Lecture 1: Trees, tree metric and tree spaces

Piotr Zwiernik

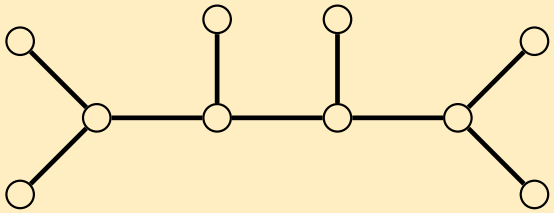
University of Genoa

Algebraic Statistics 2015
Genova
June 11, 2015

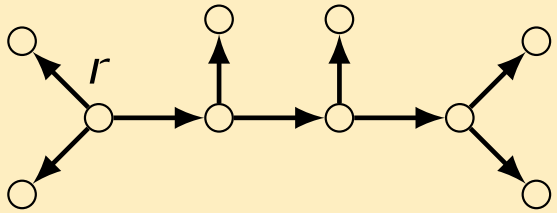
UNIVERSITÀ DEGLI STUDI
DI GENOVA

Trees

- **Definition**: Tree = undirected graph without cycles
  - tree $T = (V, E)$: V vertices, E edges

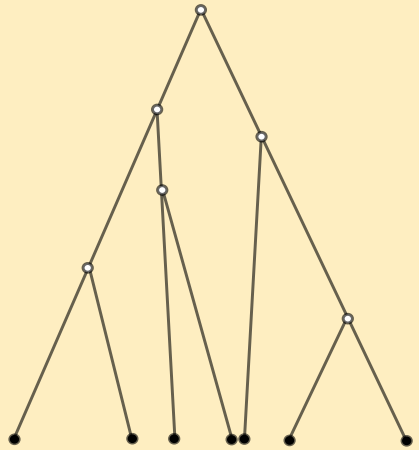


undirected



rooted

- rooted tree often depicted as...
- leaves = degree one nodes
- inner nodes = degree ≥ 2 nodes



Latent tree models

- Graphical models on trees have many nice properties
  - exponential families with explicit formulas for the MLE
  - dynamic programming for efficient computation of various probabilistic quantities
- Making some of the variables hidden gives greater flexibility

- **Definition***: Tree-decomposable distribution = marginal distribution of a tree distribution.
  - hidden variables are marginalized out
- Tree-decomposable distributions discussed by Judea Pearl as a natural extension of star-decomposable distributions (naive Bayes model, latent class model)

Judea Pearl, *Fusion, Propagation, and Structuring in Belief Networks*, Artificial Inteligence, 1986.

Motivation

- Applications in:
  - linguistics and bioinformatics – to model evolutionary processes
  - hierarchical clustering
  - image processing
- Important concept in causality
- Many well known statistical models are special cases
  - examples: hidden Markov models, naive Bayes models
  - general results can be used for these special cases

- Understand models with hidden data
  - the most tractable family of models with hidden variables
  - identifiability, geometry of the likelihood function

Alan S. Willsky, *Multiresolution Markov Models for Signal and Image Processing*, 2002.
Martin J. Wainwright, Michael I. Jordan, *Graphical Models, Exponential Families, and Variational Inference*, 2008.

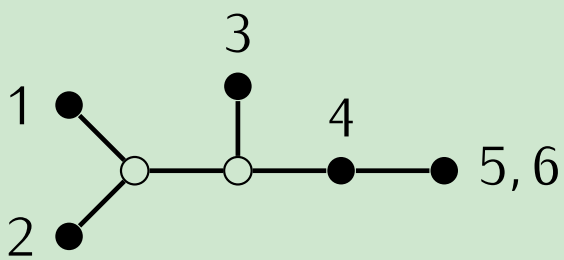
Short overview

- **Lecture 1: Trees, tree metrics and tree spaces**
- Lecture 2: Latent tree graphical models
- Lecture 3: Tree inference and parameter estimation
- Lecture 4: Likelihood geometry and model identifiability

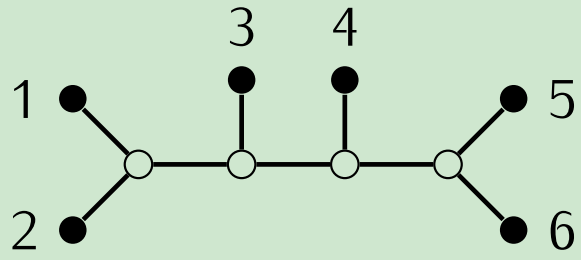
Main theme: phylogenetic combinatorics and results on tree metrics give a greater insight into the class of latent tree models

Semi-labeled trees and phylogenetic trees

- semi-labeled tree $\mathcal{T} = (T, \phi)$: $\phi : \{1, \ldots, m\} \to V$
  - all degree ≤ 2 nodes *need* to be labeled
  - multiple labels at a node are allowed
- phylogenetic tree = semi-labeled tree such that:
  - only leaves are labeled (there are no degree 2 nodes)
  - no multiple labels allowed



semi-labeled



phylogenetic

- this makes sense for both rooted and undirected trees

Charles Semple, Mike Steel, Phylogenetics, 2003.

Binary phylogenetic trees are universal

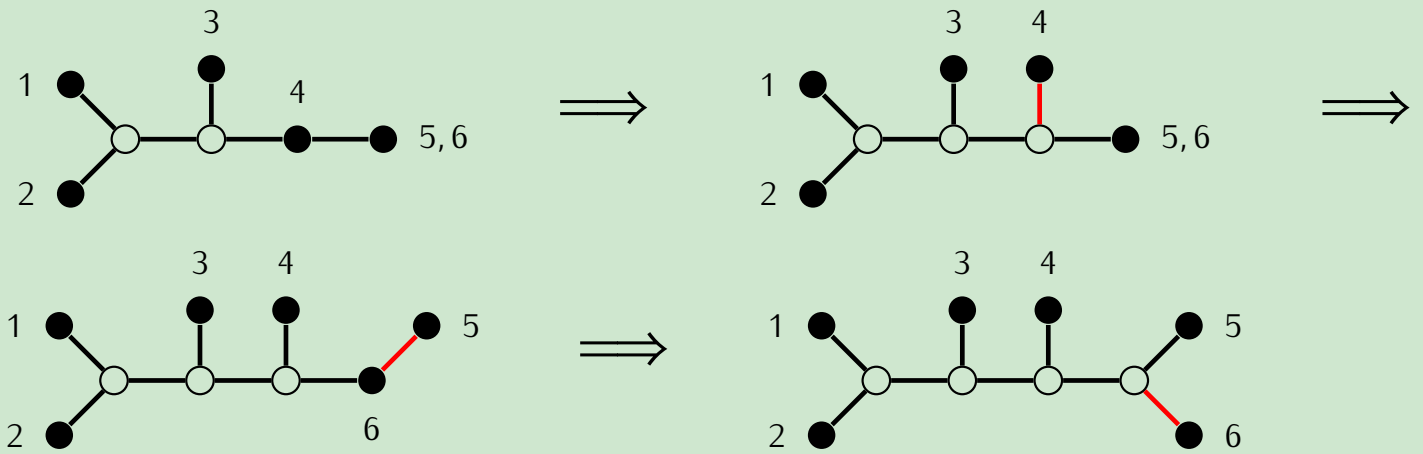
- Undirected binary tree = every inner node has degree three
- Rooted binary tree = every internal node has two children

- Let $e = u - v$ be an edge of a semi-labeled tree $\mathcal{T}$.
- $\mathcal{T}/e$ is the semi-labeled tree obtained from $\mathcal{T}$ by identifying u and v and removing e. The labeling sets of u, v are joined.
  - this operation is called edge contraction

- **Remark:** Every semi-labeled tree can be obtained from a binary phylogenetic tree by edge contractions.

