Tree-based mixtures: Application in epidemiology

Stéphane Robin

joint work with P. Barbillon, L. Schwaller
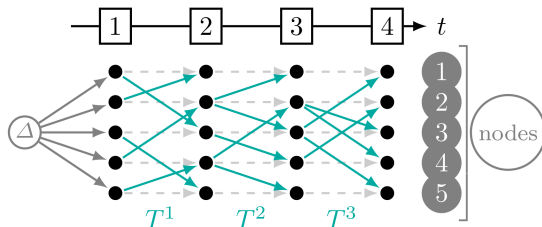
Sorbonne Université / LPSM

Bayesian Methods for the Social Sciences, IHP, Oct. 2022

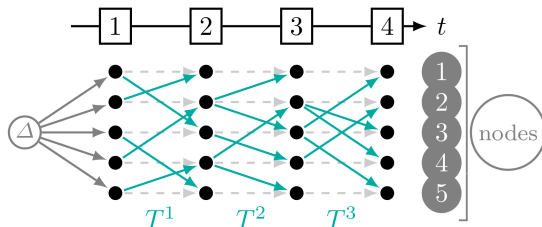# Spred of an 'epidemics' [BSR+19]

Data: n times, p individuals,

$$Y_{tj} = \text{status ('sick' / 'healthy') of individual } j \text{ at time } t$$

Spred of an 'epidemics' [BSR+19]

Data: n times, p individuals,

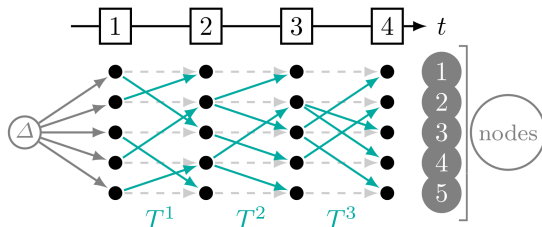$$Y_{tj} = \text{status ('sick' / 'healthy') of individual } j \text{ at time } t$$

Assumption: 'Contamination' spreads according to some 'social' network (not anybody can be 'contaminated' by anybody)

Spred of an 'epidemics' [BSR+19]

Data: n times, p individuals,

$$Y_{tj} = \text{status ('sick' / 'healthy') of individual } j \text{ at time } t$$

Assumption: 'Contamination' spreads according to some 'social' network (not anybody can be 'contaminated' by anybody)

Final aim: Learn something about this network.

For now: Observe that the path of the 'epidemics' is a tree.

Outline

Reminder on graphical models and trees

Tree shaped distributions and mixtures

Spread of an epidemics

Graphical models

Undirected graphical model [Lau96]: the multivariate distribution p is *faithful* to the graph G iff

$$p(Y) \propto \prod_{C \in \mathcal{C}(G)} \psi_C(Y^C)$$

(i.e., iff p can be factorized according to the set $\mathcal{C}(G)$ of maximal cliques of G)

Graphical models

Undirected graphical model [Lau96]: the multivariate distribution p is *faithful* to the graph G iff

$$p(Y) \propto \prod_{C \in \mathcal{C}(G)} \psi_C(Y^C)$$

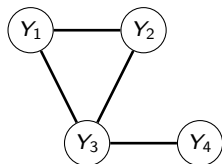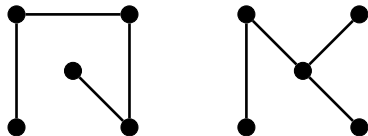(i.e., iff p can be factorized according to the set $\mathcal{C}(G)$ of maximal cliques of G)

Example:

$p(Y) \propto \psi_1(Y_1, Y_2, Y_3) \, \psi_2(Y_3, Y_4)$

▸ Connected graph: all variables are dependent

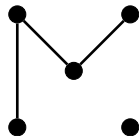▸ $Y_3 =$ separator: $(Y_4 \perp (Y_1, Y_2)) \mid Y_3$

