diagrams: Graphical models for decision making and causal reasoning**

Two popular classes of graphical models

Undirected Graph (UG; MRF; Markov Network)

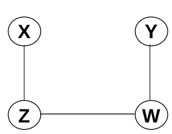


Directed acyclic graph (DAG; Bayesian Network)

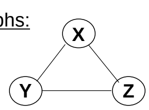


Other types of graphical models

Chain graphs:



Directed cyclic graphs:



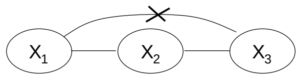
Graphical Model

Assumption for intro: Joint distribution known with certainty

- **Domain: $X = (X_1,\ldots,X_n)$**
- **Graphical model = structure + collection of local distributions**
- **Structure:**
  - **Nodes ~ variables**
  - **Missing arcs ~ conditional independence**
- **Independencies + local distributions => joint distribution (“modularity”)**

Directed Acyclic Graphs

e.g. Wright, 1921; Good, 1961; Howard & Matheson 1981; Pearl 1988



$$p(x_1, x_2, x_3) = p(x_1)\ p(x_2 \mid x_1)\ p(x_3 \mid \cancel{x_1}, x_2)$$

Directed Acyclic Graphs

The DAG structure encodes those independencies that permits the factorization:

$$p(\mathbf{x}) = \prod_i p(x_i \mid \mathbf{pa}_i)$$

parents of x_i

Namely, for any total ordering of the variables consistent with the DAG:

$$p(x_i \mid x_1, \ldots, x_{i-1}) = p(x_i \mid \mathbf{pa}_i)$$

Equivalently, each variable is independent of its non-descendants given its parents (Howard & Matheson 1981).

Directed Acyclic Graphs

Thus, independencies + local distributions yield joint:

$$p(\mathbf{x}) = \prod_i p(x_i \mid \mathbf{pa}_i)$$

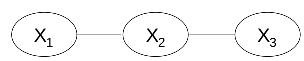
local distributions

Caveat: Local distributions may exist but joint does not.

Undirected Graphs

e.g. Darroch, Lauritzen, & Speed 1980; Whittaker, 1990

Assumption to simplify presentation: p(x) is positive.



$$X_1 \perp X_3 \mid X_2$$

Undirected Graphs

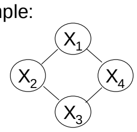
Each variable is independent of all other variables given its neighbors in the graph.

If p(x) is positive, then (Hammersly-Clifford-Besag):

$$p(\mathbf{x}) = \prod_i f_i(\mathbf{x}_{c_i})$$

maximal cliques of the graph

Example:



$$p(\mathbf{x}) = f_1(x_1, x_2) \cdot f_2(x_2, x_3) \cdot f_3(x_3, x_4) \cdot f_4(x_4, x_1)$$

Undirected Graphs

$$p(\mathbf{x}) = \prod_i f_i(\mathbf{x}_{c_i})$$

When working with contingency tables or the case where p(x) is a multivariate Gaussian:

Can generate joint from clique marginals $p(x_{ci})$ using Iterative Proportional Scaling (Deming and Stephan 1940).

Note: $p(x_{ci})$ are local distributions.

Iterative Proportional Scaling

E.g., for contingency table:

- Intialize $p_{old}(x)$ to be uniform

- Iterate, cycling over cliques c_i:

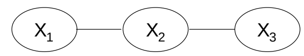
$$p_{new}(\mathbf{x}) \leftarrow p_{old}(\mathbf{x}) \frac{p(\mathbf{x}_{c_i})}{p_{old}(\mathbf{x}_{c_i})}$$

Undirected Graphs – Alternate Form

e.g., Levy 1948, Besag 1974

Each variable is independent of all other variables given its neighbors in the graph.

Use "local distributions" $p(x_i|\text{neighbors}_i)$

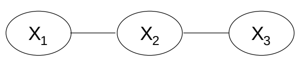


$p(x_1|x_2)$ $p(x_2|x_1,x_3)$ $p(x_3|x_2)$

Undirected Graphs – Alternate Form

Each variable is independent of all other variables given its neighbors in the graph.

Use "local distributions" $p(x_i|\text{neighbors}_i)$



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

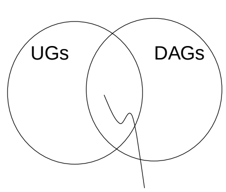
Generate p(x) via Gibbs sampling
(Heckerman, Chickering, Meek, Rounthwaite, and Kadie 2000)

$$p(x_i \mid \text{neighbors}_i) = p(x_i \mid x \setminus x_i)$$

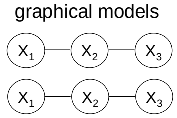
Summary

Model	Local distrbns	Joint recovery
DAG	$p(x_i\|pa_i)$	multiplication
UG_{IPS}	$p(x_{ci})$	IPS
UG_{MC}	$p(x_i\|n_i)$	Gibbs sampling

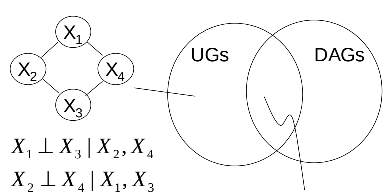
DAGs <> UGs



decomposable graphical models

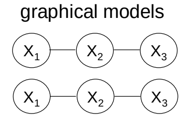


DAGs <> UGs

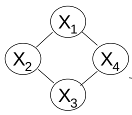


$$X_1 \perp X_3 \mid X_2, X_4$$
$$X_2 \perp X_4 \mid X_1, X_3$$

decomposable graphical models

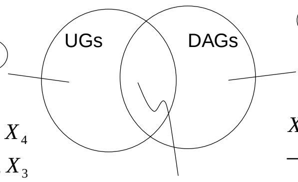


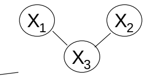
DAGs <> UGs



$X_1 \perp X_3 \mid X_2, X_4$

$X_2 \perp X_4 \mid X_1, X_3$



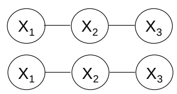


"v-structure"