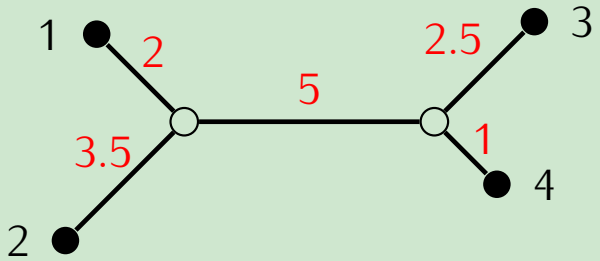
Binary expansion

- A binary expansion of a semi-labeled tree $\mathcal{T}$ is a binary phylogenetic tree $\mathcal{T}^*$ such that $\mathcal{T}$ can be obtained from $\mathcal{T}^*$ by edge contractions. (typically not unique)



Tree metrics

- $\mathcal{T}$ semi-labeled tree with labeling set $[m] := \{1, \ldots, m\}$
- Attach a positive number d_e to each edge e of $\mathcal{T}$
- For every two labeled nodes $i, j \in [m]$
  - $\overline{ij}$ denotes the path between i and j in $\mathcal{T}$
  - $d_{ij} := \sum_{e \in \overline{ij}} d_e$ is the $\mathcal{T}$-distance between i and j in $\mathcal{T}$



$$\begin{bmatrix} 0 & 5.5 & 9.5 & 8 \\ \cdot & 0 & 11 & 9.5 \\ \cdot & \cdot & 0 & 3.5 \\ \cdot & \cdot & \cdot & 0 \end{bmatrix}$$

Tree metrics (2)

- $\mathcal{T}$ a semi-labeled tree with labeling set $[m]$.
- $D = [d_{ij}] \in \mathbb{R}^{m \times m}$ a symmetric matrix with zeros on the diagonal.
- **Definition:** D is a $\mathcal{T}$-metric if there exists a collection of edge lengths d_e of $\mathcal{T}$ such that $d_{ij} = \sum_{e \in \overline{ij}} d_e$ for all $i, j \in [m]$.
- **Definition:** D is a tree metric if it is a $\mathcal{T}$-metric for some semi-labeled tree $\mathcal{T}$.

- Question: Given a symmetric matrix D with $d_{ii} = 0$ and $d_{ij} > 0$ for $i \neq j$, can we say if it is a tree metric? If yes, can we identify the underlying tree $\mathcal{T}$ and the edge lengths d_e?

Tree metric theorem

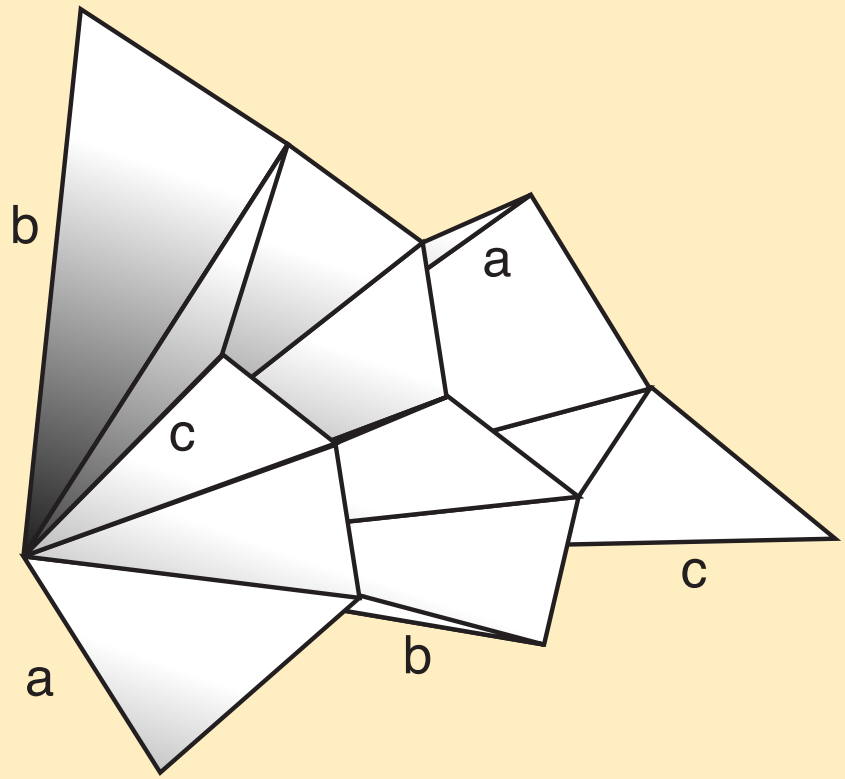
- Theorem[Buneman,1974]: A symmetric matrix $D = [d_{ij}]$ with $d_{ii} = 0$ is a tree metric if and only if for any four (not necessarily distinct) $i, j, k, l \in [m]$

$$d_{ij} + d_{kl} \quad \leq \quad \max \begin{cases} d_{ik} + d_{jl} \\ d_{il} + d_{jk}. \end{cases}$$

  Moreover, a tree metric defines the defining $\mathcal{T}$ and the edge lengths d_e uniquely.

- Every tree metric is a metric $\equiv$ satisfies the triangle inequality.

The space of tree metrics



Billera, L. J., Holmes, S. P., & Vogtmann, K. (2001). Geometry of the Space of Phylogenetic Trees. Advances in Applied Mathematics, 27(4).

Phylogenetic oranges

- $\mathcal{T}$ a semi-labeled tree with labeling set $[m] = \{1, \ldots, m\}$
- Attach a number $\rho_e \in [0, 1]$ to each edge of $\mathcal{T}$.
- For every two labeled nodes $i, j \in [m]$, $\rho_{ij} := \prod_{e \in \overline{ij}} \rho_e$.
- Write $\Sigma = [\rho_{ij}] \in \mathrm{PO}(\mathcal{T})$, $\rho_{ii} = 1$.
  - That Σ is positive semidefinite will be shown later.

- $\mathrm{PO}(m) := \bigcup_{\mathcal{T} \text{ semi-labeled}} \mathrm{PO}(\mathcal{T})$

Moulton, Steel, *Peeling phylogenetic oranges*, 2004.
Kim, *Slicing hyperdimensional oranges: the geometry of phylogenetic estimation*, 2000.
Engström, Hersh, and Sturmfels, *Toric cubes*, 2012.

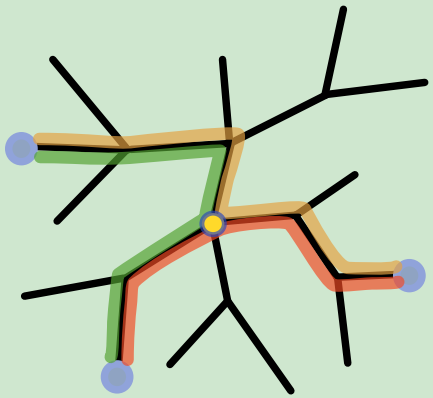
Relation to tree metrics

- Note: all $\rho_e \neq 0$ if and only if all $\rho_{ij} \neq 0$
  - $\mathrm{PO}_{>}(m) := \mathrm{PO}(m) \cap (0,1]^{\binom{m}{2}}$
- **Proposition**: Points in $\mathrm{PO}_{>}(m)$ are in one-to-one correspondence with tree metrics over $[m]$.
  - define $d_{ij} := -\log \rho_{ij}$, $d_e := -\log \rho_e$, then
  - $d_{ij}, d_e \geq 0$ and $d_{ij} = \sum_{e \in \overline{ij}} d_e$ (because $\rho_{ij} = \prod_{e \in \overline{ij}} \rho_e$)

- The space of phylogenetic oranges arises naturally for various statistical models on trees, which we will see later.
- Tree metrics are well studied and many authors exploit this link to propose efficient learning algorithms.

