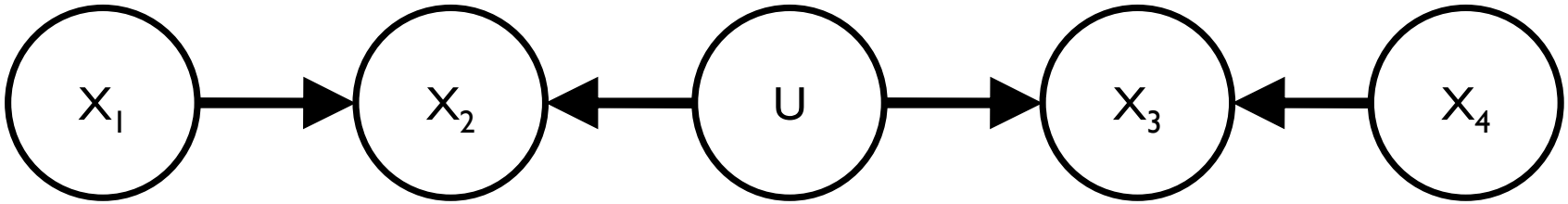
Mixed Cumulative Distribution Networks

Ricardo Silva, Charles Blundell and Yee Whye Teh
University College London

AISTATS 2011 – Fort Lauderdale, FL

Directed Graphical Models



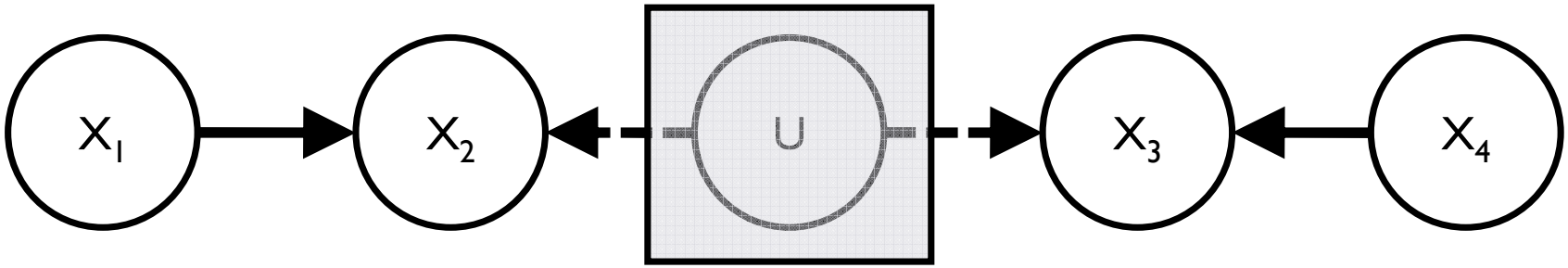
$$X_2 \perp\!\!\!\perp X_4$$

$$X_2 \not\perp\!\!\!\perp X_4 \mid X_3$$

$$X_2 \perp\!\!\!\perp X_4 \mid \{X_3, U\}$$

...

Marginalization



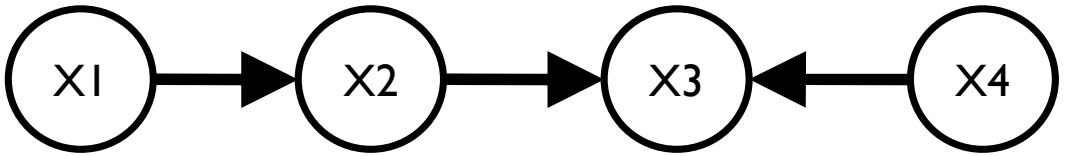
$$X_2 \perp\!\!\!\perp X_4$$

$$X_2 \not\perp\!\!\!\perp X_4 \mid X_3$$

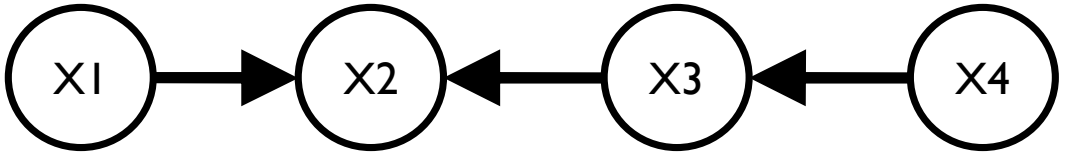
~~$X_2 \perp\!\!\!\perp X_4 \mid \{X_3, U\}$~~

...

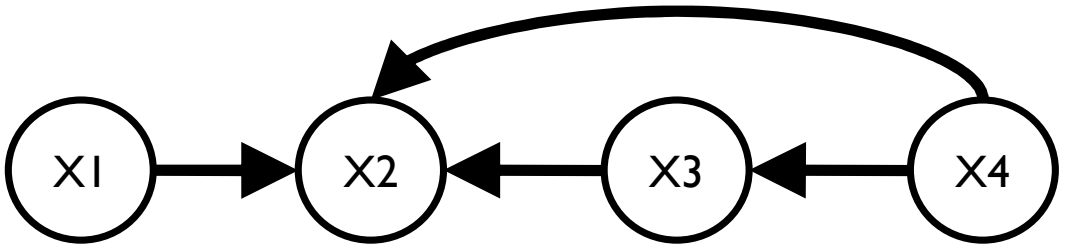
Marginalization



? No: $X_1 \perp\!\!\!\perp X_3 \mid X_2$



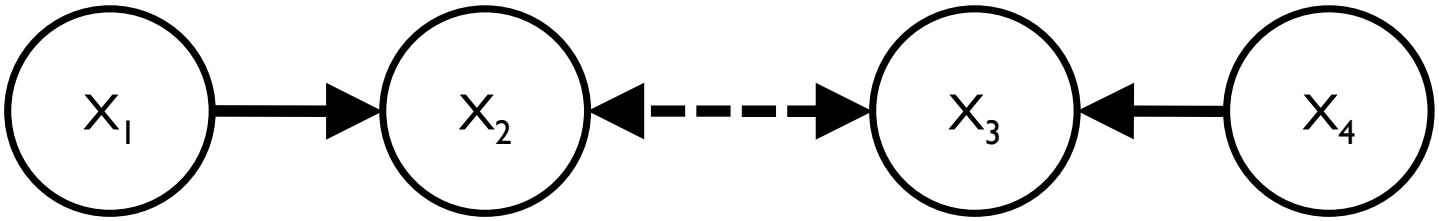
? No: $X_2 \perp\!\!\!\perp X_4 \mid X_3$