$X_1 \perp X_2 \mid \phi$

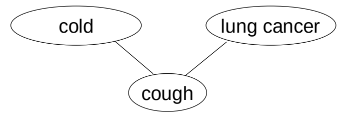
$\neg(X_1 \perp X_2 \mid X_3)$

decomposable
graphical models



Explaining Away

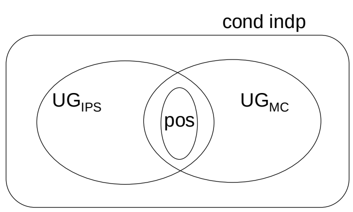
Pearl 1988



$\text{cold} \perp \text{lung cancer}$

$\neg(\text{cold} \perp \text{lung cancer} \mid \text{cough})$

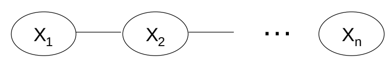
UGs and non-positive distributions



UG_{IPS}: Those distributions encoded by $p(\mathbf{x}) = \prod_i f(\mathbf{x}_{c_i})$

UG_{MC}: Those distributions encoded by irreducible MC

"Inference" in graphical models



$$p(x_n) = \sum_{x_1,\ldots,x_{n-1}} p(x_1)\, p(x_2 \mid x_1) \cdots p(x_n \mid x_{n-1})$$
$$= \sum_{x_1} p(x_1) \sum_{x_2} p(x_2 \mid x_1) \cdots \sum_{x_{n-1}} p(x_{n-1} \mid x_{n-2})\, p(x_n \mid x_{n-1})$$

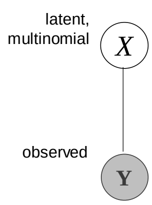
Inference in graphical models

- **Exact methods that exploit UG/DAG structure**
  e.g., Laurtizen and Spiegelhalter 1988
  - **Convert to triangulated (decomposable) UG**
  - **Create tree of cliques (running int property)**
  - **Perform tree version of dynamic programming**
- **Approximation methods needed when largest cliques contains too many variables**
  - **MCMC (e.g., Geman and Geman 1984)**
  - **Variational methods (e.g., Jordan et al. 1999)**
  - **Loopy propagation (e.g., Murphy et al. 1999)**

Graphical models are a common representation for many models

- **Finite mixture models**
- **Factor analysis**
- **Hidden Markov model**
- **Kalman filter**
- **Hierarchical models**

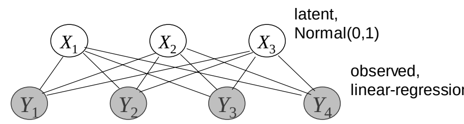
Finite mixture model



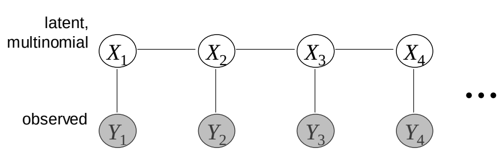
$$p(\mathbf{y}) = \sum_k p(x=k)\; p(\mathbf{y} \mid x=k)$$

mixing weight of kth component model

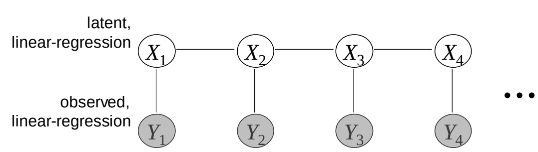
Factor Analysis



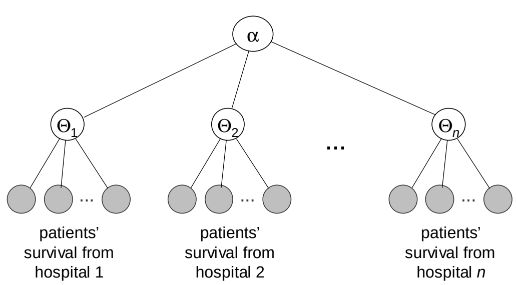
Hidden Markov model



Kalman filter



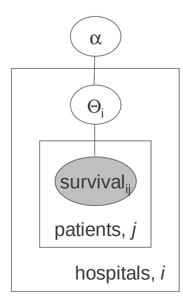
Hierarchical models



Plates (BUGS)