Semi-labeled forests

- If some $\rho_{ij} = 0$, then Σ does not map to a tree metric.
  - if $\rho_{ij} \to 0$, then $-\log \rho_{ij} \to \infty$
- $\rho_{ij} = \prod_{e \in \overline{ij}} \rho_e$ and so $\rho_{ij} = 0$ if and only if $\rho_e = 0$ for some $e \in \overline{ij}$.

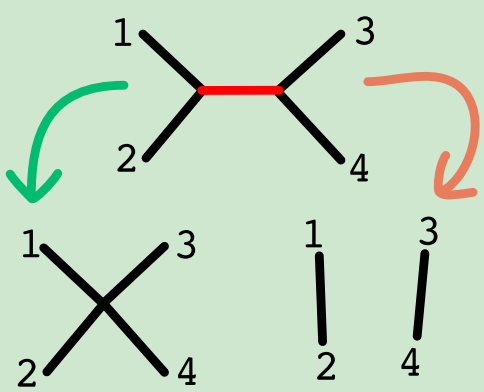


- if $\rho_{ij} \neq 0$ and $\rho_{jk} \neq 0$ then $\rho_{ik} \neq 0$ and so
  - $i \sim j$ iff $\rho_{ij} \neq 0$ defines an equivalence relation
- Every equivalence relation on $[m]$ gives a partition $B_1/\cdots/B_r$ of $[m]$ into equivalence classes (blocks).

- A semi-labeled forest $\mathcal{F}$ with labeling set $[m]$ is a collection of semi-labeled trees with labeling sets $B_1, \ldots, B_r$ that are disjoint and $\bigcup B_i = [m]$.

Tuffley poset

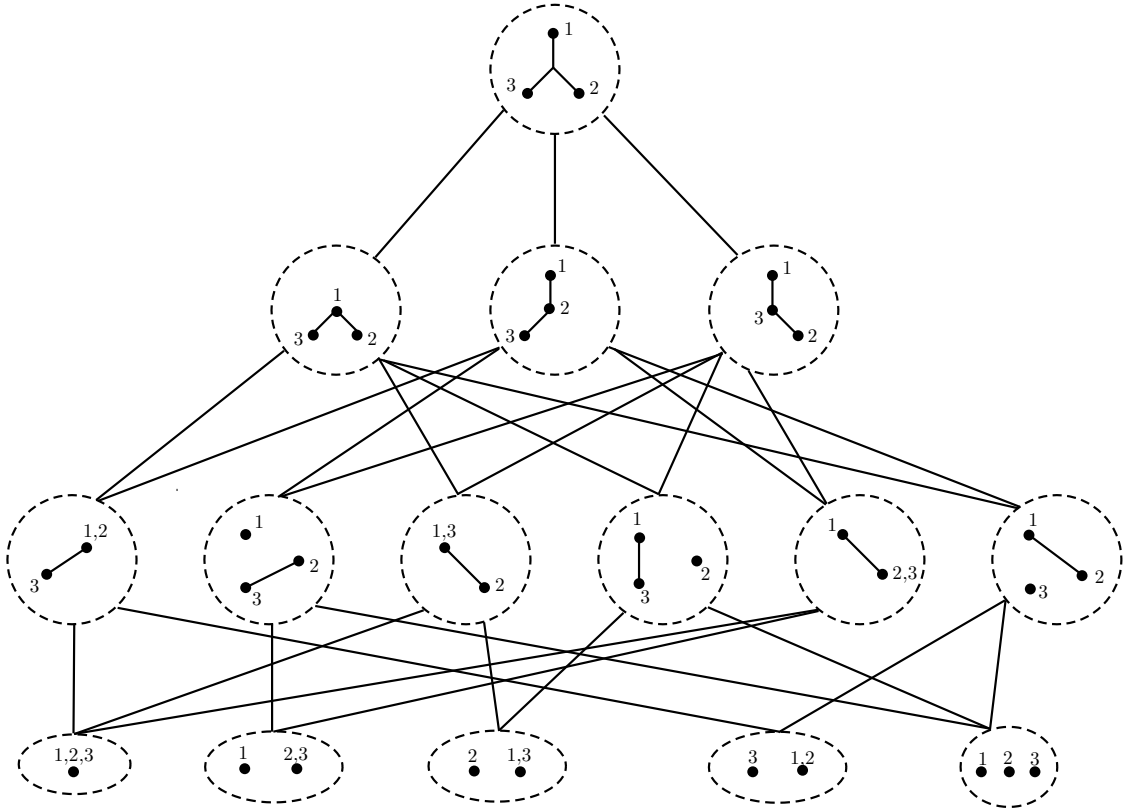
- Consider all semi-labeled forests on $[m]$.
- They form a partially ordered set, called the Tuffley poset.



- If $\mathcal{F}$ is a semi-labeled forest then $\mathcal{F}/e$ is a semi-labeled forest obtained from $\mathcal{F}$ by contracting e
- If $\mathcal{F}$ is a semi-labeled forest then $\mathcal{F} \setminus e$ is a semi-labeled forest obtained from $\mathcal{F}$ by removing e (some post-processing is needed)

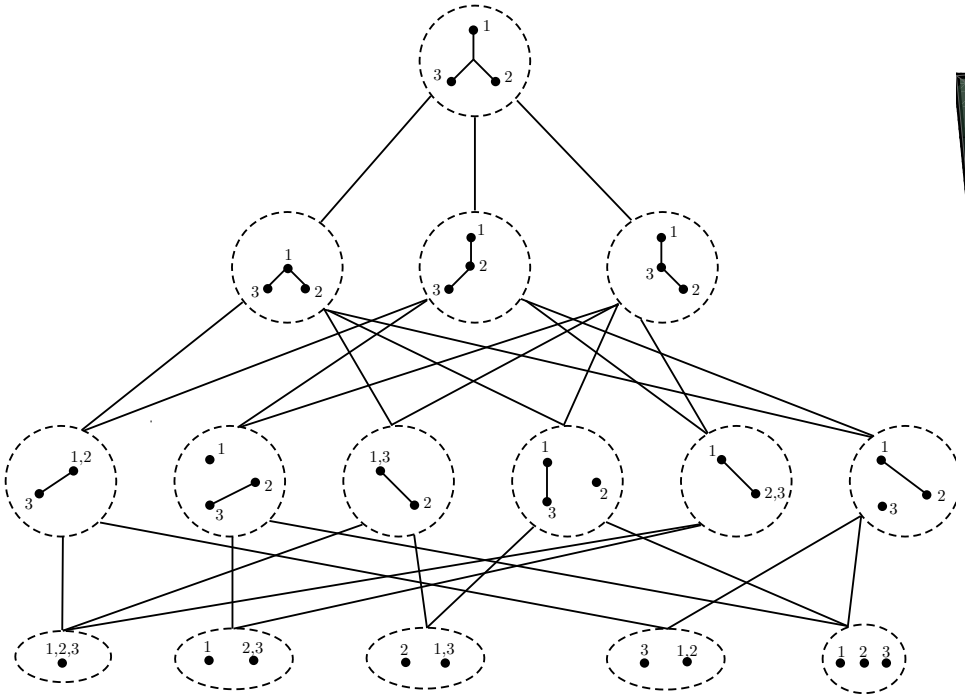
- We say that $\mathcal{T} \leq \mathcal{T}'$ in the Tuffley poset if $\mathcal{T}$ can be obtained from $\mathcal{T}'$ by edge contractions and edge deletions