? OK, but not ideal

$X_2 \not\perp\!\!\!\perp X_4$

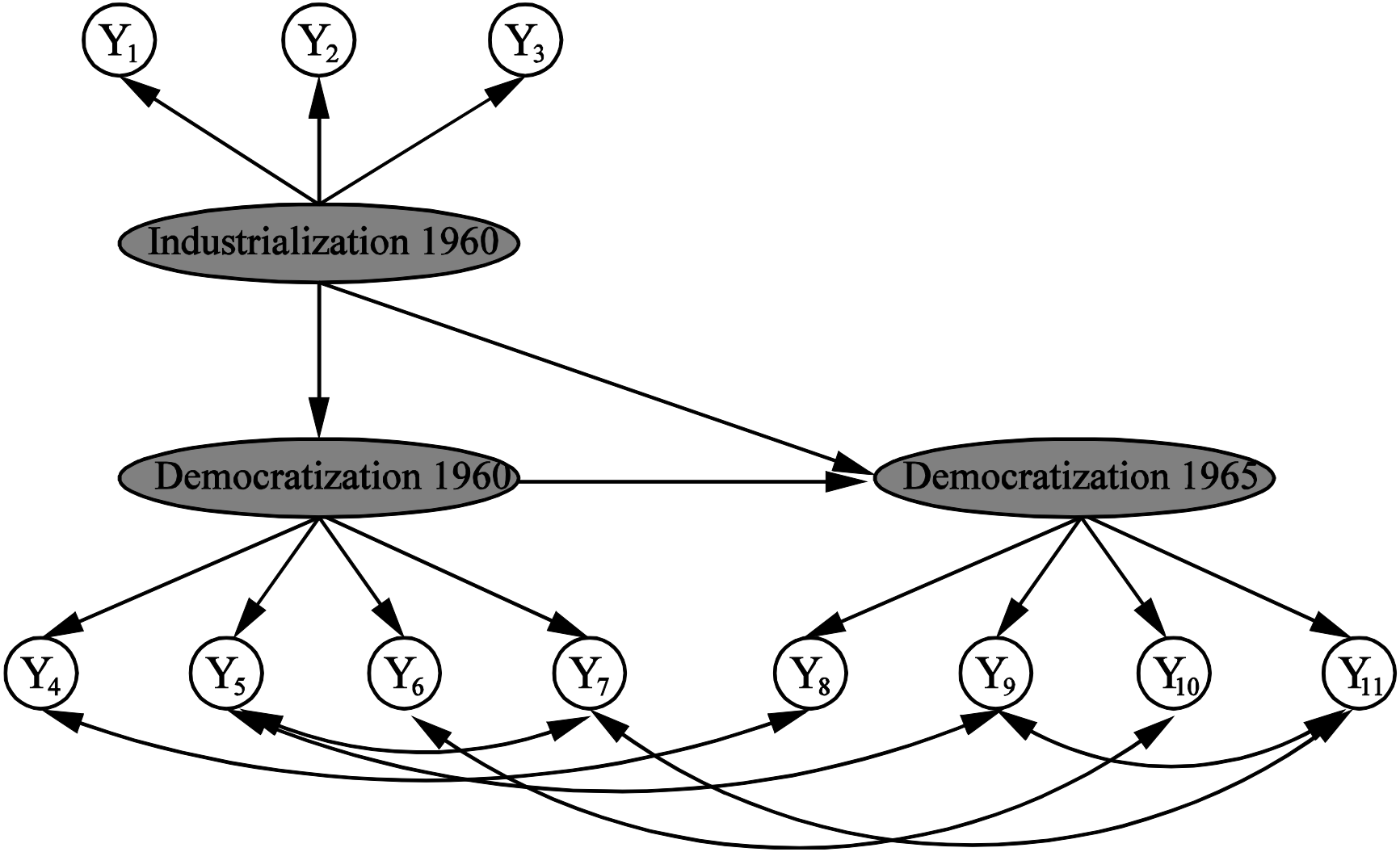
The Acyclic Directed Mixed Graph (ADMG)



- "Mixed" as in directed + bi-directed
- "Directed" for obvious reasons
  - See also: chain graphs
- "Acyclic" for the usual reasons
- Independence model is
  - Closed under marginalization (generalize DAGs)
  - Different from chain graphs/undirected graphs
  - Analogous inference calculus as DAGs: m-separation

(Richardson and Spirtes, 2002; Richardson, 2003)

Why do we care?



(Bollen, 1989)

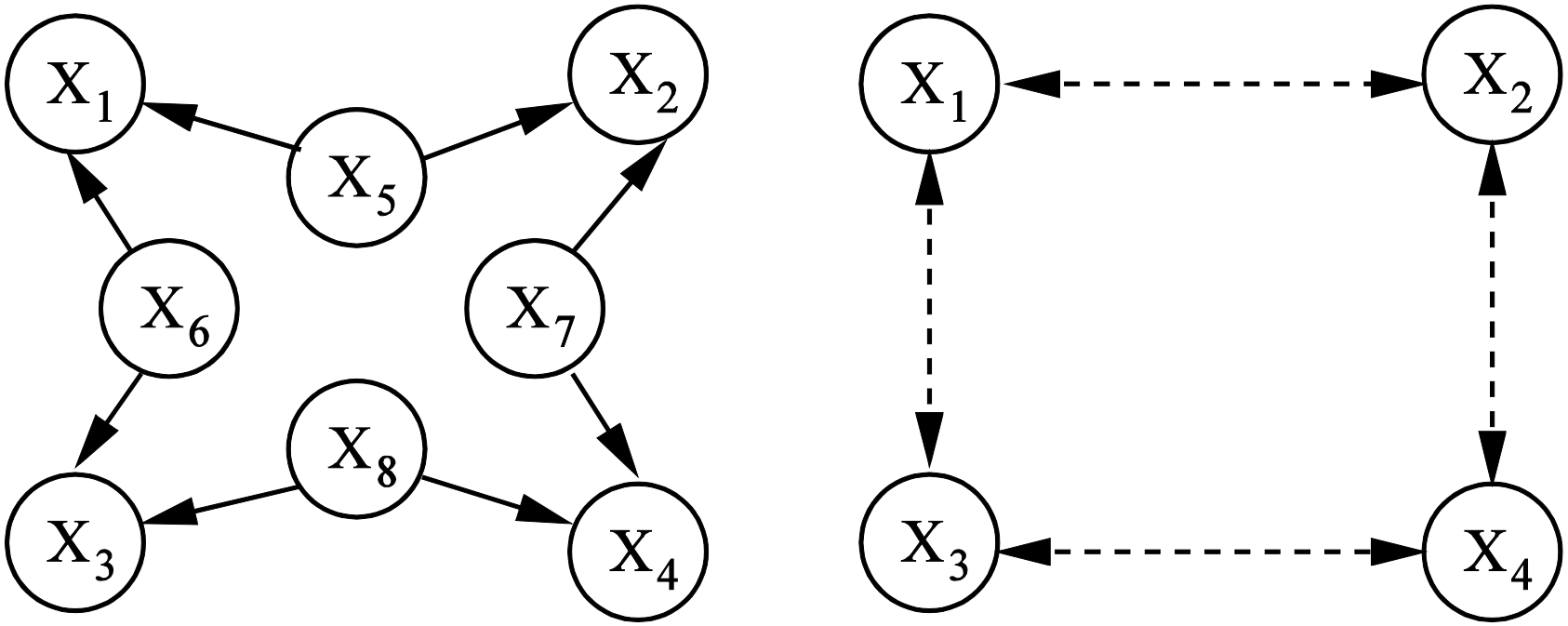
Why do we care?