Spiegelhalter, Thomas, & Best 1999

A compact representation for repetitive structure

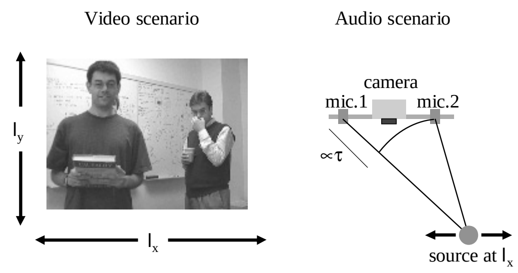


Advantages of common representation

- Transfer ideas between research communities
- Design new models

Example: Audio-video fusion

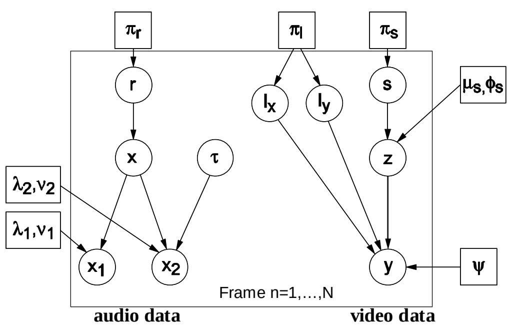
Beal, Attias, & Jojic 2002



Goal: detect and track speaker

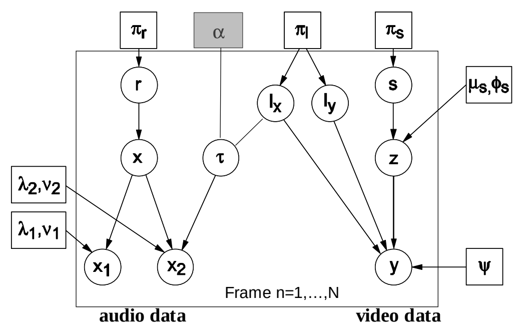
Slide courtesy Beal, Attias and Jojic

Separate audio-video models



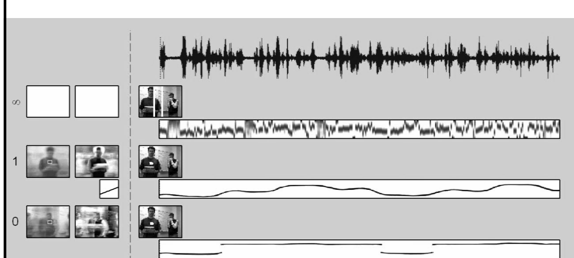
Slide courtesy Beal, Attias and Jojic

Combined model



Slide courtesy Beal, Attias and Jojic

Tracking Demo



Slide courtesy Beal, Attias and Jojic

Applications of graphical models

DAGs and UGs:
- Density estimation
- Classification and regression
- Clustering (finite mixture models)

UGs:
- Acausal models
- Spatial processes

DAGs:
- Acausal and causal models
- Expert systems

DAGs or “Bayesian Networks” and expert systems

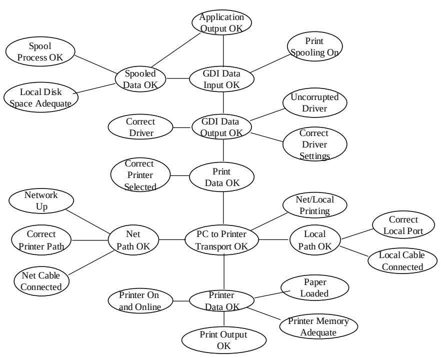
Early competitors to representing uncertainty in expert systems (late 70s, early 80s)
- **MYCIN certainty-factor model (rule-based systems)**
- **Dempster-Shafer theory**
- **Fuzzy set theory**
- **Bayesian probability**

Bayesian probability dominant by 1987 (in large part due to Bayesian Networks)

Examples of expert systems

- **MUNIN: Neuromuscular diagnosis**
  Andreassen, Woldbye, Falck, and Andersen 1987
- **Pathfinder: Lymph-node pathology diagnosis**
  Heckerman, Horvitz, & Nathwani 1989
- **QMR-DT: Internal medicine diagnosis**
  Shwe et al. 1991
- **Microsoft Windows Troubleshooters**
  Heckerman et al. 1995

Print Troubleshooter



So simple, a child could do it…

Teenager Designs Award-Winning Science Project

.. For her science project, which she called "Dr. Sigmund Microchip," Tovar wanted to create a computer program to diagnose the probability of certain personality types. With only answers from a few questions, the program was able to accurately diagnose the correct personality type 90 percent of the time.

Elena Tovar stands proudly in front of "Dr. Sigmund Microchip," the science project she created using the advanced mathematical formulas that Microsoft Research uses to build artificial intelligence programs.

Software

http://www.cs.berkeley.edu/~murphyk/Bayes/bnsoft.html

Learning graphical models from data

Uncertainty in parameters: $p(\theta \mid \mathbf{m})$ (assumed smooth)

Uncertainty in model: $p(\mathbf{m})$

Given finite sample of inf exchangeable data $\mathbf{d} = (\mathbf{x}_1,\ldots,\mathbf{x}_N)$:

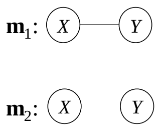
$$p(q \mid \mathbf{d}) = \sum_{\mathbf{m}} p(\mathbf{m} \mid \mathbf{d}) \int p(q \mid \theta, \mathbf{m})\; p(\theta \mid \mathbf{d}, \mathbf{m})\; d\theta$$

$$p(\mathbf{m} \mid \mathbf{d}) \propto p(\mathbf{m}) \underbrace{\int p(\mathbf{d} \mid \theta, \mathbf{m})\; p(\theta \mid \mathbf{m})\; d\theta}_{\text{marginal likelihood}}$$

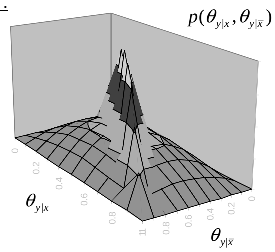
Marginal parameter prior is not smooth

$$p(\theta) = p(\theta \mid \mathbf{m}_1)\, p(\mathbf{m}_1) + p(\theta \mid \mathbf{m}_2)\, p(\mathbf{m}_2)$$

Example for binary X, Y:



$\theta_y = \theta_{y|x} = \theta_{y|\bar{x}}$



Methods and approximations

Only parameters uncertain:

- Bayesian – MCMC (e.g., BUGS)
- MAP/ML – EM (e.g., NIPS community)

Both parameters and structure uncertain:

- Bayesian – RJMCMC (Green 1995); MC[3] (Madigan and York 1995)
- Bayesian model selection for complete data (e.g., Cooper and Herskovits 1992; Spiegelhalter et al. 1993; Buntine 1994; Heckerman et al. 1995)
- Approx Bayesian model selection for incomplete data (e.g. Friedman 1997; Attias 2000)

- Constraint-based methods (Spirtes et al. 2001, Pearl 2000)

Computationally attractive paremeter priors for DAG models

Geiger and Heckerman 1997, 2002

Challenge: The number of DAG models for n variables grows super exponentially with n
- Want priors for all DAG models for X to come from a small number of assessments
- Want closed form for marginal likelihood