Tree-shaped graphical model

Spanning tree = acyclic graph connecting all the nodes

spanning trees   not spanning   not a tree

Tree-shaped graphical model

Spanning tree = acyclic graph connecting all the nodes

| spanning trees | not spanning | not a tree |
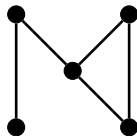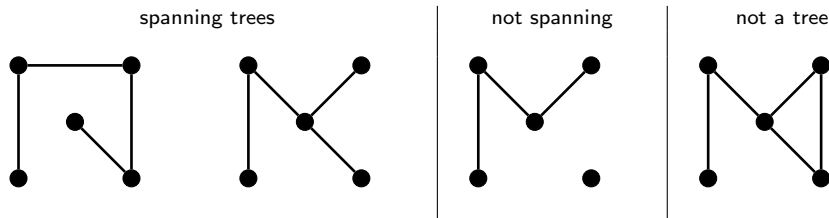
Tree-shaped graphical model: p faithful to the spanning tree T

$$\Leftrightarrow \qquad p(Y) \propto \prod_{(j,k)\in T} \psi_{jk}(Y_j, Y_k)$$

(cliques are only edges)

$\rightarrow$ All variables are dependent (*spanning*) but few are conditionnaly dependent (*tree*)

Maximum likelihood inference

Consider $Y \sim p$, faithful to T:

$$\log p_T(Y) = \sum_{(j,k) \in T} \log \psi_{jk}(Y_j, Y_k) + \mathrm{cst}$$

---

[1]Actually $\psi_{jk}(Y_j, Y_k) = p(Y_j, Y_k)/\big(p(Y_j)p(Y_k)\big)$, so edge weight = mutual information

Maximum likelihood inference

Consider $Y \sim p$, faithful to T:

$$\log p_T(Y) = \sum_{(j,k) \in T} \log \psi_{jk}(Y_j, Y_k) + \text{cst}$$

Suppose each ψ_{jk} has a parametric form and that MLEs $\widehat{\psi}_{jk}$ is available.

---

[1]Actually $\psi_{jk}(Y_j, Y_k) = p(Y_j, Y_k)/\left(p(Y_j)p(Y_k)\right)$, so edge weight = mutual information

Maximum likelihood inference

Consider $Y \sim p$, faithful to T:

$$\log p_T(Y) = \sum_{(j,k) \in T} \log \psi_{jk}(Y_j, Y_k) + \text{cst}$$

Suppose each ψ_{jk} has a parametric form and that MLEs $\widehat{\psi}_{jk}$ is available.

Tree MLE. Finding

$$\widehat{T} = \arg\max_{T \in \mathcal{T}} \log p_T(Y)$$

amounts to solve a maximum spanning tree problem, where edge (j, k) has weight

$$\log \widehat{\psi}_{jk}(Y_j, Y_k)$$

(Chow & Liu algorithm[1]: [CL68])

---

[1] Actually $\psi_{jk}(Y_j, Y_k) = p(Y_j, Y_k) / \big(p(Y_j) p(Y_k) \big)$, so edge weight = mutual information

Outline

Reminder on graphical models and trees

Tree shaped distributions and mixtures

Spread of an epidemics

Tree-based mixture

Mixture of tree-shaped distributions [MJ06,Kir07].

- ▸ Tree marginal distribution: product form

$$p(T) = B^{-1} \prod_{(j,k) \in T} \beta_{jk}, \qquad B = \sum_{T \in \mathcal{T}} \prod_{(j,k) \in T} \beta_{jk}.$$

Tree-based mixture

Mixture of tree-shaped distributions [MJ06,Kir07].

▸ Tree marginal distribution: product form

$$p(T) = B^{-1} \prod_{(j,k) \in T} \beta_{jk}, \qquad B = \sum_{T \in \mathcal{T}} \prod_{(j,k) \in T} \beta_{jk}.$$

▸ Data conditional distribution (likelihood):

$$p(Y \mid T) \propto \prod_{(j,k) \in T} \psi_{jk}(Y_j, Y_k).$$

Tree-based mixture

Mixture of tree-shaped distributions [MJ06,Kir07].

