GRAPHICAL MODEL STRUCTURE LEARNING

- WE WANT $P(G|D)$. G IS GRAPH STRUCTURE AS $V \times V$ ADJ MATRIX. PROBLEM EXPONENTIAL IN NUMBER OF NODES $O(2^{V(V-1)/2})$
  FULL POSTERIOR IS PROHIBITIVELY LARGE.

FOR KNOWLEDGE DISCOVERY

JUST GRAPH TOPOLOGY. POSTERIOR EDGE MARGINALS $P(G_{ST}=1|D)$. EDGE 'THICKNESS' IS CONFIDENCE VALUE

FOR DENSITY ESTIMATION

MAP GRAPH $\hat{G} \in \arg\max_G P(G|D)$. HEURISTICS BECAUSE EXPTIME. FOR TREES IS OK TO FIND GLOBAL OPTIMUM EXACTLY.

- CONSIDER WHETHER A LATENT VARIABLE MODEL WOULD BE APPROPRIATE BECAUSE IT IS A LOT EASIER

FOR KNOWLEDGE DISCOVERY

QUICK AND DIRTY. NO JOINT PROBS → NO PREDICTION. NO CAN GOODNESS OF FIT. GOOD FOR VISUALIZATION.

RELEVANCE NETWORKS: PLOTS PAIRWISE MUTUAL INFORMATION. WE DRAW EDGE IF $I(X_i, X_j) > $ THRESHOLD. IF GAUSSIAN, THIS IS Σ. COVARIANCE GRAPH.
→ ALSO PLOT EDGES WHERE CONDITIONAL INDEPENDENCE.

DEPENDENCY NETWORKS: FITS A SPARSE FULL CONDITIONALS $P(X_T|X_{-T})$. CHOSEN VARS ARE INPUTS PLOTTED FOR NODE. CPD FITTED USING ANY REG/CLASS METHOD.
→ CAN BE USED FOR INFERENCE VIA GIBBS. OK-ISH IF NOT MUCH MISSING DATA. USEFUL FOR DATA IMPUTATION / INITIALIZATION

LEARNING TREES

BECAUSE MUCH EASIER THAN ANY GRAPH. AND SUPPORT EXACT INFERENCE

- SAME NUM OF PARAMS FOR DIRECTED OR UNDIRECTED TREES. UNDIR + FOR STRUCTURE LEARNING. DIRECTED + FOR PARAMETER LEARNING

TREE LOG-LIKELIHOOD:

$$\frac{\log P(D|\theta,T)}{N} = \underbrace{\sum_{t\in V}\sum_{k} P_{EMP}(x_T=k)\log P_{EMP}(x_T=k)}_{\text{DOES NOT DEPEND ON TOPOLOGY} \to \text{IGNORED WHEN LEARNING STRUCTURE}} + \underbrace{\sum_{S,T} I(x_S, x_T|\hat{\theta}_{ST})}_{\text{MUTUAL INFO GIVEN EMPIRICAL DISTR.}}$$

MAX LL: MAXIMUM WEIGHT SPANNING TREE, EDGE WEIGHTS ARE MUTUAL INFO
→ CHOW-LIU ALGORITHM

- USE ALGO TO FIND MAX SPANNING TREE. $O(E \lg V)$
- NO OVERFIT → ALL TREES HAVE SAME NUM OF PARAMS

IF WE WANT FOREST BECAUSE BP INFERENCE IS MUCH FASTER. NO CAN USE MLE (IT DOESN'T OMIT EDGES). USE MARGINAL LIKELIHOOD OR OTHER PENALIZED MEASURE (BIC)

- $\log P(D|T) = \sum \text{SCORE}(N_{T,PA(T)})$. N ARE COUNTS.
- $\log P(D|T) = \sum w_{PA(T),T} + \sum_T \text{SCORE}(T|0)$

- MOST PROBABLE TREE IS MAXIMAL BRANCHING OF MATCHING WEIGHTED DIRECTED GRAPH
- MODIFIED MST

MIXTURES OF TREES EACH COMPONENT HAS DIFFERENT TOPOLOGY. EVEN INFINITE IS OK $O(V^3)$ INTEGRATING OVER ALL POSSIBLE TREES

→ FIT USING EM: E COMPUTES RESPONSIBILITIES FOR EACH CLUSTER FOR EACH DATAPOINT.
M USES WEIGHTED CHOW-LIU

LEARNING DAG

AKA BAYESIAN STRUCTURE LEARNING, NO HIDDEN VARS.