Solution: A set of assumptions

Extension of Dawid and Lautitzen's (1993) priors for decomposable models

Assumptions

For eligible local distribution families:
- Parameter independence
- (Conjugate priors)
- Complete data
- Equivalent graphs have equivalent priors
- Parameter modularity

Eligible local distribution families

- $X=(X_1,\ldots,X_n)$ discrete (finite): $p(x_i|pa_i,\theta_i)$ is "full table"

$$p(x_i \mid \mathbf{pa}_i = j, \theta_i, \mathbf{m}) \text{ is } mult\,(\theta_{x_i^1|pa_i^j}, \ldots, \theta_{x_i^{r_i}|pa_i^j})$$

$$\theta_{ij} = (\theta_{x_i^1|pa_i^j}, \ldots, \theta_{x_i^{r_i}|pa_i^j})$$

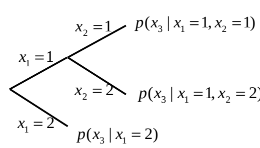
- X continuous: $p(x_i|pa_i,\theta_i)$ is linear regression

$$p(x_i \mid \mathbf{pa}_i, \theta_i, \mathbf{m}) = m_i + \sum_{x_j \in \mathbf{pa}_i} b_{ji} x_j + N(0, \sigma_i^2)$$

$\theta_i = (m_i, \mathbf{b}_i, \sigma_i^2)$ Note: $p(x|\theta)$ is m.v. Gaussian

Other eligible distribution families

- $X=(X_1,\ldots,X_n)$ discrete: $p(x_i|pa_i,\theta_i)$ a (probabilistic) decision tree



- X_i continuous: $p(x_i|pa_i,\theta_i)$ is a linear regression for each configuration of the discrete parents of X_i;
  X_i discrete: $p(x_i|pa_i,\theta_i)$ is full table (cont parents not allowed);
  $p(x|\theta)$ is conditional Gaussian (Lauritzen 1992)

First Assumption: Parameter independence

Speigelhalter and Lauritzen 1990

X discrete:

$$p(\theta) = \prod_i \prod_j p(\theta_{ij})$$

X continuous:

$$p(\theta) = \prod_i p(\theta_i)$$

Second assumption: Conjugate priors

When $p(x_i|pa_i,\theta_i)$ is a full table:

$$p(\theta_{ij}) = \mathrm{Dir}(\alpha_{ij1}, \ldots, \alpha_{ijr_i})$$

When $p(x_i|pa_i,\theta_i)$ is a linear regression

$$p(\theta_i) = \text{Normal - gamma}$$

Third assumption: Complete data

Yields fast-to-compute, closed-form formula

E.g., when $p(x_i|pa_i,\theta_i)$ is a full table (Cooper and Herskovits 1992):

$$p(\mathbf{d}\,|\,\mathbf{m}) = \prod_{i=1}^{n} \prod_{j=1} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}$$

$$N_{ijk} : \#\text{ cases where } X_i = x_i^k \text{ and } \mathbf{Pa}_i = \mathbf{pa}_i^j$$

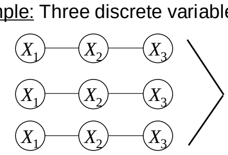
$$\alpha_{ij} = \sum_{k=1} \alpha_{ijk} \quad N_{ij} = \sum_{k=1} N_{ijk}$$

Problem with equivalent models

Two DAGs for X are equivalent if they encode the same sets of distributions for X.

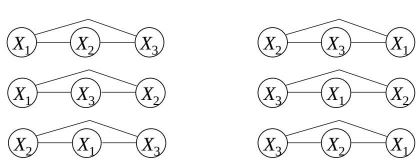
If each $p(x_i|pa_i,\theta_i)$ is full table or linear regression, then two DAGs for X are equivalent iff they encode the same independencies.

Example: Three discrete variables; full tables



$$X_1 \perp X_3 \mid X_2$$

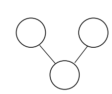
Complete network structures encode no independence

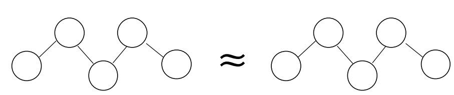


General test for equivalence

Verma & Pearl 1990: Two DAGs for X encode the same independencies iff

- They have the same skeleton
- They have the same v-structures





Problem: Equivalent graphs have different priors (for almost all hyperparameter values)

Example: X and Y binary; full tables

$\mathbf{m}_1$: X — Y $\qquad$ $\mathbf{m}_2$: X — Y

$\theta_x \qquad \theta_{y|x}, \theta_{y|\bar{x}} \qquad\qquad \theta_y = \theta_x \theta_{y|x} + (1-\theta_x)\theta_{y|\bar{x}}$

etc.

If each multinomial (bernoulli) has a Dir(1,1) prior, then

$$p(\theta_{m2} \mid \mathbf{m}_2) \neq \left| \frac{\partial \theta_{m1}}{\partial \theta_{m2}} \right| p(\theta_{m1} \mid \mathbf{m}_1)$$

Fourth assumption

Equivalent (complete) graphs have equivalent priors, and hence equal marginal likelihoods

$\mathbf{m}_1$: X — Y $\qquad$ $\mathbf{m}_2$: X — Y

$$p(\theta_{m2} \mid \mathbf{m}_2) = \left| \frac{\partial \theta_{m1}}{\partial \theta_{m2}} \right| p(\theta_{m1} \mid \mathbf{m}_1)$$

Independence + Equivalence => Conjugacy

Geiger and Heckerman 1997, 2002

- **Parameter independence**
- **Equivalent complete graphs have equivalent priors**
- **Technical conditions**