- Tree marginal distribution: product form

$$p(T) = B^{-1} \prod_{(j,k) \in T} \beta_{jk}, \qquad B = \sum_{T \in \mathcal{T}} \prod_{(j,k) \in T} \beta_{jk}.$$

- Data conditional distribution (likelihood):

$$p(Y \mid T) \propto \prod_{(j,k) \in T} \psi_{jk}(Y_j, Y_k).$$

- Data marginal distribution:

$$p(Y) = \sum_{T \in \mathcal{T}} p(T) p(Y \mid T) \propto \sum_{T \in \mathcal{T}} \prod_{(j,k) \in T} \underbrace{\beta_{jk} \psi_{jk}(Y_j, Y_k)}_{w_{jk}(Y)}.$$

Tree-based mixture

Mixture of tree-shaped distributions [MJ06,Kir07].

- ▸ Tree marginal distribution: product form

$$p(T) = B^{-1} \prod_{(j,k) \in T} \beta_{jk}, \qquad B = \sum_{T \in \mathcal{T}} \prod_{(j,k) \in T} \beta_{jk}.$$

- ▸ Data conditional distribution (likelihood):

$$p(Y \mid T) \propto \prod_{(j,k) \in T} \psi_{jk}(Y_j, Y_k).$$

- ▸ Data marginal distribution:

$$p(Y) = \sum_{T \in \mathcal{T}} p(T) p(Y \mid T) \propto \sum_{T \in \mathcal{T}} \prod_{(j,k) \in T} \underbrace{\beta_{jk} \psi_{jk}(Y_j, Y_k)}_{w_{jk}(Y)}.$$

- ▸ Tree conditional distribution: product form

$$p(T \mid Y) = C^{-1} \prod_{(j,k) \in T} w_{jk}(Y), \qquad C = \sum_{T \in \mathcal{T}} \prod_{(j,k) \in T} w_{jk}(Y).$$

Summing over trees

We need to sum over the set $\mathcal{T}$ of spanning trees, but

$$\text{card}\,(\mathcal{T}) = p^{p-2}$$

Summing over trees

We need to sum over the set $\mathcal{T}$ of spanning trees, but

$$\text{card}\,(\mathcal{T}) = p^{p-2}$$

Matrix-Tree Theorem (Kirchhoff) [Cha82]. Let $W = [w_{jk}]$ be a symmetric matrix with null diagonal and $\Delta(W)$, its Laplacian:

$$\Delta(W)_{jk} = \left\{ \begin{array}{ll} \sum_k w_{jk} & \text{if } j = k \\ -w_{jk} & \text{otherwise} \end{array} \right.$$

then

Summing over trees

We need to sum over the set $\mathcal{T}$ of spanning trees, but

$$\text{card}\,(\mathcal{T}) = p^{p-2}$$

Matrix-Tree Theorem (Kirchhoff) [Cha82]. Let $W = [w_{jk}]$ be a symmetric matrix with null diagonal and $\Delta(W)$, its Laplacian:

$$\Delta(W)_{jk} = \left\{ \begin{array}{ll} \sum_k w_{jk} & \text{if } j = k \\ -w_{jk} & \text{otherwise} \end{array} \right.$$

then

1. all the cofactors $[\Delta(W)]^{uv}$ of $\Delta(W)$ are equal

Summing over trees

We need to sum over the set $\mathcal{T}$ of spanning trees, but

$$\text{card}\,(\mathcal{T}) = p^{p-2}$$

Matrix-Tree Theorem (Kirchhoff) [Cha82]. Let $W = [w_{jk}]$ be a symmetric matrix with null diagonal and $\Delta(W)$, its Laplacian:

$$\Delta(W)_{jk} = \left\{ \begin{array}{ll} \sum_k w_{jk} & \text{if } j = k \\ -w_{jk} & \text{otherwise} \end{array} \right.$$

then

1. all the cofactors $[\Delta(W)]^{uv}$ of $\Delta(W)$ are equal

2. and

$$[\Delta(W)]^{uv} = \sum_{T \in \mathcal{T}} \prod_{(j,k) \in T} w_{jk}$$

(i.e.: $\sum_{T \in \mathcal{T}}$ can be computed at the price of a determinant: $O(p^3)$)