- I like latent variables. Why not latent variables everywhere, everytime, latent variables in my cereal, no questions asked?
  - ADMG models open up new ways of parameterizing distributions
  - New ways of computing estimators
  - Theoretical advantages in some important cases (Richardson and Spirtes, 2002)

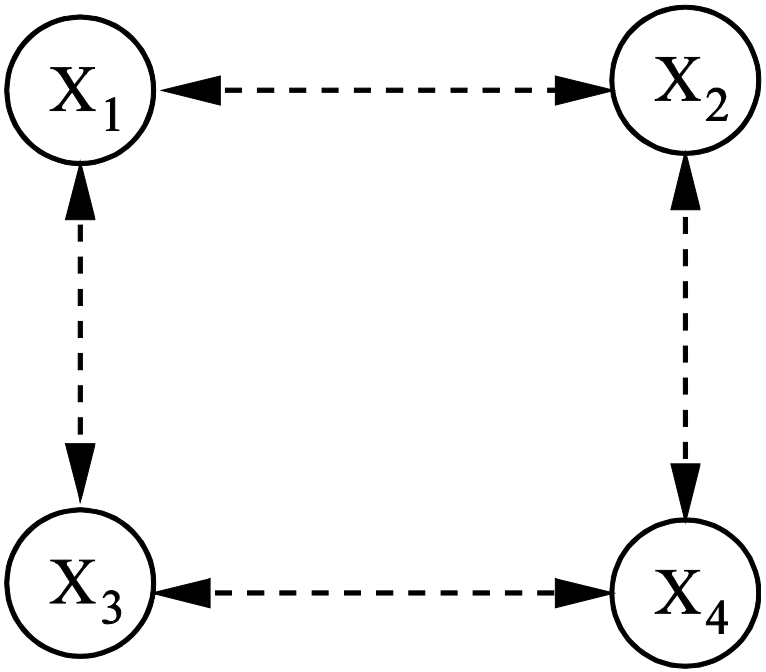
The talk in a nutshell

- The challenge:
  - How to specify families of distributions that respect the ADMG independence model, requires no explicit latent variable formulation
  - How NOT to do it: make everybody independent!
  - Needed: rich families. How rich?
- Contribution:
  - a new construction that is fairly general, easy to use, and complements the state-of-the-art
- First, a review:
  - current parameterizations, the good and bad issues
- For fun and profit: a simple demonstration on how to do Bayesianish parameter learning in these models

The Gaussian bi-directed model



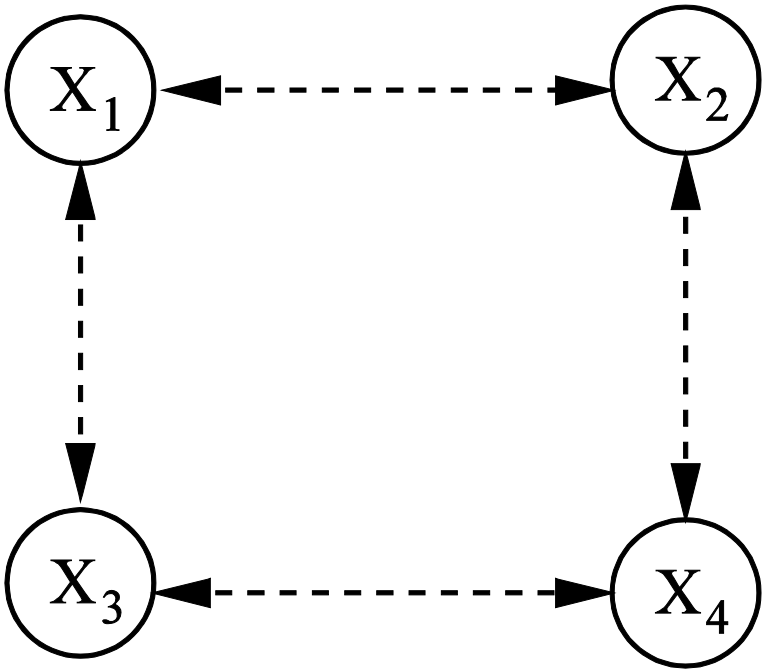
The Gaussian bi-directed case



$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & 0 \\ \sigma_{12} & \sigma_{22} & 0 & \sigma_{24} \\ \sigma_{13} & 0 & \sigma_{33} & \sigma_{34} \\ 0 & \sigma_{24} & \sigma_{34} & \sigma_{44} \end{bmatrix}$$

(Drton and Richardson, 2003)

Binary bi-directed case: the constrained Moebius parameterization



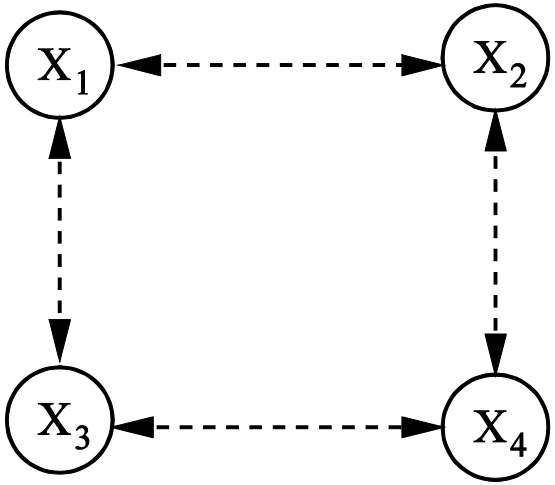
$$q_A \equiv P(X_A = 0)$$

$$P(X_A = 0, X_{V \setminus A} = 1) = \sum_{B : A \subseteq B} (-1)^{|B \setminus A|} q_B$$

(Drton and Richardson, 2008)

Binary bi-directed case: the constrained Moebius parameterization

- *Disconnected sets* are marginally independent. Hence, define q_A for connected sets only