$\Downarrow$

parameters have conjugate distributions

Example: Two binary variables

X — Y $\qquad$ X — Y

$\Downarrow$ parameter independence
equivalence property

$$\frac{f(\theta_x)\, g(\theta_{y|x})\, h(\theta_{y|\bar{x}})}{\theta_x(1-\theta_x)} = \frac{i(\theta_y)\, j(\theta_{x|y})\, k(\theta_{x|\bar{y}})}{\theta_y(1-\theta_y)}$$

$\Downarrow$ positivity

$$p(\theta_{\mathbf{x}}) = p(\theta_{xy}, \theta_{x\bar{y}}, \theta_{\bar{x}y}, \theta_{\bar{x}\bar{y}}) \propto \theta_{xy}^a \theta_{x\bar{y}}^b \theta_{\bar{x}y}^c \theta_{\bar{x}\bar{y}}^d$$

General discrete case

- **Parameter independence**
- **Equivalent complete graphs have equivalent priors**
- **p(θ) strictly positive**

$\Downarrow$

$$\mathbf{X} \sim mult; \quad p(\theta_{\mathbf{x}} \mid \mathbf{m}_{\text{complete}}) = \mathrm{Dir}(\alpha_1, \ldots, \alpha_{|\mathbf{x}|})$$

"hyper Dirichlet"

Characterization of the Dirichlet

- **Parameter independence**
- **Equivalent complete graphs have equivalent priors**
- **p(θ) strictly positive**

$\Updownarrow$

$$\mathbf{X} \sim mult; \quad p(\theta_{\mathbf{x}} \mid \mathbf{m}_{\text{complete}}) = \mathrm{Dir}(\alpha_1, \ldots, \alpha_{|\mathbf{x}|})$$

"hyper Dirichlet"

Characterization of Normal-Wishart

- **Parameter independence**
- **Equivalent complete graphs have equivalent priors**
- **n>2; no element of Σ^{-1} is zero**

$$\Updownarrow$$

$$\mathbf{X} \sim m.v.Gaussian; \quad p(\theta_{\mathbf{x}} \mid \mathbf{m}_{\text{complete}}) = \text{NW}(\boldsymbol{\mu}, \Sigma^{-1})$$

"hyper Normal-Wishart"

Hyperparameters are highly constrained

E.g., discrete case:

$$p(\theta_{\mathbf{x}} \mid \mathbf{m}_{\text{complete}}) = \text{Dir}(\alpha_1, \ldots, \alpha_{|\mathbf{x}|}), \quad \alpha_i = \alpha \cdot p(\mathbf{x} = i)$$

$$\Downarrow$$

$$p(\theta_{x_i^1 | pa_i^j}, \ldots, \theta_{x_i^{r_i} | pa_i^j} \mid \mathbf{m}_{\textbf{complete}}) = \text{Dir}(\alpha_{ij1}, \ldots, \alpha_{ijr_i})$$

$$\alpha_{ijk} = \alpha \cdot p(x_i = k \mid pa_i = j)$$

So far…

$\Downarrow$ parameter independence
equivalence property

hyper distribution

$\Downarrow$

(conjugate) p(θ|**m**) for all complete **m**

Fifth assumption: Parameter modularity

Heckerman, Geiger, and Chickering 1995

If a variable X_i in two DAG models have the same parents, then

$$p(\theta_i \mid \mathbf{m}_1) = p(\theta_i \mid \mathbf{m}_2)$$

Parameter Modularity: Example

$\mathbf{m}_1$: X_1 — X_2

$\mathbf{m}_2$: X_1 X_2

$$p(\theta_1 \mid \mathbf{m}_1) = p(\theta_1 \mid \mathbf{m}_2)$$

The whole story

$\Downarrow$ parameter independence
equivalence property

hyper distribution

$\Downarrow$

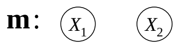
(conjugate) p(θ|**m**) for all complete **m**

$\Downarrow$ parameter independence
parameter modularity

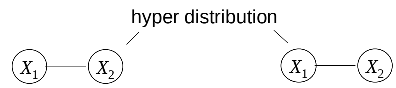
(conjugate) p(θ|**m**) for all **m**

Example: Empty graph for two variables

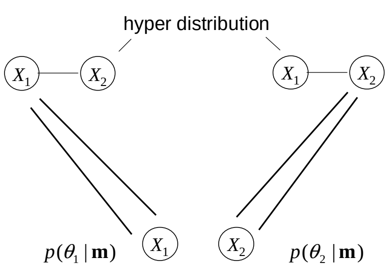
Given a hyper distribution for (X_1,X_2), compute parameter prior for



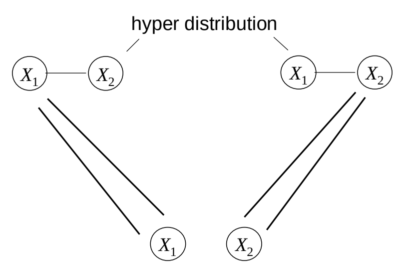
Step 1: Change of variable



Step 2: Parameter modularity



Step 3: Parameter independence



$$p(\theta_1, \theta_2 \mid \mathbf{m}) = p(\theta_1 \mid \mathbf{m}) \, p(\theta_2 \mid \mathbf{m})$$

College plans of high-school seniors

Sewall and Shah 1968

SEX **Sex: male, female**

SES **Socioeconomic status: low, low mid, high mid, high**

IQ **IQ: low, low mid, high mid, high**

PE **Parental encouragement: low, high**

CP **College plans: yes, no**

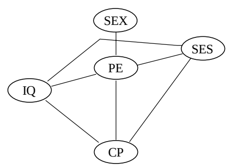
Data: ~10,000 students

College plans of high-school seniors

Analysis:

- **Consider DAG models with no latent variables**
- **$p(\theta|m_{complete})$ is hyper Dirichlet with uniform mean and sample size 5**
- **p(m): Uniform**

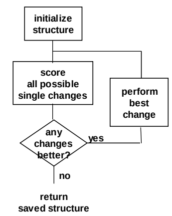
Results



$p(\mathbf{m} \mid \mathbf{d}) = 1.000000$

DAG search for large domains

- **Finding the DAG model with the highest marginal likelihood among those structures with at most *k* parents is NP hard for *k*>1.**
  Chickering 1996
- **Monte-Carlo methods**
- **Greedy/local search**
  Heckerman et al. 1995



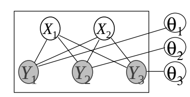
DAG model selection given incomplete data

- **Large sample approximations**
  - **BIC (Friedman 1997)**
  - **Laplace (Thiesson, Meek, Chickering, & Heckerman 1998)**
    Caveat: DAG models (discrete) with latent variables are Stratified Exponential Family (Geiger, Heckerman, King, & Meek 2001)
- **MCMC methods (e.g., DiCiccio et al. 1995)**
- **Variational methods (e.g., Attias 2000)**

Variational methods for model selection

e.g. Attias 2000, Ghahramani & Beal 2000

Example: Factor analysis



$$\ln p(\mathbf{y} \mid \mathbf{m}) = \ln \int dx\, d\theta\; p(\mathbf{y}, \mathbf{x}, \theta \mid \mathbf{m})$$

$$= \ln \int dx\, d\theta\; q(\mathbf{x}, \theta)\, \frac{p(\mathbf{y}, \mathbf{x}, \theta)}{q(\mathbf{x}, \theta)}$$

$$\geq \int dx\, d\theta\; q(\mathbf{x}, \theta)\, \ln \frac{p(\mathbf{y}, \mathbf{x}, \theta)}{q(\mathbf{x}, \theta)} \qquad \text{(Jensen ineq.)}$$

Using a simple, factorized $q(x,\theta)=q_x(x)q_\theta(\theta)$:

$$\ln p(\mathbf{y} \mid \mathbf{m}) \geq \int dx\, d\theta\; q_x(\mathbf{x})\, q_\theta(\theta) \ln \frac{p(\mathbf{y}, \mathbf{x}, \theta)}{q_x(\mathbf{x})\, q_\theta(\theta)}$$

Variational versus Laplace methods

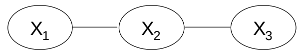
- **Laplace: Approximates p(θ|d) around one mode; full dependence**
- **Variational: p(θ|d) can be any convenient (possibly multi-model) distribution; convenience usually demands independence assumptions**

Averaging/selecting among UG models

- **Decomposable: special case of DAGs;**
  e.g., Dawid and Lauritzen 1993
- **Non-decomposable**
  - **No closed-form marginal likelihood; MCMC used (e.g. Dellaportas & Foster 1999)**
  - **BIC via IPS + EM (e.g., Lauritzen 1996)**
  - **Heuristic method**

Heuristic method

Heckerman et al. 2000



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

Each local distribution

$$p(x_i \mid \text{neighbors}_i) = p(x_i \mid x \setminus x_i)$$

learned separately

Each local distribution can be learned efficiently (e.g., decision tree learned with Bayesian model selection)

Resulting conditionals are inconsistent, although "almost consistent"

Microsoft applications of "Dependency Networks"

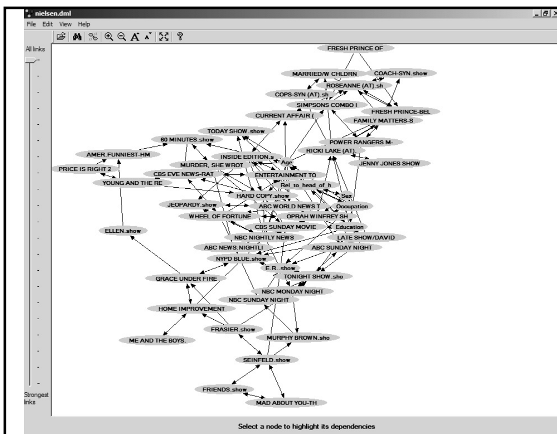
- **Collaborative filtering (Commerce Server 2002)**
- **Exploratory data analysis (SQL Server 2000)**

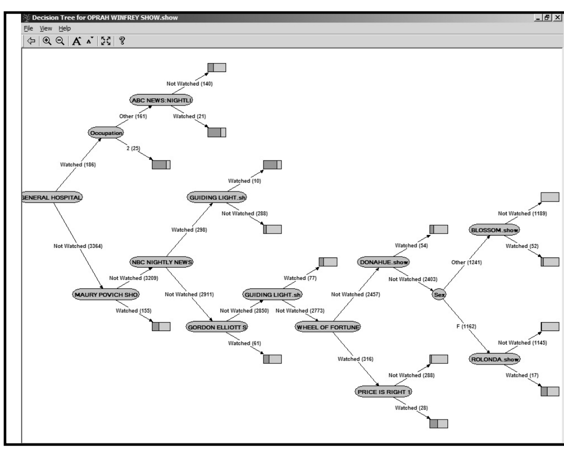
Exploratory data analysis

Example: Nielsen data, 2/6/95-2/19/95

	Age	Show1	Show2	Show3	
viewer 1	73	y	n	n	
viewer 2	16	n	y	y	...
viewer 3	35	n	n	n	
	etc.				