Edge probability

We may not be interested in the whole (conditional) distribution of T, but rather on edge probabilities:

$$\Pr\{(j, k) \in T\} \qquad \text{or} \qquad \Pr\{(j, k) \in T \mid Y\}.$$

Edge probability

We may not be interested in the whole (conditional) distribution of T, but rather on edge probabilities:

$$\Pr\{(j,k) \in T\} \qquad \text{or} \qquad \Pr\{(j,k) \in T \mid Y\}.$$

Theorem [Kir07]. Denote W^{ab} the same matrix as W but setting $w_{ab} = w_{ba} = 0$, then

$$[\Delta(W^{ab})]^{ab} = \sum_{T \in \mathcal{T} : (a,b) \notin T} \prod_{(j,k) \in T} w_{jk}$$

Edge probability

We may not be interested in the whole (conditional) distribution of T, but rather on edge probabilities:
$$\Pr\{(j, k) \in T\} \qquad \text{or} \qquad \Pr\{(j, k) \in T \mid Y\}.$$

Theorem [Kir07]. Denote W^{ab} the same matrix as W but setting $w_{ab} = w_{ba} = 0$, then
$$[\Delta(W^{ab})]^{ab} = \sum_{T \in \mathcal{T} : (a,b) \notin T} \prod_{(j,k) \in T} w_{jk}$$

▸ Gives access to $\Pr\{(j, k) \notin T\}$ or $\Pr\{(j, k) \notin T\}$

▸ All $[\Delta(W^{ab})]^{ab}$ can be computed at the price of one matrix inversion: $O(p^3)$

Outline

Reminder on graphical models and trees

Tree shaped distributions and mixtures

Spread of an epidemics

Susceptible-Infected-Susceptible (SIS) model

Data. n times, p individuals

$$Y_{tj} = \left\{ \begin{array}{ll} 1 & \text{if individual } j \text{ is infected at time } t \\ 0 & \text{otherwise} \end{array} \right.$$

$\rightarrow$ Complete observations

Susceptible-Infected-Susceptible (SIS) model

Data. n times, p individuals

$$Y_{tj} = \left\{ \begin{array}{ll} 1 & \text{if individual } j \text{ is infected at time } t \\ 0 & \text{otherwise} \end{array} \right.$$

$\rightarrow$ Complete observations

Model. At each time t, each node k

▸ picks up a parent j at time $t-1$ with probability $\propto \beta_{jk}$ and,

▸ denoting $\psi_{jk}^t = \Pr\{Y_{t+1,k} \mid Y_{t,k}, Y_{t,j}\}$, it evolves according to

ψ_{jk}^t		$Y_{t+1,k} = 0$	$Y_{t+1,k} = 1$
$Y_{t,k} = 1$	$Y_{t,j} = 1$	e	$1 - e$
$Y_{t,k} = 1$	$Y_{t,j} = 0$	e	$1 - e$
$Y_{t,k} = 0$	$Y_{t,j} = 1$	$1 - c$	c
$Y_{t,k} = 0$	$Y_{t,j} = 0$	1	0

▸ c = contamination rate
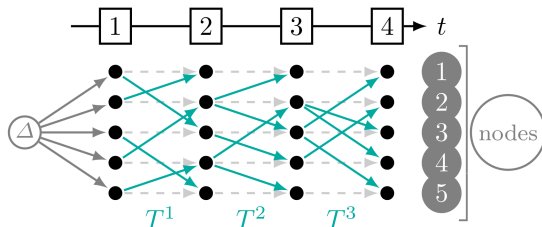
▸ e = extinction rate (become susceptible again)