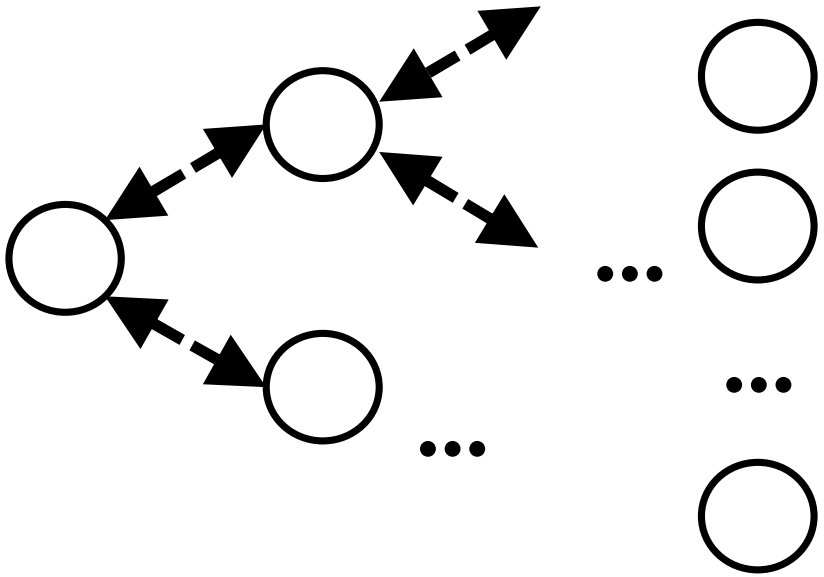
$$P(X_1 = 0, X_4 = 0) = P(X_1 = 0)P(X_4 = 0)$$

$$q_{14} = q_1 q_4$$

(However, notice there *is* a parameter q_{1234})

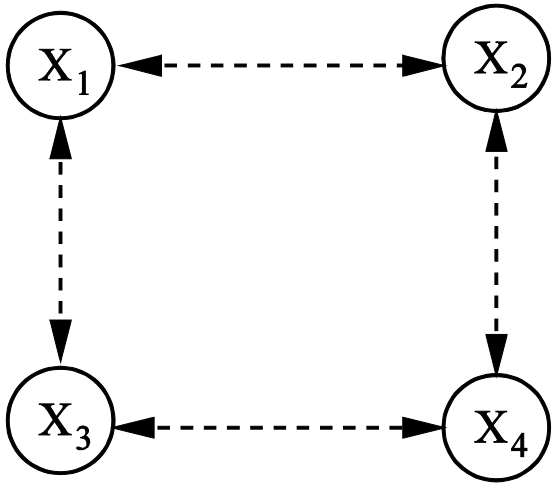
Binary bi-directed case: the constrained Moebius parameterization

- The good:
  - this parameterization is *complete*. *Every single binary bi-directed model* can be represented with it
- The bad:
  - Moebius inverse is intractable, and number of connected sets can grow exponentially even for trees



The Cumulative Distribution Network (CDN) approach

- Parameterizing cumulative distribution functions (CDFs) by a product of functions defined over subsets
  - Sufficient condition: each factor is a CDF itself
  - Independence model: the "same" as the bi-directed graph... but with extra constraints



$$F(X_{1234}) = F_1(X_{12})F_2(X_{24})F_3(X_{34})F_4(X_{13})$$

$$X_1 \perp\!\!\!\perp X_4$$

$$X_1 \not\perp\!\!\!\perp X_4 \mid X_2 \quad \textbf{etc}$$

(Huang and Frey, 2008)

Relationship

- CDN: the resulting PMF (usual CDF2PMF transform)

$$\sum_{z_1=0}^{1} \cdots \sum_{z_d=0}^{1} (-1)^{z_1+z_2+\ldots z_d} F(x_1 - z_1, \ldots, x_d - z_d)$$

- Moebius: the resulting PMF is equivalent

$$P(X_A = 0, X_{V \setminus A} = 1) = \sum_{B:A \subseteq B} (-1)^{|B \setminus A|} q_B$$

- Notice: $q_B = P(X_B = 0) = P(X_{\setminus B} \leq 1, X_{\setminus B} \leq 0)$
- However, in a CDN, parameters further factorize over cliques

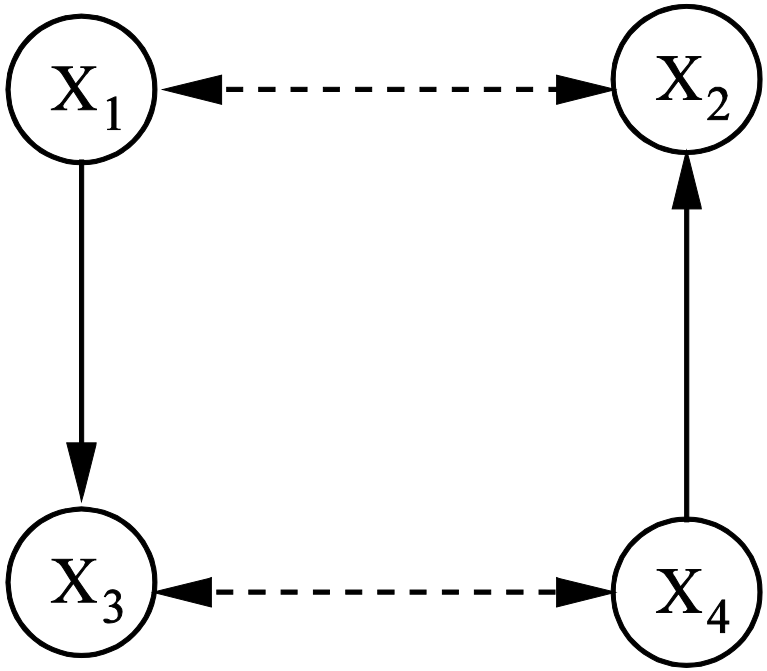
$$q_{1234} = q_{12} q_{13} q_{24} q_{34}$$

Relationship

- In the binary case, CDN models are a strict subset of Moebius models
- Moebius should still be the approach of choice for small networks where independence constraints are the main target
  - E.g., jointly testing the implication of independence assumptions
- But...
  - CDN models have a reasonable number of parameters, they are flexible, for small treewidths any fitting criterion is tractable, and learning is trivially tractable anyway by marginal composite likelihood estimation
    - Take-home message: a still flexible bi-directed graph model with no need for latent variables to make fitting "tractable"

The Mixed CDN model (MCDN)

- How to construct a distribution Markov to this?



- The binary ADMG parameterization by Richardson (2009) is complete, but with the same computational shortcomings
  - And how to easily extend it to non-Gaussian, infinite discrete cases, etc.?