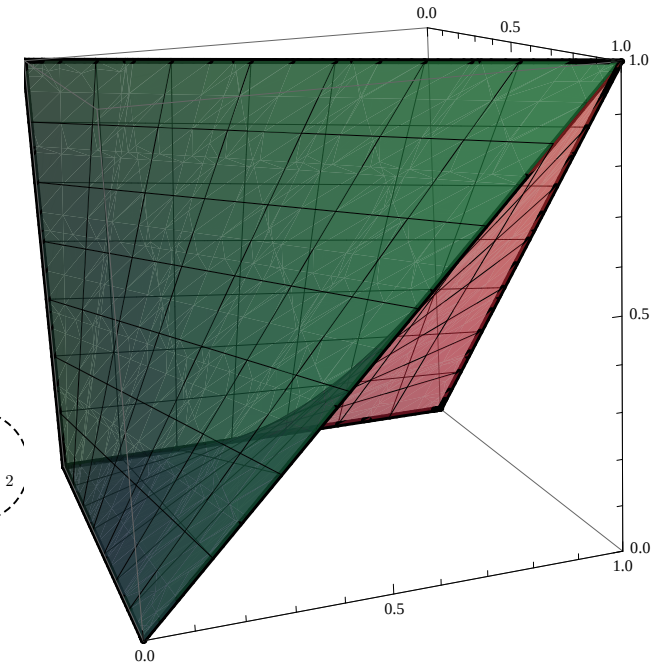
Tuffley poset for $m = 3$



Tuffley poset and the face structure

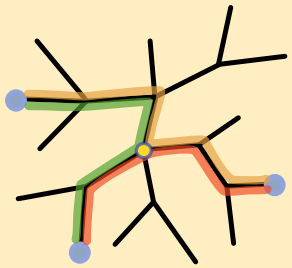
- Contracting an edge corresponds to $\rho_e = 1$. Deleting an edge corresponds to $\rho_e = 0$.
- The Tuffley poset describes the face structure of the boundary of PO(m). Each element corresponds to a strata.





Tree correlations

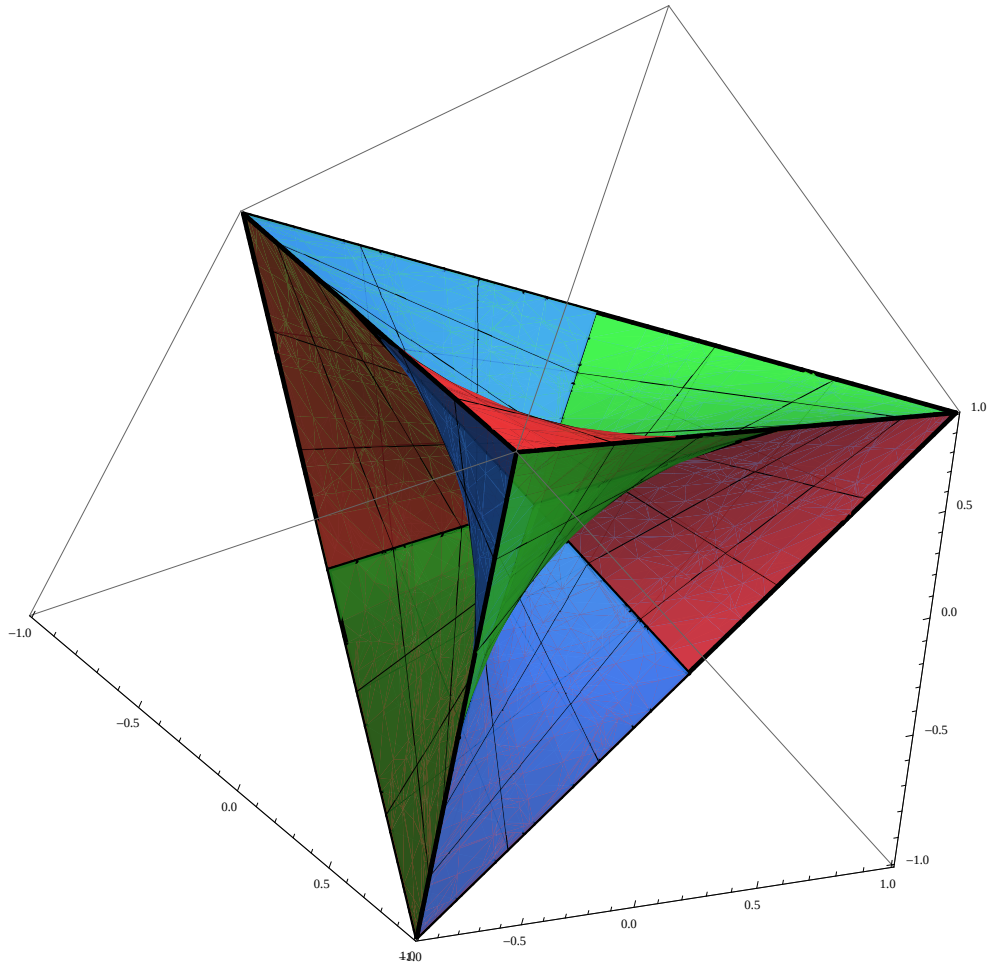
- In many contexts it will be more natural to assume that the edge correlations can be negative, $\rho_e \in [-1, 1]$.
- Call this space the space of tree correlations, $\mathcal{T}$-correlations
- Note that $\rho_{ij}\rho_{ik}\rho_{jk} = \prod_{e\in\overline{ij}} \rho_e \prod_{e\in\overline{ik}} \rho_e \prod_{e\in\overline{jk}} \rho_e$ and so



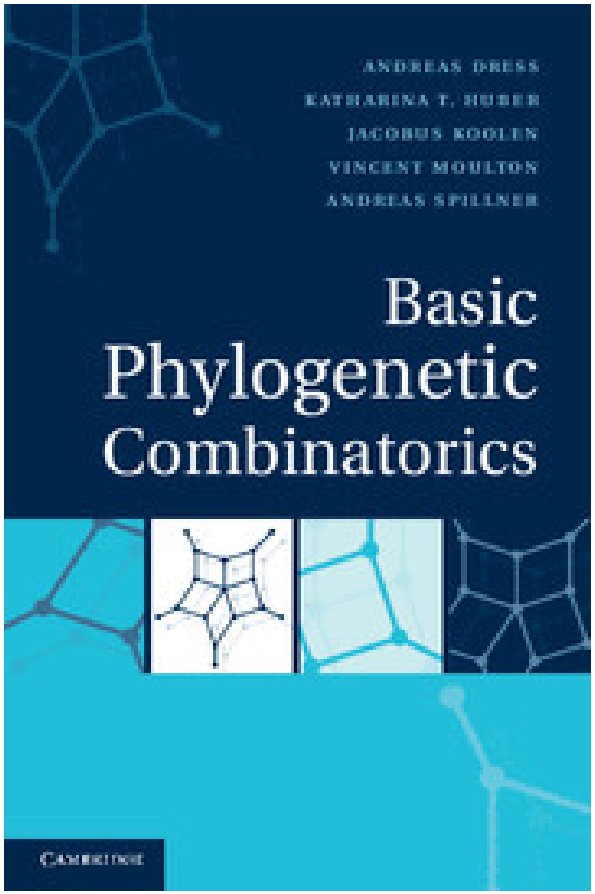
$$\rho_{ij}\rho_{ik}\rho_{jk} = \prod_{e\in\overline{ri}} \rho_e^2 \prod_{e\in\overline{rj}} \rho_e^2 \prod_{e\in\overline{rk}} \rho_e^2 \geq 0.$$