A tree-shaped path

Adding a fictitious root Δ (at time 0), a path is a tree $T = (T^1, \ldots, T^{n-1})$:
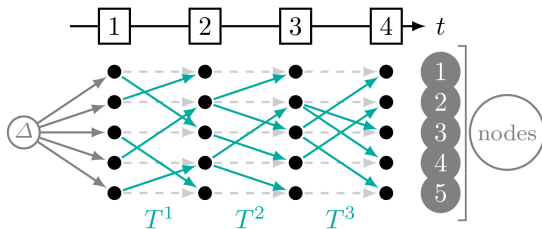
A tree-shaped path

Adding a fictitious root Δ (at time 0), a path is a tree $T = (T^1, \ldots, T^{n-1})$:

Define $\mathcal{T}_\Delta$ the set of oriented spanning tree

- over the nodes $\Delta \cup \{(t, k) : 1 \leqslant t \leqslant n, 1 \leqslant k \leqslant p\}$,
- rooted in Δ,
- with edges connecting only time-adjacent nodes $(j \neq k)$

$\rightarrow$ $\mathcal{T}_\Delta$ = set of spanning trees going 'forward' in time $(|\mathcal{T}_\Delta| = (p-1)^{p(n-1)})$

Inference

EM algorithm. Denote

$$\theta = (\beta = [\beta_{jk}], e, c)$$

Inference

EM algorithm. Denote

$$\theta = (\beta = [\beta_{jk}], e, c)$$

E Step: Compute the conditional distribution

$$p(T_\Delta = (T^1, \dots T^{n-1}) \mid Y) = \frac{\prod_{t=1}^{n-1} \prod_{(j,k) \in T^t} w_{jk}^t}{\sum_{T \in \mathcal{T}_\Delta} \prod_{t=1}^{n-1} \prod_{(j,k) \in T^t} w_{jk}^t}, \qquad w_{jk}^t = \beta_{jk} \psi_{jk}^t$$

Inference

EM algorithm. Denote

$$\theta = (\beta = [\beta_{jk}], e, c)$$

E Step: Compute the conditional distribution

$$p(T_\Delta = (T^1, \dots T^{n-1}) \mid Y) = \frac{\prod_{t=1}^{n-1} \prod_{(j,k) \in T^t} w_{jk}^t}{\sum_{T \in \mathcal{T}_\Delta} \prod_{t=1}^{n-1} \prod_{(j,k) \in T^t} w_{jk}^t}, \qquad w_{jk}^t = \beta_{jk} \psi_{jk}^t$$

M Step: Update the parameters as

$$\widehat{\theta} = \arg\max \mathbb{E}\left[\log p_\theta(Y, T) \mid Y\right]$$

Inference

EM algorithm. Denote
$$\theta = (\beta = [\beta_{jk}], e, c)$$

E Step: Compute the conditional distribution
$$p(T_\Delta = (T^1, \dots T^{n-1}) \mid Y) = \frac{\prod_{t=1}^{n-1} \prod_{(j,k) \in T^t} w_{jk}^t}{\sum_{T \in \mathcal{T}_\Delta} \prod_{t=1}^{n-1} \prod_{(j,k) \in T^t} w_{jk}^t}, \qquad w_{jk}^t = \beta_{jk} \psi_{jk}^t$$

M Step: Update the parameters as
$$\widehat{\theta} = \arg\max \mathbb{E}\left[\log p_\theta(Y, T) \mid Y\right]$$
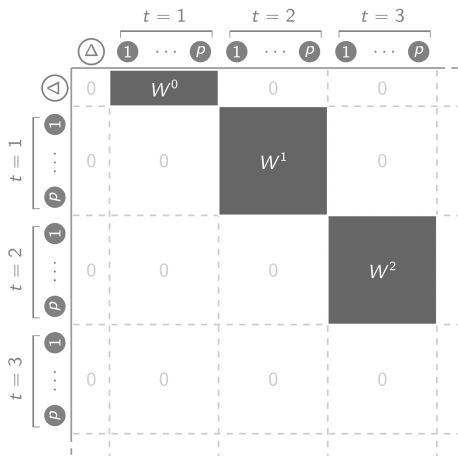
Critical step = E step.