Step 1: The high-level factorization

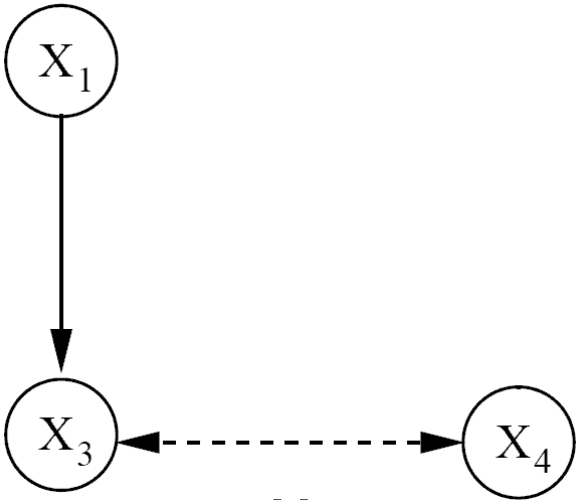
- A *district* is a maximal set of vertices connected by bi-directed edges
- For an ADMG G with vertex set X_V and districts $\{D_i\}$, define

$$P(X_V) = \prod_{i=1}^{K} P_i(X_{D_i} \mid pa_{\mathcal{G}}(X_{D_i}) \backslash X_{D_i})$$

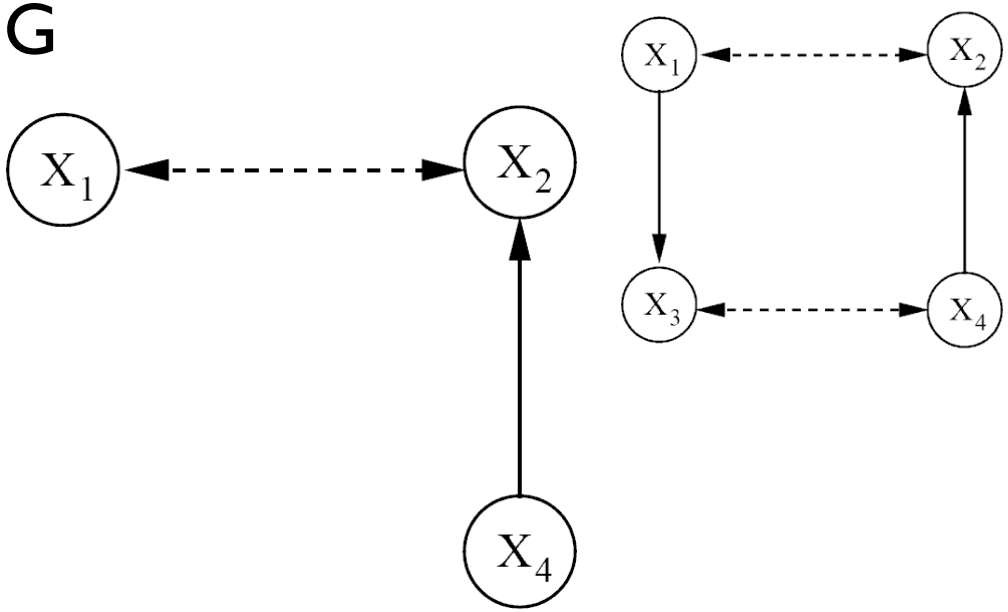
where $P(\cdot)$ is a density/mass function and $pa_G(\cdot)$ are parent of the given set in G

Step 1: The high-level factorization

- Also, assume that each $P_i(\cdot \mid \cdot)$ is Markov with respect to subgraph G_i – the graph we obtain from the corresponding subset
- We can show the resulting distribution is Markov with respect to the ADMG



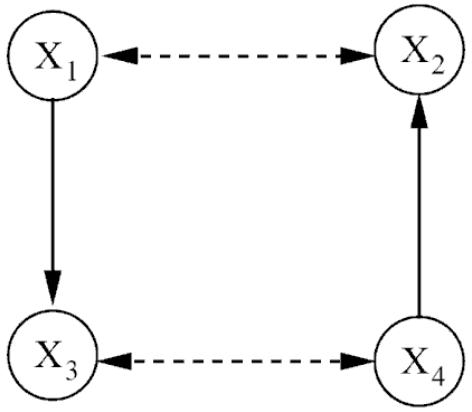
$X_4 \perp\!\!\!\perp X_1$



$X_4 \perp\!\!\!\perp X_1$

Step 1: The high-level factorization

- Despite the seemingly "cyclic" appearance, this factorization always gives a valid $P(\cdot)$ for any choice of $P_i(\cdot \mid \cdot)$

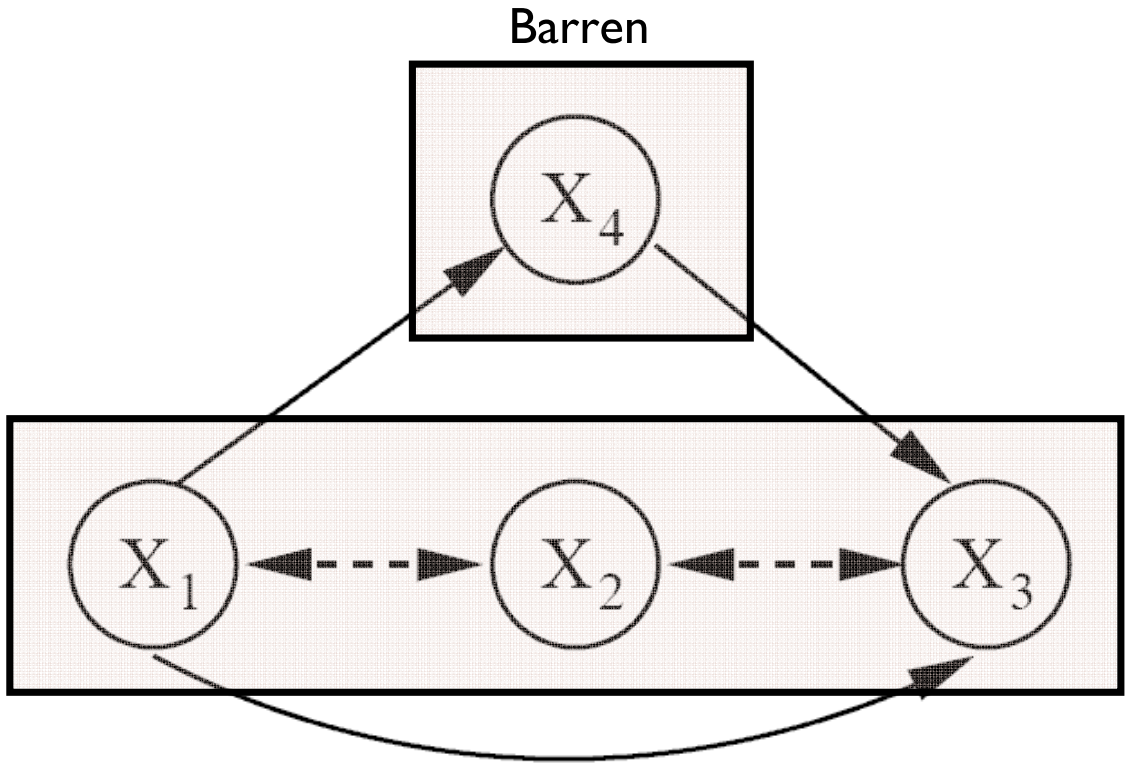


$$P(X_{134}) = \Sigma_{x2} P(X_1, x_2 \mid X_4) P(X_3, X_4 \mid X_1)$$
$$\equiv P(X_1 \mid X_4) P(X_3, X_4 \mid X_1)$$

$$P(X_{13}) = \Sigma_{x4} P(X_1 \mid x_4) P(X_3, x_4 \mid X_1)$$
$$= \Sigma_{x4} P(X_1) P(X_3, x_4 \mid X_1)$$
$$\equiv P(X_1) P(X_3 \mid X_1)$$