MARKOV EQUIVALENCY: IF THEY ENCODE SAME SET OF CI ASSUMPTIONS → PDAG (ESSENTIAL GRAPH, PATTERN) SOME EDGES ARE REVERSIBLE | UNDIRECTED, OTHERS ARE DIRECTED (COMPELLED)
SAME PATTERN, SAME V-STRUCTURES

- WHEN LEARNING DAG STRUCTURE → UP TO MARKOV EQUIVALENCY

EXACT INFERENCE

- LIKELIHOOD $P(D|G,\theta) = \prod_{t}^{V}\prod_{c}^{C}\prod_{k}^{K} \theta_{TCK}^{N_{TCK}}$ K STATE OF NODE, C STATE OF PARENTS
- ML WILL YIELD FULLY CONNECTED GRAPHS (B/C MAX PARAMS) → MAX MARGINAL LIKELIHOOD $P(D|G)$, WE NEED PRIORS ON PARAMS

ASSUME: GLOBAL PRIOR PARAM INDEPENDENCE $P(\theta) = \prod^{V} P(\theta_t)$
LOCAL " " " $P_t(\theta) = \prod_{c=1}^{C_t} P(\theta_{TC})$
PRIOR FOR EACH ROW MUST BE DIRICHLET $P(\theta_{TC}) = \text{DIR}(\theta_{TC}|\alpha_{TC})$

- $P(D|G) = \prod^{V} \text{SCORE}(N_{T,PA(T)})$
  $\text{SCORE}(N_{T,PA(T)}) = \prod_{c=1}^{C_t} \frac{B(N_{TC}+\alpha_{TC})}{B(\alpha_{TC})}$

$N_{TC} = \sum N_{TCK}$, $\alpha_{TC} = \sum \alpha_{TCK}$; $N_{T,PA(T)}$ IS VECTOR OF COUNTS (SUFF. STATS)

- MARGINAL LIKELIHOOD DECOMPOSES / FACTORIZES OVER GRAPH

HOW TO SET α

WHY NOT JEFFREYS? ⟶ VIOLATES LIKELIHOOD EQUIVALENCE (~~SAME~~ MARKOV EQUIVALENCE ⟶ SAME MARG. LIKELIHOOD), ONLY DIRICHLET WORKS

- BDE: $\alpha_{TCK} = \alpha \, p_0(x_T = k, x_{PA(T)} = c)$, $\alpha > 0$. p_0 IS PRIOR JOINT
- BDEU: N HAS SAME PARENTS IN $G_1, G_2 \longleftrightarrow p(\theta_t|G_1) = p(\theta_t|G_2)$

  p_0 ASSUMED UNIFORM $\alpha_{TCK} = \frac{\alpha}{K_T C_T}$

OBS! DAGS W/O HIDDENS MAY STILL HAVE LATENT HIDDENS IN DATA, CAREFUL WHEN INTERPRETING CAUSALLY

- K2 ALGORITHM WHEN NODES ARE TOTALLY ORDERED; ENUMERATE OVER ALL POSSIBLE ANCESTORS SUBSET AND COMPUTE MARGINAL LIKELIHOODS.
- GAUSSIAN CPD (VS TABULAR) ⟶ CONDITIONAL GAUSSIAN DAG CAN MIX GAUSSIAN + DISCRETE NODES. BIC APPROX OF MARGINAL LIKELIHOOD.

$$p(D|G) = \sum_T \log p(D_t|\hat{\theta}_t) - \frac{K_T C_T}{2} \log N$$

LARGE SCALE APPROXIMATIONS

- HYPOTHESIS SPACE IS HUGELY MOTHERFUCKING LARGE: $f(D) = \sum_{i}^{D} (-1)^{i+1} \binom{D}{i} 2^{i(D-i)} f(D-i)$ WAT DO?
- MAP ⟶ DYNAMIC PROGRAMMING WORKS UP TO 16 NODES. THEN GREEDY HILL-CLIMBING. CAN INIT WITH BEST RANDOM TREE (FOUND ANALYTICALLY). RANDOM RESTARTS, ETC.
- OTHER STUFF ⟶ IE KNOWLEDGE DISCOVERY. SAMPLE FROM POSTERIOR AND COMPUTE EMPIRICAL COUNTS. USE MH SAMPLING, COLLAPSED.

DAG WITH LATENT VARS

- MARGINAL LIKELIHOOD: $p(D|G) = \sum_H \int p(D,h|\theta,G)\,p(\theta|G)\,d\theta$ INTRACTABLE, USE DETERMINISTIC APPROXIMATIONS
  - BIC: $BIC(G) = \log p(D|\hat{\theta},G) - \frac{\log N}{2} \dim(G)$
  - CHEESEMAN-STUTZ: COMPUTE MAP PARAMS $\hat{\theta}$, $\bar{D} = D(\hat{\theta})$ (→ EM), $p(D|G) \approx p(\bar{D}|G)$. DOES NOT SUM FOR h. $\log p(D|G) \approx \underbrace{\log p(\bar{D}|G)}_{\text{PLUG-IN}} + \underbrace{\log p(D|\hat{\theta},G)}_{\text{INFERENCE}} - \underbrace{\log p(\bar{D}|\hat{\theta},G)}_{\text{PLUGIN LIKELIHOOD}}$
  - VARIATIONAL BAYES EM: $p(\theta, z_{1:N}|D) = q(\theta)\prod_i q(z_i)$
  - STRUCTURAL EM: FILL-IN DATA ONCE, USE IT TO EVALUATE NEIGHBOR SCORES. GOOD APPROX FOR DIFFERENCE IN MARGINAL LIKELIHOOD BETWEEN MODELS. → OK FOR PICKING NEIGHBORS

HOW TO DISCOVER HIDDEN VARS) LOOK FOR SIGNS IN THE DATA. IE CLUSTERS. INTRODUCE HIDDEN VARS. SEE HOW IT PERFORMS. HIDDENS MAY HAVE HIERARCHY → AGGLOMERATE. DATA MAY BE CONSTRAINED AT LEAVES OR NOT. WORKS EXCELLENTLY IF TRUE GRAPH IS A TREE.