▸ We now need to sum over the (huge) set of rooted oriented trees $\mathcal{T}_\Delta$.

▸ Hopefully, a alternative version of the matrix-tree theorem enables to sum over all *directed* trees (W asymmetric) with given root [Cha82].

An easy situation
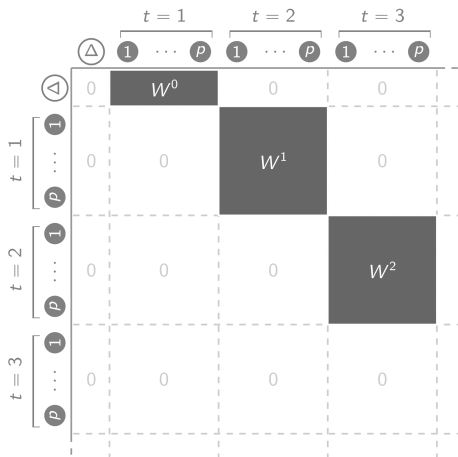
$W = [w_{jk}^t]$ where $w_{jk}^t = \beta_{jk}\psi_{jk}^t$: Edges connect only time-adjacent nodes

An easy situation

$W = [w_{jk}^t]$ where $w_{jk}^t = \beta_{jk}\psi_{jk}^t$:
Edges connect only time-adjacent nodes

$\rightarrow$ Laplacian $\Delta(W)$
is upper triangular

$\rightarrow$ Matrix-tree theorem:

$$\Delta(W)^{00} = \prod_{t,j}\left(\sum_k \beta_{jk}\psi_{jk}^t\right)$$

$\rightarrow$ computable in $O(np^2)$

In practice

Edge probabilities:

$$\text{At a given time } t: \qquad \Pr\{(j, k) \in T^t \mid Y\}$$
$$\text{At least in one time:} \qquad \Pr\{\exists t : (j, k) \in T^t \mid Y\}$$

In practice

Edge probabilities:

$$\text{At a given time } t: \qquad \Pr\{(j, k) \in T^t \mid Y\}$$
$$\text{At least in one time:} \qquad \Pr\{\exists t : (j, k) \in T^t \mid Y\}$$

Alternatives.

▸ Bayesian inference can be carried out for e an c

▸ Iterating the EM steps does not improve the performances very much

▸ Observing multiple waves of the epidemics (even over a shortest time-range) improves the accuracy (see next)

Simulation study

Design: Consider a graph G and launch the 'epidemics' along its edges.

Method: Predict if $(j, k) \in G$ based on edge probability $\Pr\{\exists t : (j, k) \in T^t \mid Y\}$.

Simulation study

Design: Consider a graph G and launch the 'epidemics' along its edges.

Method: Predict if $(j, k) \in G$ based on edge probability $\Pr\{\exists t : (j, k) \in T^t \mid Y\}$.

Parameters: $p = 20$ nodes

Simulation study

Design: Consider a graph G and launch the 'epidemics' along its edges.

Method: Predict if $(j, k) \in G$ based on edge probability $\Pr\{\exists t : (j, k) \in T^t \mid Y\}$.

Parameters: $p = 20$ nodes

$G = $ ER: Erdös,
$G = $ PA: preferential attachment

Simulation study

Design: Consider a graph G and launch the 'epidemics' along its edges.

Method: Predict if $(j, k) \in G$ based on edge probability $\Pr\{\exists t : (j, k) \in T^t \mid Y\}$.

Parameters: $p = 20$ nodes

G = ER: Erdös,
G = PA: preferential attachment

d = network density
e = .05

Simulation study

Design: Consider a graph G and launch the 'epidemics' along its edges.