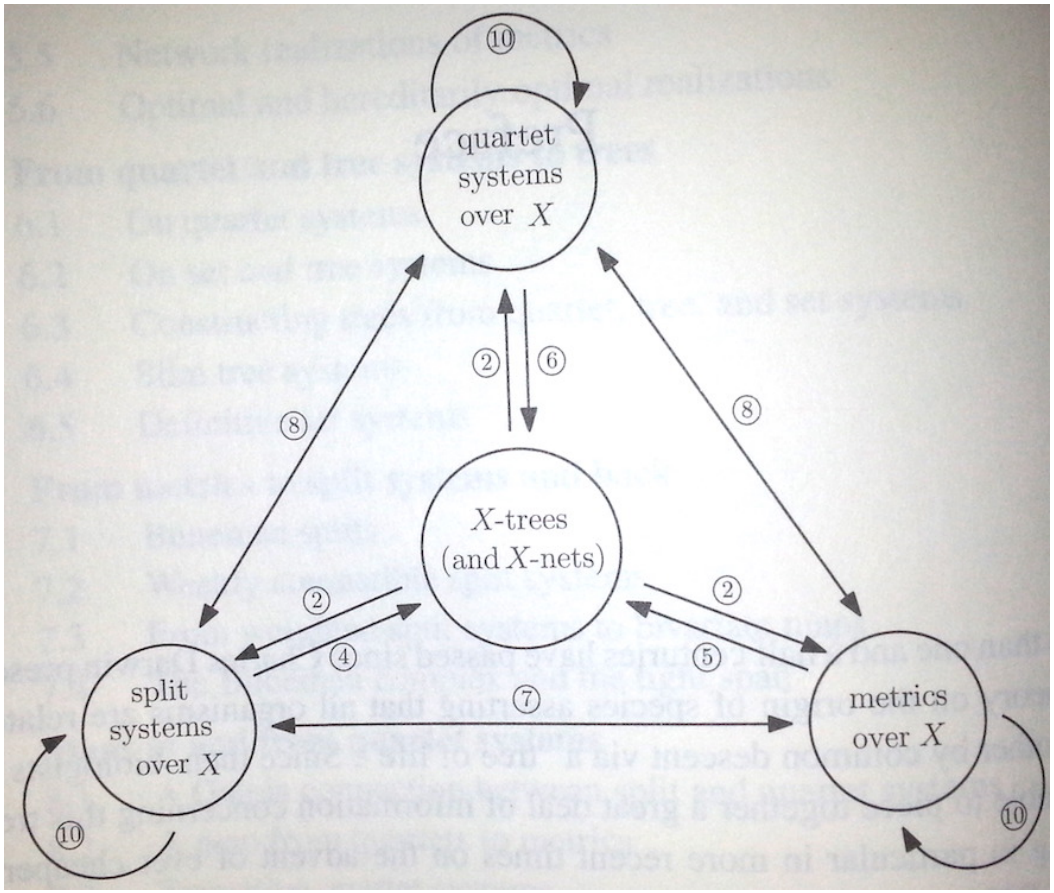
- **Proposition**: A correlation matrix $\Sigma = [\rho_{ij}]$ lies in the space of tree correlations if and only if:
  - (i) $[|\rho_{ij}|]$ lies in the space of phylogenetic oranges $\mathrm{PO}(m)$
  - (ii) for all i, j, k we have $\rho_{ij}\rho_{ik}\rho_{jk} \geq 0$

Tree correlations for three leaves



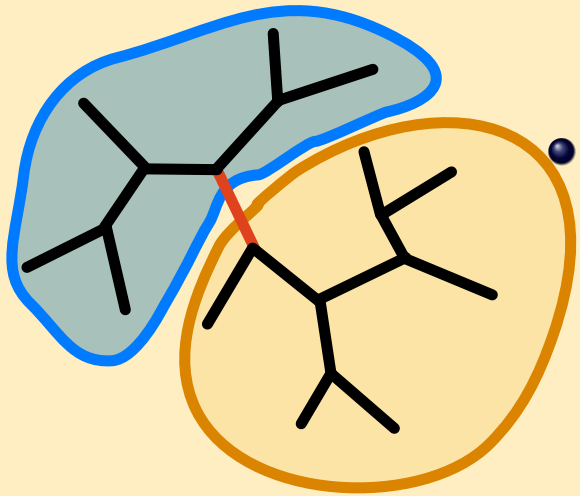
Alternative descriptions of semi-labeled trees





Split systems

- $[m] = \{1, \ldots, m\}$ = the labeling set of the semi-labeled tree $\mathcal{T}$
- Let A/B be a split of $[m]$, i.e. $A \cup B = [m]$, $A \cap B = \emptyset$

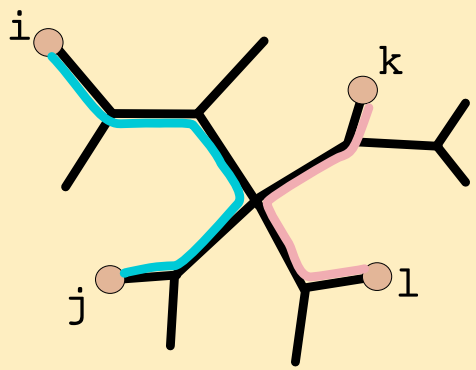


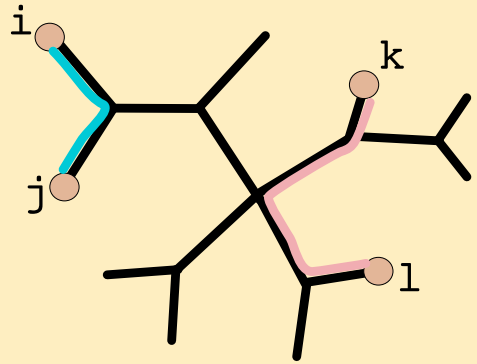
- We say that A/B is a $\mathcal{T}-$*split* if A/B is induced after removing an edge from $\mathcal{T}$ and considering the two connected components of so obtained forest.

- Let Π be the set of all $\mathcal{T}$-splits. Then Π identifies $\mathcal{T}$ uniquely.

Quartet systems

- Let $\mathcal{T}$ be a semi-labeled tree and i, j, k, l any four distinct labeled nodes.
- We say that ij/kl is a quartet of $\mathcal{T}$ if the paths $\overline{ij}$ and $\overline{kl}$ have no vertex in common.





- Let Q be the set of quartets if $\mathcal{T}$. Then Q identifies $\mathcal{T}$ uniquely.

Lecture 2: Latent tree graphical models

Piotr Zwiernik

University of Genoa

Algebraic Statistics 2015
Genova
June 11, 2015

UNIVERSITÀ DEGLI STUDI
DI GENOVA

Short overview

- Lecture 1: Trees, tree metrics and tree spaces
- **Lecture 2: Latent tree graphical models**
- Lecture 3: Tree inference and parameter estimation
- Lecture 4: Likelihood geometry and model identifiability

Graphical models formalism

- graph $G = (V, E)$; V vertex set, E edge set.
- With each vertex $v \in V$ we associate a random variable Y_v with values in $\mathcal{Y}_v$, $\quad Y = (Y_v)$, $\mathcal{Y} = \prod_{v \in V} \mathcal{Y}_v$.
- Missing edges of G indicate some sort of independence.

- for $A \subset V$ denote $Y_A = (Y_v)_{v \in A}$ and $\mathcal{Y}_A = \prod_{v \in A} \mathcal{Y}_v$

Two important classes of graphical models:

- undirected: $f(y) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(y_C)$ for some nonnegative functions ψ_C
  - $\mathcal{C}$ = set of cliques, Z = normalizing constant
- directed acyclic graphs: $f(y) = \prod_{v \in V} f_{v|\mathrm{pa}(v)}(y_v | y_{\mathrm{pa}(v)})$, $y \in \mathcal{Y}$

Graphical model on trees

- Let $T = (V, E)$ be an undirected tree.
- We consider two situations:
  - $Y = (Y_v)$ is multivariate Gaussian
  - $Y = (Y_v)$ is a finite discrete vector with state space $\mathcal{Y} = \prod_{v \in V} \mathcal{Y}_v$

- Fix $\mathcal{Y}$ and T. An undirected tree model $\mathbf{N}(T, \mathcal{Y})$ is the family of densities of the form

$$f(y) = \frac{1}{Z} \prod_{v \in V} \psi_v(y_v) \prod_{u-v \in E} \psi_{uv}(y_u, y_v) \quad \text{for all } y \in \mathcal{Y}$$

  for some nonnegative functions ψ_v, ψ_{uv}.
- we write $\mathbf{N}(T)$ in the Gaussian case

Some alternative formulations

The density f lies in $\mathcal{N}(T, \mathcal{Y})$ if and only if for disjoint $A, B, C \subset V$

- $Y_A \perp\!\!\!\perp Y_B | Y_C \, [f]$ whenever C separates A and B in T
  - i.e. when every path from A to B crosses C

- Fix a vertex $r \in V$ and consider the rooted version T^r of T with root r. Consider the Bayesian network (DAG model) on T^r

$$f(y) = f_r(y_r) \prod_{v \in V \setminus r} f_{v|\mathrm{pa}(v)}(y_v | y_{\mathrm{pa}(v)}) \quad \text{for all } y \in \mathcal{Y},$$

  where $\mathrm{pa}(v)$ is the unique parent of v.