~400 shows, ~3000 viewers



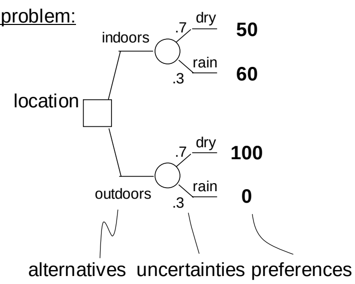


Software

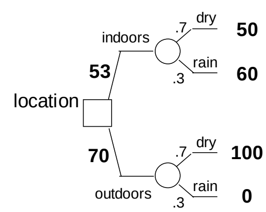
- **Dependency networks, DAGs:** **http://research.microsoft.com/~dmax/WinMine/Tooldoc.htm**
- **Many others:** **http://www.cs.berkeley.edu/~murphyk/Bayes/bnsoft.html**

Decision making
vonNeumann & Morgenstern, 1948

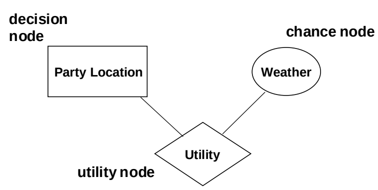
The party problem:



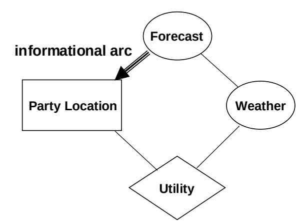
MEU principle



Graphical model for decision making: Influence Diagrams
Howard & Matheson 1981



Influence diagrams

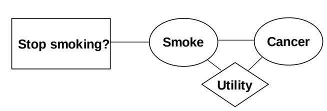


Solving influence diagrams

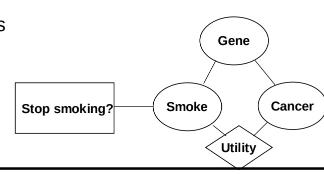
- **Concert to decision tree and then solve (Howard & Matheson 1981); computation grows exponentially with number of nodes**
- **Solve using influence diagram as data structure (Shachter 1986)**

Causal reasoning, decision making, and influence diagrams

Causal reasoning is about prediction in the face of interventions (a.k.a. alternatives)



versus



Page 15

Causal reasoning and influence diagrams

Alternative formulations of causal reasoning

- Rubin (e.g., 1978)
- Pearl (e.g. 2000)
- Spirtes, Glymous, & Scheines (e.g. 2001)

involve couterfactuals

Dawid 2000: Don't need couterfactuals

Heckerman & Shacther 1995: If you want couterfactuals, they are consistent with decision theory and can be encoded with influence diagrams

For more information…

Tutorials:

W. Buntine. Operations for learning with graphical models. Journal of Artificial Intelligence Research, 2, 159-225 (1994).

D. Heckerman (1999). A tutorial on learning with Bayesian networks. In Learning in Graphical Models (Ed. M. Jordan). MIT Press.

Books:

R. Cowell, A. P. Dawid, S. Lauritzen, and D. Spiegelhalter. Probabilistic Networks and Expert Systems. Springer-Verlag. 1999.

F. Jensen (2001). Bayesian Networks and Decision Diagrams. Springer-Verlog, New York.

M. I. Jordan (ed, 1988). Learning in Graphical Models. MIT Press.

S. Lauritzen (1996). Graphical Models. Claredon Press.

J. Pearl (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann.

J. Pearl (2000). Causality: Models, Reasoning, and Inference. Cambridge University Press.

P. Spirtes, C. Glymour, and R. Scheines (2001). Causation, Prediction, and Search, Second Edition. MIT Press.

Software:

http://www.cs.berkeley.edu/~murphyk/Bayes/bayes.html

Bibliography

S. Andreassen and M. Woldbye and B. Falck and S. Andersen. MUNIN: A causal probabilistic network for interpretation of electromyographic findings. In Proceedings of the Tenth International Joint Conference on Artificial Intelligence, Milan, Italy, 366-372. Morgan Kaufmann, San Mateo, CA, August 1987.

H. Attias (2000). A Variational Bayesian framework for graphical models. Advances in Neural Information Processing Systems 12 (Eds. S. Solla et al.), 209-215. MIT Press, Cambridge, MA.

M. Beal, H. Attias, N. Jojic (2002). Audio-video sensor fusion with probabilistic graphical models . Proc. ECCV 2002.

J. Besag. Spatial interaction and the statistical analysis of lattice systems, Journal of the Royal Statistical Society, B, 36, 192-236 (1974).

W. Buntine. Operations for learning with graphical models. Journal of Artificial Intelligence Research, 2, 159-225 (1994).

D. Chickering (1996). Learning Bayesian networks is NP-complete. In Learning from Data (Eds. D. Fisher and H. Lenz), 121-130. Springer-Verlag, New York.

G. Cooper and E. Herskovitz. A Bayesian method for the induction of probabilistic networks from data. Machine Learning, 9, 309-347 (1992).

J. Darroch, S. Lauritzen, and T. Speed (1980). Markov fields and log-linear models for contingency tables. Annals of Statistics, 8, 522-539.

J. Darroch, S. Lauritzen, and T. Speed (1980). Markov fields and log-linear models for contingency tables. Annals of Statistics, 8, 522-539.

A.P. Dawid and S. Lauritzen. Hyper Markov laws in the statistical analysis of decomposable graphical models. Annals of Statistics, 21, 1272-1317 (1993).

A.P. Dawid. Causal inference without counterfactuals. With discussion. Journal of the American Statistical Association, 95 407-448 (2000).

J. Darroch, S. Lauritzen, and T. Speed (1980). Markov fields and log-linear models for contingency tables. Annals of Statistics, 8, 522-539.

A.P. Dawid and S. Lauritzen. Hyper Markov laws in the statistical analysis of decomposable graphical models. Annals of Statistics, 21, 1272-1317 (1993).

A.P. Dawid. Causal inference without counterfactuals. With discussion. Journal of the American Statistical Association, 95 407-448 (2000).