Method: Predict if $(j, k) \in G$ based on edge probability $\Pr\{\exists t : (j, k) \in T^t \mid Y\}$.

Parameters: $p = 20$ nodes

$G = $ ER: Erdös,
$G = $ PA: preferential attachment

$d = $ network density
$e = .05$

uW: one wave ($n = 200$)
mW: 10 waves ($n = 20$)

Simulation study

Design: Consider a graph G and launch the 'epidemics' along its edges.

Method: Predict if $(j, k) \in G$ based on edge probability $\Pr\{\exists t : (j, k) \in T^t \mid Y\}$.

Parameters: $p = 20$ nodes

$G = $ ER: Erdös,
$G = $ PA: preferential attachment

$d = $ network density
$e = .05$

uW: one wave ($n = 200$)
mW: 10 waves ($n = 20$)

s: forcing edge probabilities to be symmetric

Simulation study

Design: Consider a graph G and launch the 'epidemics' along its edges.

Method: Predict if $(j, k) \in G$ based on edge probability $\Pr\{\exists t : (j, k) \in T^t \mid Y\}$.

Parameters: $p = 20$ nodes

G = ER: Erdös,
G = PA: preferential attachment

d = network density
e = .05

uW: one wave ($n = 200$)
mW: 10 waves ($n = 20$)

s: forcing edge probabilities to be symmetric

AUC:
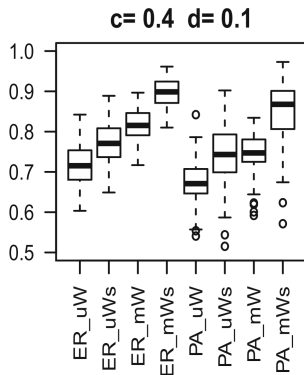
Illustration: Seed exchange network

Question: decipher the social structure underlying
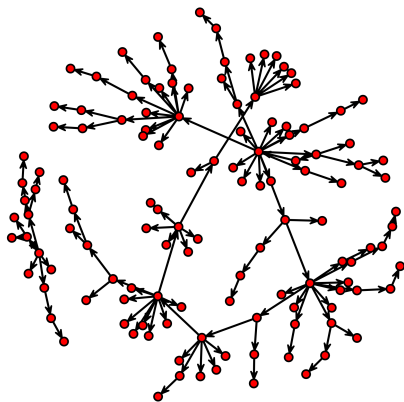seed exchanges between farmers

Telangana region (India) data:
- $p = 127$ farmers
- $n = 3$ years
- 14 seed varieties (waves)

$Y_{ti}^h = 1$ if farmer i holds variety h at time t.

No symmetry assumption.

More exchanges
- within the same caste
- within the same village
- from younger to older

Most probable donor for each farmer

Conclusion

Tree shaped mixtures

- Flexible model for multivariate distributions
- Base on a mixture with exponentially many components (p^{p-2})
- But giving access to edge probabilities at a low computational cost

Conclusion

Tree shaped mixtures

- ▸ Flexible model for multivariate distributions
- ▸ Base on a mixture with exponentially many components (p^{p-2})
- ▸ But giving access to edge probabilities at a low computational cost

Extensions

- ▸ 'Network' inference (= structure inference) [SRS19]
- ▸ Network comparison or network changes along time [SR17]
- ▸ Accounting for missing nodes ('actors') [RAR19,MRA20,MRA21]
- ▸ S-I-S model can be extended to more that two states (e.g. S-I-R models)

Conclusion

Tree shaped mixtures

- ▸ Flexible model for multivariate distributions
- ▸ Base on a mixture with exponentially many components (p^{p-2})
- ▸ But giving access to edge probabilities at a low computational cost

Extensions

- ▸ 'Network' inference (= structure inference) [SRS19]
- ▸ Network comparison or network changes along time [SR17]
- ▸ Accounting for missing nodes ('actors') [RAR19,MRA20,MRA21]
- ▸ S-I-S model can be extended to more that two states (e.g. S-I-R models)

Some questions

- ▸ Theoretical guaranties (e.g.: consistency of the estimated graph)?
- ▸ Numerical issues arising for large p or n (use tempering?)