- **Proposition**: Every choice of r leads to the same family of densities. This family is equal to $\mathcal{N}(T, \mathcal{Y})$.

Steffen Lauritzen, Graphical models, 1996.

Model parametrization: discrete case

- We parametrize $\mathbf{N}(T, \mathcal{Y})$ by rooting T at r and specifying the root distribution $\theta_r(y_r)$ together with conditional probabilities $\theta_{v|\mathrm{pa}(v)}(y_v|y_{\mathrm{pa}(v)})$ for all $v \in V \setminus \{r\}$.

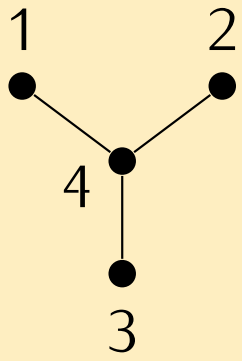
$$f(y;\theta) = \theta_r(y_r) \prod_{v \in V\setminus\{r\}} \theta_{v|\mathrm{pa}(v)}(y_v|y_{\mathrm{pa}(v)}).$$

- probability simplex: $\Delta_k = \{x \in \mathbb{R}^k : x_i \geq 0, \sum_i x_i = 1\}$
- the root distribution lies in $\Delta_{|\mathcal{Y}_r|}$
- for $u \to v$ and every $y \in \mathcal{Y}_u$ we have $\theta_{v|u}(\cdot|y) \in \Delta_{|\mathcal{Y}_v|}$
- the parameter space $\Theta = \Delta_{|\mathcal{Y}_r|} \times \prod_{v \in V\setminus r} (\Delta_{|\mathcal{Y}_v|})^{|\mathcal{Y}_{\mathrm{pa}(v)}|}$

Markov process on T^r

- If all state spaces $\mathcal{Y}_v$ are equal then $\boldsymbol{N}(T, \mathcal{Y})$ is called a Markov process on T^r and denoted by $\boldsymbol{N}(T, d)$, where $d := |\mathcal{Y}_v|$.
- In this case the conditional probabilities $\theta_{v|\mathrm{pa}(v)} \in \mathbb{R}^{d \times d}$ are called transition matrices .
- We can think about this model as a generalization of a Markov chain.

Example: tripod tree model



- $Y \in \{0,1\}^4$, $\theta_4 \in \Delta_2$, $\theta_{1|4}, \theta_{2|4}, \theta_{3|4} \in (\Delta_2)^2$
- $\dim \Theta = 7$
- e.g. $\theta_{1|4} = \begin{bmatrix} \theta_{1|4}(0|0) & \theta_{1|4}(1|0) \\ \theta_{1|4}(0|1) & \theta_{1|4}(1|1) \end{bmatrix}$
- $p(y_1, y_2, y_3, y_4) = \theta_4(y_4)\theta_{1|4}(y_1|y_4)\theta_{2|4}(y_2|y_4)\theta_{3|4}(y_3|y_4)$ for all $(y_1, y_2, y_3, y_4) \in \{0,1\}^4$

- By the separation criterion $1 \perp\!\!\!\perp \{2,3\}|4$ and $2 \perp\!\!\!\perp 3|4$ in $\boldsymbol{N}(T, 2)$ and thus $1 \perp\!\!\!\perp 2 \perp\!\!\!\perp 3|4$.

The Gaussian case: standard definitions

- In the standard language of Gaussian graphical models: $\mathbf{N}(T)$ is the set of all concentration matrices $K = \Sigma^{-1}$ such that $K_{uv} = 0$ whenever u, v are not neighbors in T.
- The dimension of the model is $|V| + |E|$.

- Alternatively, we can describe the model using linear structural equations.
- Let $(\epsilon_v)_{v \in V}$ be independent $\epsilon_v \sim \mathcal{N}(0, \sigma_v)$
- Let $Y_r = \epsilon_r$, and suppose that

$$Y_v = \lambda_v Y_{\mathrm{pa}(v)} + \epsilon_v \quad \text{for all } v \in V \setminus \{r\}$$

ans some (λ_v), then the distribution of Y lies in $\mathbf{N}(T)$.

Alternative parametrization: edge correlations

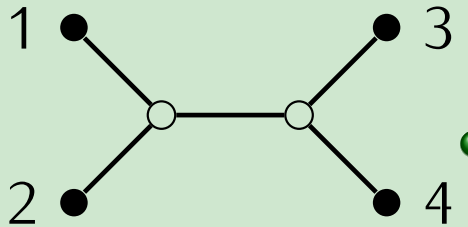
- Suppose Y is jointly Gaussian. We have $Y_u \perp\!\!\!\perp Y_v | Y_w$ if and only if $\rho_{uv} = \rho_{uw}\rho_{wv}$, where $\rho_{uv} = \mathrm{corr}(Y_u, Y_v)$.
- In $\mathbf{N}(T)$ we have $Y_u \perp\!\!\!\perp Y_v | Y_w$ whenever w lies on the path $\overline{uv}$
- Using this recursively we get:

$$(*) \qquad \rho_{uv} = \prod_{e \in \overline{uv}} \rho_e \quad \text{for all } u, v \in V.$$

- new parameters: edge correlations $\rho_e \in [-1, 1]$ for $e \in E$ and variances σ_{vv} for $v \in V$

Latent tree graphical model $M(\mathcal{T}, \mathcal{Y})$

- Let $\mathcal{T}$ be a semi-labeled tree with the underlying tree T and labeling set $[m]$
- $Y = (X, H)$, $\mathcal{Y} = \mathcal{X} \times \mathcal{H}$
  - observed (labeled) subvector of Y: $X \in \mathcal{X}$
  - hidden (unlabeled) subvector of Y: $H \in \mathcal{H}$
- **Definition**: Fix $\mathcal{Y}$ and $\mathcal{T}$. The corresponding latent tree graphical model $M(\mathcal{T}, \mathcal{Y})$ is the set of margins of the densities in $N(T, \mathcal{Y})$ over the labeled nodes of $\mathcal{T}$.



- Consider a distribution $p \in N(T, \mathcal{Y})$ over a quartet tree (6 nodes).
- Summing over all possible values of the two inner nodes gives a distribution in $M(\mathcal{T}, \mathcal{Y})$, where $\mathcal{T}$ is the semi-labeled tree on the left.

Parametrization of $M(\mathcal{T}, \mathcal{Y})$

- In the discrete case the parametrization becomes:

$$p(x; \theta, \mathcal{T}) = \sum_{v \text{ unlabeled}} \sum_{h_v \in \mathcal{Y}_v} p((x,h); \theta, T),$$

where $y = (x, h)$ and

$$p(y; \theta, T) = \theta_r(y_r) \prod_{u \to v} \theta_{v|u}(y_v|y_u) \quad \text{for } y = (y_v)_{v \in V} \in \mathcal{Y}.$$

- In the Gaussian case take simply the corresponding submatrix of the covariance matrix. If $Y \sim \mathcal{N}_{|V|}(0, \Sigma)$ then $X \sim \mathcal{N}_m(0, \Sigma_{XX})$.
  - $\rho_{ij} = \prod_{e \in \overline{ij}} \rho_e$ for all $i, j \in [m]$, variances σ_{ii} unconstrained
  - σ_{vv} for unlabeled v does not appear; assume $\sigma_{vv} = 1$

On the definition of semi-labeled trees

- In our definition of semi-labeled trees we assumed that all nodes of degree ≤ 2 are necessarily labeled.