STRUCTURAL EQUATION MODELS (SEM): IS DIRECTED MIXED GRAPH WHERE CPD ARE ALL LINEAR GAUSSIAN. PATH DIAGRAMS.

$x = Wx + \mu + \epsilon$, $p(x) = N(\mu, \Sigma)$, $\Sigma = (I-W)^{-1}\Psi(I-W)^{-T}$.

W LOWER TRIANGULAR ⟶ ACYCLIC. Ψ NOT DIAGONAL ⟶ BIDIRECTED EDGES, CORRELATIONS

RELATED TO FACTOR ANALYSIS, WHERE NOISE HAS FULL COVARIANCE MATRIX

CAUSAL DAG

STRONGER THAN ASSOCIATIVE CLAIMS. $A \to B$ IS NOW A DIRECTLY CAUSES B. CAUSAL MARKOV ASSUMPTION. NO CONFOUNDERS

- ASSIGNMENT $\neq$ OBSERVATION. DO NOTATION.
- GRAPH SURGERY: JOINT IS USUAL DGM, CUT ARCS COMING INTO NODES INTERVENED
- GRAPH SURGERY, THEN CAN COMPUTE $P(X_i \mid DO(X_j))$
- DIFFERENT CAUSAL ASSUMPTIONS $\to$ DIFFERENT CAUSAL CONCLUSIONS. (SIMPSON'S PARADOX)

LEARN FROM OBSERVATIONAL DATA: LEARN A PDAG. IF WE KNOW TRUE DAG $\to$ COMPUTE CAUSAL EFFECTS. IF WE DON'T KNOW TRUE DAG $\to$ COMPUTE FOR ALL DAGS IN CLASS OR LOCAL NEIGHBORHOOD
IDA

LEARN FROM INTERVENTION DATA: SKIP CASES WHERE INTERVENTION HAPPENED BEFORE LEARNING θ.
OR AUGMENT NORMAL DAG WITH NODES FOR INTERVENTIONAL FACTORS

LEARNING U-GGM

- EASIER THAN DAG $\to$ NO CYCLE PROBLEM.
- HARDER THAN DAG $\to$ LIKELIHOOD DOES NOT DECOMPOSE. NO GREEDY SEARCH/MCMC
- MLE $\ell(\Omega) = \log \det(\Omega) - \operatorname{tr}(S\Omega)$. $\Omega = \Sigma^{-1}$, S = EMPIRICAL COVARIANCE MATRIX $\to \nabla\ell(\Omega) = \Omega^{-1} - S$; COVARIANCE ESTIMATION
- 1:1 CORRESPONDANCE BETWEEN ZEROS IN Ω AND ABSENT EDGES IN GRAPH.
- LET'S USE A SPARSE INDUCING OBJECTIVE: $J(\Omega) = -\log \det(\Omega) + \operatorname{tr}(S\Omega) + \lambda \|\Omega\|_1$ GRAPHICAL LASSO!!! CONVEX BUT NONSMOOTH BECAUSE ℓ_1
- COMPUTING EXACTLY $P(G \mid D)$. POSTERIOR INFERENCE IN SPACE OF GRAPHS. CAN DO IF GRAPH IS DECOMPOSABLE. NOT DECOMPOSABLE $\to$ V. HARD. MODIFIED GRADIENT DESCENT + DIAGONAL LAPLACE PRIOR
- NON GAUSSIAN STILL CONTINUOUS DATA: USE D MONOTONIC TRANSFORMATIONS f_j SO THAT RESULTING DATA IS JOINTLY GAUSSIAN NONPARANORMAL DISTRIBUTION

DISCRETE UGM

HARDESTEST. STOCHASTIC LOCAL SEARCH IS NOT TRACTABLE. BUT STUFF HAS BEEN TRIED

GLASSO: FOR DISCRETE CRF/MRF. MODIFIED OBJECTIVE. V. COSTLY. APPROX FOR EVERYTHING.

LEARNING SPARSE MODELS MAY NOT BE ENOUGH BECAUSE IF TREEWIDTH IS LARGE, INFERENCE IS INTRACTABLE. $\to$ BOUND THE TREEWIDTH

- THIN JUNCTION TREES, OR LOW CIRCUIT COMPLEXITY