References I

P. Barbillon, L. Schwaller, S. Robin, A. Flachs, and G.D. Stone. Epidemiologic network inference. *Statistics and Computing*, pages 1–15, 2019.

S. Chaiken. A combinatorial proof of the all minors matrix tree theorem. *SIAM Journal on Algebraic Discrete Methods*, 3(3):319–329, 1982.

C.K. Chow and C.N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, IT-14(3):462–467, 1968.

S. Kirshner. Learning with tree-averaged densities and distributions. In *NIPS*, pages 761–768, 2007.

S. Lauritzen. *Graphical Models*. Oxford Statistical Science Series. Clarendon Press, 1996.

M. Meilă and T. Jaakkola. Tractable Bayesian learning of tree belief networks. *Statistics and Computing*, March 2006.

R. Momal, S. Robin, and C. Ambroise. Tree-based inference of species interaction networks from abundance data. *Methods in Ecology and Evolution*, 11(5):621–632, 2020.

R. Momal, S. Robin, and C. Ambroise. Accounting for missing actors in interaction network inference from abundance data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2021.

G. Robin, C. Ambroise, and S. Robin. Incomplete graphical model inference via latent tree aggregation. *Statistical Modelling*, 19(5):545–568, 2019.

L. Schwaller and S. Robin. Exact Bayesian inference for off-line change-point detection in tree-structured graphical models. *Statistics and Computing*, 27(5):1331–1345, 2017.

L. Schwaller, S. Robin, and M. Stumpf. Bayesian inference of graphical model structures using trees. *J. Soc. Franc. Stat.*, 160(2):1–23, 2019.
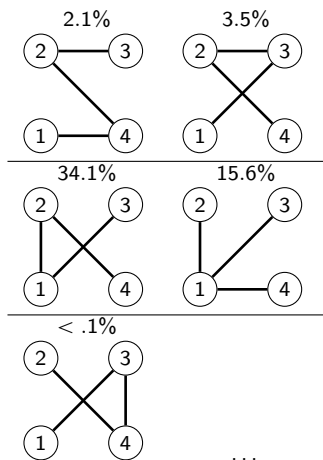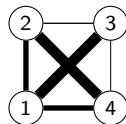
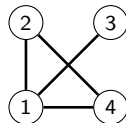# 'Tree-averaging' principle

Tree conditional distribution. $p(T|Y)$
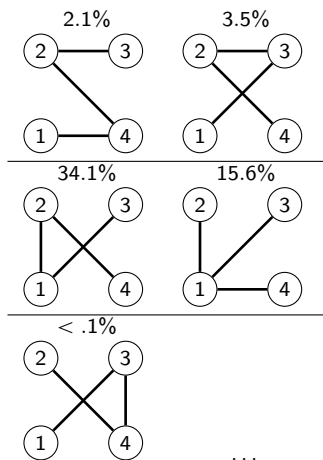
'Tree-averaging' principle

Tree conditional distribution. $p(T|Y)$
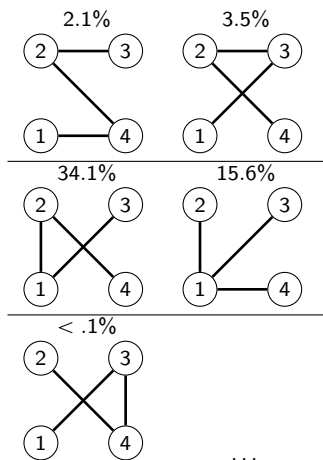
Edge probabilities. $\Pr\{(j, k) \in T \mid Y\}$
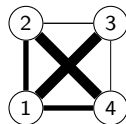
'Tree-averaging' principle

Tree conditional distribution. $p(T|Y)$

Edge probabilities. $\Pr\{(j,k) \in T \mid Y\}$

Most probable edges.

(not a tree) [SRS19]