- If v is a degree one unlabeled node then the formula for $p(x;\theta,\mathcal{T})$ contains: $\sum_{h_v} \theta_{v|\mathrm{pa}(v)}(h_v|y_{\mathrm{pa}(v)}) = 1$ so we can remove v from T without affecting the margin $M(\mathcal{T},\mathcal{Y})$.
- If v is a degree two unlabeled node, then (w.l.o.g.) $u \to v \to w$ is an induced subgraph of T, and the formula for $p(x;\theta,\mathcal{T})$ contains: $\sum_{h_v} \theta_{v|u}(h_v|y_u)\theta_{w|v}(y_w|h_v) = \tilde{\theta}_{w|u}(y_w|y_u)$ so we can suppress v from T without affecting the margin $M(\mathcal{T},\mathcal{Y})$.

- There is a finite number of semi-labeled trees on $[m]$.

Latent forest models

- Let $\mathcal{F}$ be a semi-labeled forest whose tree components are $\mathcal{T}_1, \ldots, \mathcal{T}_k$ with labeling sets $B_1, \ldots, B_k$, $\bigcup B_i = [m]$.
- The latent tree models can be extended to forests. Every density in $\boldsymbol{M}(\mathcal{F}, \mathcal{Y})$ is of the form

$$p(x; \theta, \mathcal{F}) = \prod_{i=1}^{k} p(x_B; \theta, \mathcal{T}_i),$$

  where $p(x_B; \theta, \mathcal{T}_i)$ is a density $\boldsymbol{M}(\mathcal{T}_i, \mathcal{Y}_i)$.
- In particular $X_{B_1} \perp\!\!\!\perp \cdots \perp\!\!\!\perp X_{B_k}$

General Markov model

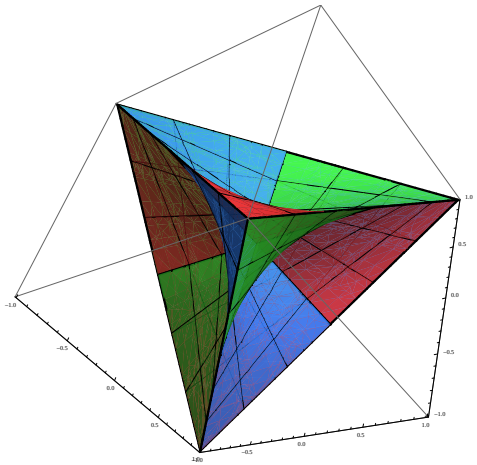
- We focus on two cases
  - the Gaussian case
  - general Markov model: where all $\mathcal{Y}_v$ are equal.

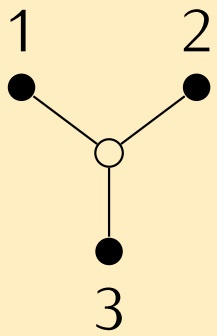
- Write $\boldsymbol{M}(\mathcal{T}, d)$, where $d = |\mathcal{Y}_v|$.
- The matrix of the conditional distribution $\theta_{v|u}$ for the edge $e = u \to v$ is denoted by θ_e and is called a transition matrix.

- The case $d = 4$ is of some interest.

Link to tree correlations

- **Theorem**: The Gaussian latent tree model on a phylogenetic tree $\mathcal{T}$ is equal to the space of tree correlations on $\mathcal{T}$ with $\rho_{ij} \in (-1, 1)$.





- Consider the tripod tree model.
- $Y = (X_1, X_2, X_3, H)$
- $Y \sim \mathcal{N}_4(0, \Sigma)$, $\Sigma \in \mathbf{N}(T)$

- $\rho_{12} = \rho_{h1}\rho_{h2}$, $\rho_{13} = \rho_{h1}\rho_{h3}$, $\rho_{23} = \rho_{h2}\rho_{h3}$ and $\rho_{h1}, \rho_{h2}, \rho_{h3} \in [-1, 1]$.

Edge contraction and removal

- Let $\mathcal{T}$ be a semi-labeled tree and $M(\mathcal{T}, d)$ the corresponding general Markov model.
  - $\mathcal{T}/e$ = the semi-labeled tree with the edge e contracted.
  - $\mathcal{T} \setminus e$ = the semi-labeled forest with e removed.

- Fix an edge $e = u \to v$ and consider the image of all parameters satisfying $\theta_{v|u} = \mathbb{I}_d$. This submodel is equal to $M(\mathcal{T}/e, d)$.
- Fix an edge $e = u \to v$ and consider the image of all parameters satisfying $\text{rank}(\theta_{v|u}) = 1$. This submodel is equal to $M(\mathcal{T} \setminus e, d)$.

- In the Gaussian case the same is obtained by taking $\rho_e = \pm 1$ (contraction) and $\rho_e = 0$ (deletion).

Reduction to binary phylogenetic tree

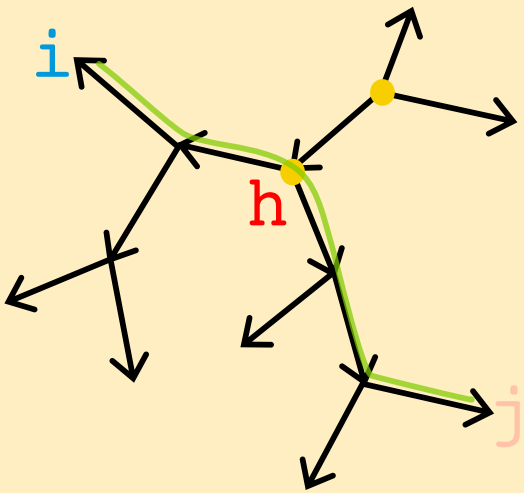
- Recall: a tree is called binary if every inner edge has degree 3

- A binary expansion of $\mathcal{T}$, is any binary phylogenetic tree $\mathcal{T}^*$ such that $\mathcal{T}$ is obtained from $\mathcal{T}^*$ by contracting some edges.

- Using the same argument as on the previous slide, we can show the following result:
- **Proposition**: If $\mathcal{T}^*$ is a binary expansion of a semi-labeled tree $\mathcal{T}$ then $\boldsymbol{M}(\mathcal{T}, \mathcal{Y}) \subseteq \boldsymbol{M}(\mathcal{T}^*, \mathcal{Y})$.
- The same holds in the Gaussian case.

Two-way margins

- Let $M(\mathcal{T}, d)$ be a general Markov model on $\mathcal{T}$ parametrized by the root distribution and the transition matrices θ_e.
- For any distribution in $M(\mathcal{T}, d)$ and any two labels i, j we have



$$\operatorname{diag}(p_i) = \operatorname{diag}(p_h) \prod_{e \in \overline{hi}} \theta_e, \quad \text{and}$$

$$p_{ij} = \prod_{e \in \overline{hi}} \theta_e^T \operatorname{diag}(p_h) \prod_{e \in \overline{hj}} \theta_e.$$

- In particular $\det p_{ij} = \prod_{e \in \overline{ij}} \det \theta_e \prod_{k=1}^{d} p_h(k)$

Link to phylogenetic oranges

- Define $u_{ij} := \frac{\det p_{ij}}{\sqrt{\det(\mathrm{diag}(p_i))\det(\mathrm{diag}(p_j))}} = \frac{\prod_{e\in\overline{ij}} \det\theta_e}{\sqrt{|\prod_{e\in\overline{ij}} \det\theta_e|}}$
- Then for $p \in M(\mathcal{T}, d)$

$$|u_{ij}| = \prod_{e\in\overline{ij}} \sqrt{|\det\theta_e|}.$$

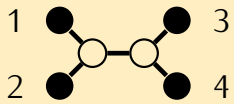
- Since θ_e is a stochastic matrix, $\det\theta_e \in [-1, 1]$.
- Note $\sqrt{|\det\theta_e|} \in [0, 1]$ and so $(|u_{ij}|)$ lies in the space of phylogenetic oranges

Link to tree correlations

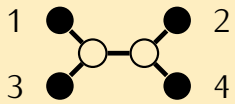
- check: $u_{ij}u_{ik}u_{jk} \geq 0$ for all $i, j, k \in [m]$
- **Proposition**: The space of all possible $u = (u_{ij})$ is equal to the space of all tree correlations.
  - Proof: use a proposition from the previous lecture

- θ_e is a stochastic matrix, $\det \theta_e \in [-1, 1]$ and it is equal to ± 1 if and only if it is a permutation matrix.
- It follows that $|u_{ij}| \in [0, 1]$ and $u_{ij} = \pm 1$ only if X_i and X_j are functionally related
- If $d = 2$ (binary variables), then $u_{ij} = \mathrm{corr}(X_i, X_j)$, so $\rho_{ij} = \prod_{e \in \overline{ij}} \rho_e$ like in the Gaussian case.