Step 2: Parameterizing P_i (barren case)

- D_i is a "barren" district is there is no directed edge within it



Step 2: Parameterizing P_i (barren case)

- For a district D_i with a clique set C_i (with respect bi-directed structure), start with a product of conditional CDFs

$$F_i(x_{D_i} \mid pa_{\mathcal{G}}(X_{D_i})) \equiv \prod_{X_S \in \mathcal{C}_i} F_S(x_S \mid pa_{\mathcal{G}}(X_{D_i}))$$

- Each factor $F_S(x_S \mid x_P)$ is a conditional CDF function, $P(X_S \leq x_S \mid X_P = x_P)$. (They have to be transformed back to PMFs/PDFs when writing the full likelihood function.)
- On top of that, each $F_S(x_S \mid x_P)$ is defined to be Markov with respect to the corresponding G_i
- We show that the corresponding product is Markov with respect to G_i

Step 2a: A copula formulation of P_i

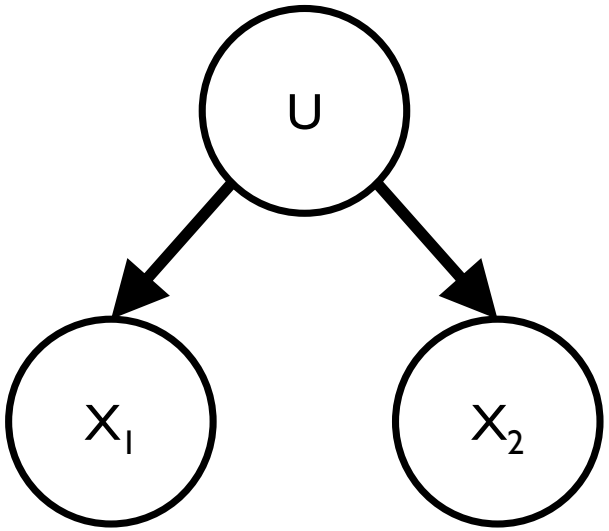
- Implementing the local factor restriction could be potentially complicated, but the problem can be easily approached by adopting a *copula* formulation

- A *copula function* is just a CDF with uniform [0, 1] marginals

- Main point: to provide a parameterization of a joint distribution that unties the parameters from the marginals from the remaining parameters of the joint

Step 2a: A copula formulation of P_i

- Gaussian latent variable analogy:



$$U \sim N(0, I)$$

$$X_1 = \lambda_1 U + e_1,\ e_1 \sim N(0, v_1)$$

$$X_2 = \lambda_2 U + e_2,\ e_2 \sim N(0, v_2)$$

Parameter sharing

Marginal of X_1: $N(0, \lambda_1^2 + v_1)$

Covariance of X_1, X_2: $\lambda_1 \lambda_2$

Step 2a: A copula formulation of P_i

- Copula idea: start from

$$F(X_1, X_2) = F(\, F_1^{-1}(F_1(X_1)), F_2^{-1}(F_2(X_2)))$$

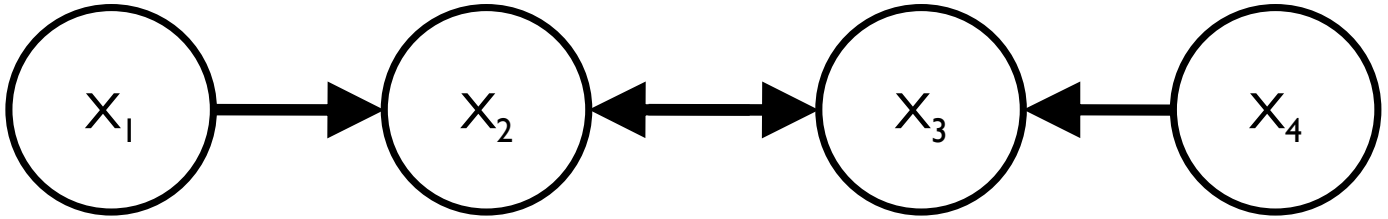
then define $H(Y_a, Y_b)$ accordingly, where $0 \le Y_* \le 1$

$$H(Y_a, Y_b) \equiv F(\, F_1^{-1}(Y_a), F_2^{-1}(Y_b))$$

- $H(\cdot, \cdot)$ will be a CDF with uniform [0, 1] marginals
- For any $F_i(\cdot)$ of choice, $U_i \equiv F_i(X_i)$ gives an uniform [0, 1]
- We mix-and-match any marginals we want with any copula function we want

Step 2a: A copula formulation of P_i

- The idea is to use a conditional marginal $F_i(X_i \mid pa(X_i))$ within a copula
- Example



$$U_2(x_1) \equiv P_2(X_2 \le x_2 \mid x_1) \qquad U_3(x_4) \equiv P_2(X_3 \le x_3 \mid x_4)$$

$$P(X_2 \le x_2, X_3 \le x_3 \mid x_1, x_4) = H(U_2(x_1), U_3(x_4))$$

- Check:

$$P(X_2 \le x_2 \mid x_1, x_4) = H(U_2(x_1), 1) = H(U_2(x_1))$$
$$= U_2(x_1) = P_2(X_2 \le x_2 \mid x_1)$$

Step 2a: A copula formulation of P_i

- Not done yet! We need this

$$F_i(x_{D_i} \mid pa_{\mathcal{G}}(X_{D_i})) \equiv \prod_{X_S \in \mathcal{C}_i} F_S(x_S \mid pa_{\mathcal{G}}(X_{D_i}))$$