P. Dellaportas and J. Forester. Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. Biometrika, 86, 615-633 (1999).

W. Deming and F. Stephan. On a least square adjustment of a sampled frequency table when the expected marginal totals are known. Annals of Mathematical Statistics, 11, 427-44 (1940).

T. DiCiccio, R. Kass, A. Raftery, and L. Wasserman. Computing Bayes factors by combining simulation and asymptotic approximations. Technical Report 630, Department of Statistics, Carnegie Mellon University, PA, July, 1995.

D. Geiger and D. Heckerman. A characterization of the Dirichlet distribution through global and local parameter independence. The Annals of Statistics, 25, 1344-1369 (1997).

D. Geiger and D. Heckerman. Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. Annals of Statistics, to appear.

D. Geiger, D. Heckerman, H. King, and C. Meek. Stratified exponential families: Graphical models and model selection. The Annals of Statistics, 29, 505-529 (2001).

Z. Ghahramani and M. Beal (2000). Variational Inference for Bayesian Mixtures of Factor Analysers. In Advances in Neural Information Processing Systems 12 (Eds. S. A. Solla, T.K. Leen, K. Miller), 449-455. MIT Press, Cambridge, MA.

I.J. Good. A causal calculus I, British Journal of Philosophy of Science, 11, 305-318 (1961).

P. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, Biometrika, 82, 711-732 (1995).

N. Friedman (1997). Learning belief networks in the presence of missing values and hidden variables. In Proceedings of the Fourteenth International Conference on Machine Learning. Morgan Kaufmann, San Mateo, CA.

J. Hammersley and P. Clifford (1971). Markov fields on finite graphs and lattices. Unpublished manuscript.

D. Heckerman, E. Horvitz, and B. Nathwani. Toward normative expert systems: Part I. The Pathfinder project. Methods of Information in Medicine, 31, 90-105 (1992).

D. Heckerman, J. Breese, and K. Rommelse. Decision-theoretic troubleshooting. Communications of the ACM, 38, 49-57 (1995).

D. Heckerman, D. Geiger, and D. Chickering. Learning Bayesian networks: The combination. Machine Learning, 20, 197-243 (1995).

D. Heckerman and R. Shachter. Decision-theoretic foundations for causal reasoning. Journal of Artificial Intelligence Research, 3, 405-430 (1995).

D. Heckerman, D. Chickering, C. Meek, R. Rounthwaite, and C. Kadie. Dependency networks for inference, collaborative filtering, and data visualization, Journal of Machine Learning Research, 1, 49-75 (2000).

R. Howard and J. Matheson (1981). Influence diagrams. In Readings on the Principles and Applications of Decision Analysis (Eds. Howard and Matheson), 721-762. Strategic Decisions Group, Menlo Park, CA.

M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul (1999). An introduction to variational methods for graphical models. In Learning in graphical models (Ed. M. Jordan. MIT Press, Cambridge, MA.

S. Lauritzen and D. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems, J. Royal Statistical Society B, 50, 157-224 (1988).

S. Lauritzen. Propagation of probabilities, means, and variances in mixed graphical association models. Journal of the American Statistical Association, 87, 1098-1108 (1992).

S. Lauritzen (1996). Graphical Models. Claredon Press, Oxford.

P. Levy. Chaines doubles de Markoff et fonctions aleatories de deux variables, Academcy of Science, Paris, 226, 53-55 (1948).

D. Madigan and J. York. Bayesian graphical models for discrete data. International Statistical Review, 63, 215-232 (1995).

K. Murphy, Y. Weiss, and M. Jordan (1999). Loopy belief propagation for approximate inference: an empirical study. In Proceedings of Fifteenth Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden.

J. Pearl (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Mateo, CA.

J. Pearl (2000). Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge, MA.

D. Rubin. Bayesian inference for causal effects: The role of randomization. Annals of Statistics, 6, 34-58 (1978).

W. Sewell and V. Shah. Social class, parental encouragement, and educational aspirations. American Journal of Sociology, 73, 559-572 (1968).

R. Shachter. Evaluating influence diagrams. Operations Research, 34, 871-882 (1986).

M. Shwe, B. Middleton, D. Heckerman, M. Henrion, E. Horvitz, H. Lehmann, and G. Cooper. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base: Part I. The probabilistic model and inference algorithms. Methods in Information and Medicine, 20, 241-250 (1991).

D. Spiegelhalter and S. Lauritzen. Sequential updating of conditional probabilities on directed graphical structures. Networks, 20, 579-605 (1990).

D. Spiegelhalter, A.P. Dawid, S. Lauritzen, and R. Cowell. Bayesian analysis in expert systems. Statistical Science, 8, 219-282 (1993).

D. Spiegelhalter, A. Thomas, and N. Best (1999). WinBUGS Version 1.2 User Manual. Cambridge: MRC Biostatistics Unit.

P. Spirtes, C. Glymour, and R. Scheines (2001). Causation, Prediction, and Search, Second Edition. MIT Press, Cambridge, MA.

B. Thiesson, C. Meek, D. Chickering, and D. Heckerman (1999). Computationally efficient methods for selecting among mixtures of graphical models, with discussion. In Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting (Eds. J. Bernardo, J. Berger, A.P. Dawid, and A.F.M. Smith), 631-656. Clarendon Press, Oxford.

T. Verma and J. Pearl. Equivalence and synthesis of causal models. In Proceedings of Sixth Conference on Uncertainty in Artificial Intelligence, Boston, MA, 220-227. Morgan Kaufmann, San Mateo, CA, July, 1990.

J. von Neumann and O. Morgenstern (1947). Theory of Games and Economic Behavior. Princeton University Press, Princeton, NJ.

J. Whittaker (1990). Graphical Models in Applied Multivariate Statistics. John Wiley and Sons, New York.

S. Wright (1921). Correlation and causation, Journal of Agricultural Research, 20, 557-585.