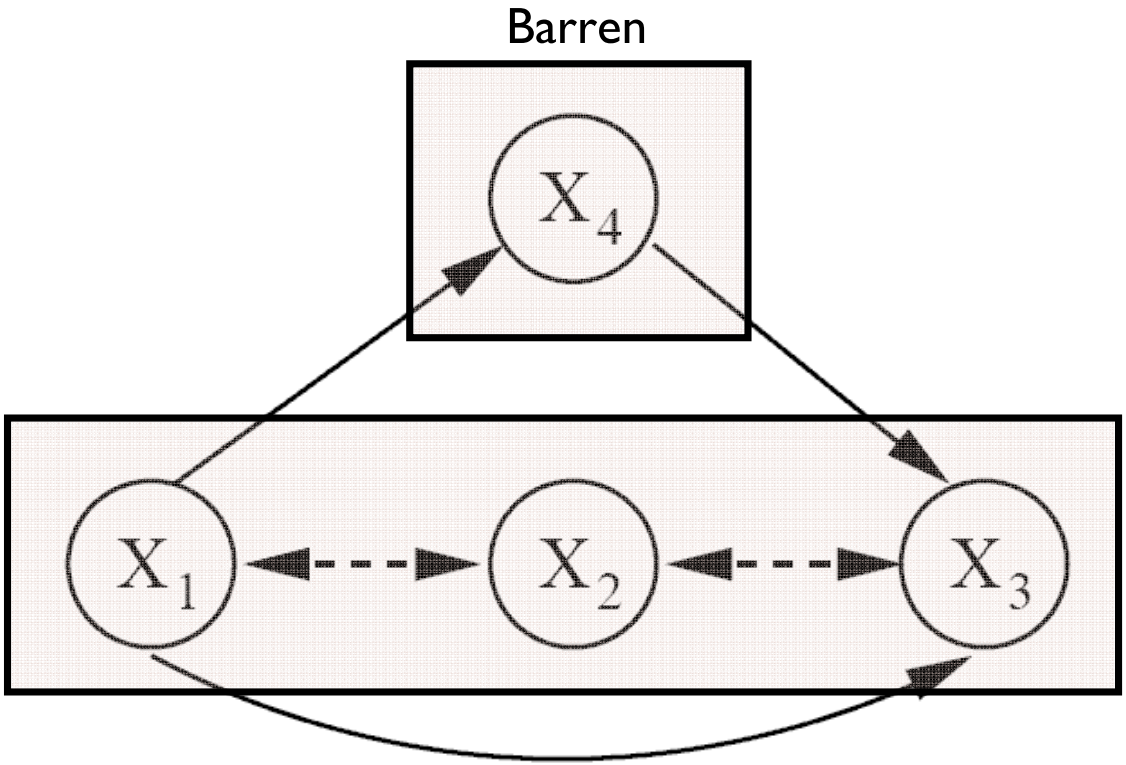
- Product of copulas is not a copula

- However, results in the literature are helpful here. It can be shown that plugging in $U_i^{1/d(i)}$, instead of U_i will turn the product into a copula
  - where *d(i)* is the number of bi-directed cliques containing X_i

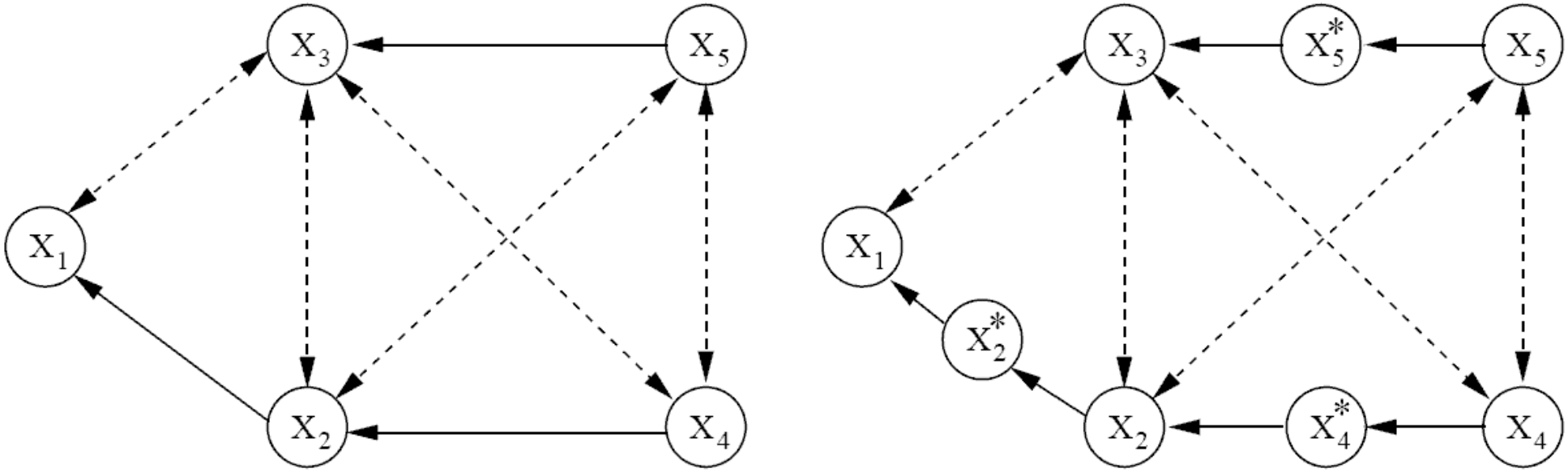
Liebscher (2008)

Step 3: The non-barren case

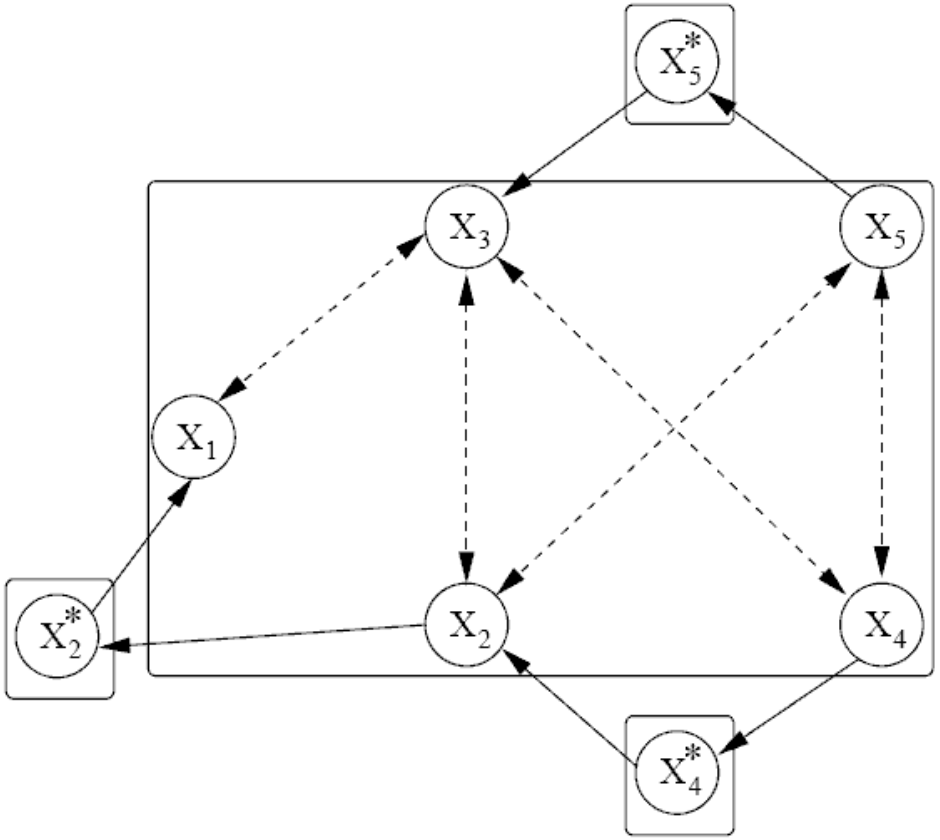
- What should we do in this case?



Step 3: The non-barren case



Step 3: The non-barren case



$$P_v^\star(X_v^\star = x \mid X_v = x) = 1$$

$$\begin{aligned} P(X_V, X_V^\star) &= \prod_{i=1}^{K} P_i(X_{D_i} \mid pa_{\mathcal{G}^\star}(X_{D_i}) \backslash X_{D_i}) \\ &\times \prod_{X_v \in X_V} P_v^\star(X_v^\star \mid X_v) \end{aligned}$$

Parameter learning

- For the purposes of illustration, assume a finite mixture of experts for the conditional marginals for continuous data

$$f_v(x_v \mid pa_{\mathcal{G}}(X_v)) = \sum_{z=1}^{K} \pi_{z;v} \mathcal{N}(x_v;\ \mu_{z;v}, \sigma^2_{z;v})$$

$$\mu_{z;v}(pa_{\mathcal{G}}(X_v)) = \theta_{v0} + \theta_v^{\mathsf{T}} pa_{\mathcal{G}}(X_v)$$

$$\pi_{z;v}(pa_{\mathcal{G}}(X_v)) \propto \exp(w_{v0} + w_v^{\mathsf{T}} pa_{\mathcal{G}}(X_v))$$

- For discrete data, just use the standard CPT formulation found in Bayesian networks

Parameter learning

- Copulas: we use a bi-variate formulation only (so we take products "over edges" instead of "over cliques").
- In the experiments: Frank copula

$$C_F(u_i, u_j; \alpha) = -\frac{1}{\alpha} \ln \left(1 + \frac{(e^{-\alpha u_i} - 1)(e^{-\alpha u_j} - 1)}{e^{-\alpha} - 1} \right)$$

Parameter learning

- Suggestion: two-stage quasiBayesian learning
  - Analogous to other approaches in the copula literature
  - Fit marginal parameters using the posterior expected value of the parameter for each individual mixture of experts
  - Plug those in the model, then do MCMC on the copula parameters
- Relatively efficient, decent mixing even with random walk proposals
  - Nothing stopping you from using a fully Bayesian approach, but mixing might be bad without some smarter proposals
- Notice: needs constant CDF-to-PDF/PMF transformations!

Experiments

Data set	**Data type**	**#V**	**#D**	$\mathbb{E}[\#\leftrightarrow]$	$\mathbb{E}[\#\rightarrow]$
SPECT	Binary	23	267	4.1	25.6
Breast cancer wisconsin	Ordinal	10	683	5.1	16.3
Soybean (large)	Ordinal	33	266	9.3	39.8
Parkinsons	Continuous	15	5875	8.9	18.2
Ionosphere	Continuous	32	351	12.4	32.8
Wine quality (red)	Continuous	11	1599	5.7	7.5
Wine quality (white)	Continuous	11	4898	7.3	14.5

Experiments

Data set	Gaussian/probit	Copula MCDN	Difference
SPECT	-11.32	-11.11	0.21 ± 0.06 $\star$
Breast cancer wisconsin	-12.60	-12.77	-0.17 ± 0.11
Soybean (large)	-20.17	-17.71	2.46 ± 0.20 $\star$
Parkinsons	-11.65	-3.48	8.17 ± 0.28 $\star$
Ionosphere	-41.10	-27.45	13.64 ± 0.67 $\star$
Wine quality (red)	-13.72	-11.25	2.47 ± 0.10 $\star$
Wine quality (white)	-13.76	-12.11	1.65 ± 0.09 $\star$

Conclusion

- General toolbox for construction for ADMG models
- Alternative estimators would be welcome:
  - Bayesian inference is still "doubly-intractable" (Murray et al., 2006), but district size might be small enough even if one has many variables
  - Either way, composite likelihood still simple. Combined with the Huang + Frey dynamic programming method, it could go a long way
- Structure learning: how would this parameterization help?
- Empirical applications in problems with extreme value issues, exploring non-independence constraints, relations to effect models in the potential outcome framework etc.

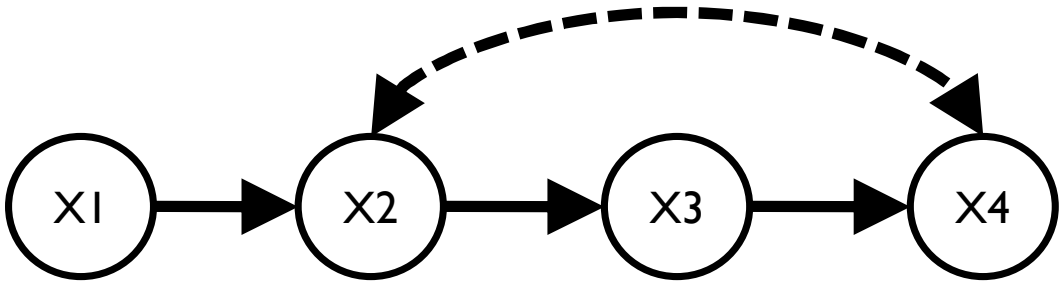
Acknowledgements

- Thanks to Thomas Richardson for several useful discussions

Thank you

Appendix: Limitations of the Factorization

- Consider the following network



$$P(X_{1234}) = P(X_2, X_4 \mid X_1, X_3)P(X_3 \mid X_2)P(X_1)$$

$$\Sigma_{x2}P(X_{1234}) \,/\, (P(X_3 \mid X_2)P(X_1)) = \Sigma_{x2}\, P(X_2, X_4 \mid X_1, X_3)$$

$$\Sigma_{x2}P(X_{1234}) \,/\, (P(X_3 \mid X_2)P(X_1)) = f(X_3, X_4)$$