Induced constraints



$u_{13}u_{24} = u_{14}u_{23}$;



$u_{12}u_{34} = u_{14}u_{23}$

- In general if ij/kl is a quartet of $\mathcal{T}$ then: $u_{ik}u_{jl} = u_{il}u_{jk}$.

- **Corollary**: We can identify the underlying tree from two-way margins only.

- More on tree inference in the next lecture.

Lecture 3: Tree inference and estimation

Piotr Zwiernik

University of Genoa

Algebraic Statistics 2015
Genova
June 11, 2015

UNIVERSITÀ DEGLI STUDI
DI GENOVA

Short overview

- Lecture 1: Trees, tree metrics and tree spaces
- Lecture 2: Latent tree graphical models
- **Lecture 3: Tree inference and parameter estimation**
- Lecture 4: Likelihood geometry and model identifiability

Three main inference problems

There are three main inference problem for $M(\mathcal{T}, \mathcal{Y})$:

- Learn the underlying tree $\mathcal{T}$.
- Learn the underlying parameter θ.
- Given an estimator $\hat{\theta}$, compute various marginal probabilities from (the fully observed distribution) $p(y; T, \hat{\theta})$.
  - Here we use the fact that $N(T, \mathcal{Y})$ and $M(\mathcal{T}, \mathcal{Y})$ share parameters.

- Depending on the application, some problems are irrelevant.

Tree models as exponential families

- the Gaussian tree model $\mathcal{N}(T)$ forms an exponential family
- In the discrete case the set of strictly positive densities in $\mathcal{N}(T, \mathcal{Y})$ forms a linear exponential family
  - in the factorization $f = \frac{1}{Z} \prod_{u-v \in E} \psi_{uv}$ all $\psi_{uv} > 0$.

- there is a closed form formula for the density at $\hat{\theta}$:

$$f(y; \hat{\theta}) = \frac{\prod_{u-v \in E} \hat{p}_{uv}(y_u, y_v)}{\prod_{v \in V} \hat{p}_v(y_v)^{\deg(v)-1}},$$

where $\deg(v)$ is the degree of v in the underlying tree T.

Why is it useful?

- By standard results on exponential families:
  - the likelihood function is strictly concave
  - conjugate duality between the cumulant function and the entropy for exponential families
- This allows to unify various known learning algorithms.

- If the sample sufficient statistic has no zeros then the MLE is guaranteed not to lie on the boundary and so we may maximize the likelihood function over the corresponding exponential family.

Wainwright, Jordan, Graphical Models, Exponential Families, and Variational Inference. 2007.

Chow-Liu algorithm

- **Problem**: Suppose that we want to find the MLE over the set of all tree models $\mathcal{N}(T, \mathcal{Y})$ for all possible trees with a fixed set of vertices.

- Mutual information $I(Y_i, Y_j)$ as the Kullback-Leibler divergence between f_{ij} and the product $f_i f_j$

$$I_f(Y_i, Y_j) = \sum_{y_i, y_j} f_{ij}(y_i, y_j) \log \frac{f_{ij}(y_i, y_j)}{f_i(y_i) f_j(y_j)}.$$

- $I_f(Y_i, Y_j) \geq 0$ and is zero precisely when Y_i and Y_j are independent.

Chow-Liu algorithm (2)

- For a fixed tree T:

$$f(y;\hat{\theta}) \quad = \quad \prod_v \hat{p}_v(y_v) \prod_{u-v\in E(T)} \frac{\hat{p}_{uv}(y_u,y_v)}{\hat{p}_u(y_u)\hat{p}_v(y_v)}.$$

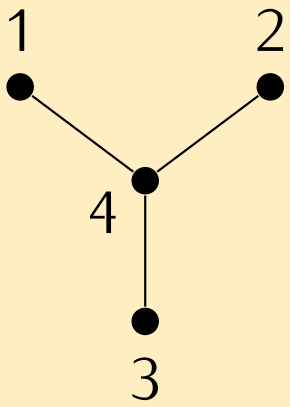
- the log-likelihood at $\hat{\theta}$ ($n\sum_y \hat{p}(y)\log f(y,\hat{\theta})$) can be rewritten as

$$n\sum_v \sum_{y_v} \hat{p}_v(y_v)\log \hat{p}_v(y_v) + n\sum_{u-v\in E(T)} I_{\hat{p}}(Y_u,Y_v).$$

- **Theorem**: The maximum likelihood tree is the maximum cost spanning tree (use Kruskal's theorem)

- the same is true in the Gaussian case
  - here also: $I_{\hat{f}}(Y_u,Y_v) = -\frac{1}{2}\log(1-\hat{\rho}_{uv}^2)$, $\hat{\rho}_{uv}$ = sample correlation

Example: Star tree



- Fixing parameter values, $\theta_4(1) = 0.6$ and $\theta_{i|4}$

$$\begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}, \quad \begin{bmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \end{bmatrix}, \quad \begin{bmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{bmatrix}$$

  we obtain the data generating distribution.

- a simulated matrix of observed mutual informations:

$$\begin{bmatrix} \cdot & 0.000 & 0.003 & 0.043 \\ \cdot & \cdot & 0.004 & 0.027 \\ \cdot & \cdot & \cdot & 0.045 \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix}.$$

- Algorithm: First add edges $3 - 4$ and $1 - 4$. Then, $2 - 4$. Since no more edges can be added without introducing cycles, we stop.

Structural EM: basic idea

- We want to find the maximum likelihood estimator over the union of latent tree models $M(\mathcal{T}, \mathcal{Y})$ for all semi-labeled trees.
- We can assume $\mathcal{T}$ are binary phylogenetic trees.
  - If in our application we are interested in more general phylogenetic trees, this can be further refined.

- If we observed all vertices, the Chow-Liu algorithm gives an efficient way to proceed.
- We use the same idea as in the EM algorithm.

Structural EM for Gaussian models

- Initialize: Choose a starting binary tree topology T^0 and edge correlations $\rho^0 = (\rho_e^0)$. Then, until a convergence criterion is satisfied, perform the two following steps for $i = 0, 1, \ldots$.
- E-step: Compute expected sample covariance of (X, H) given the parameters T^i, ρ^i and the observed vector X.
- M-step: Use the Chow-Liu algorithm to update both the tree and edge weights.

- This works subject to some technicalities...

Friedman et al., A Structural EM Algorithm for Phylogenetic Inference, Journal of Computational Biology, 2002.

The E-step

- The E-step is standard. We work with data of length n normalized to have mean zero.
- Suppose that Σ represents the full covariance matrix estimated at the previous step of the algorithm.
- Let S be the sample correlation matrix that we are trying to estimate: $S_{XX} = \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}, \quad S_{HX} = \frac{1}{n}\boldsymbol{H}^T\boldsymbol{X}, \quad S_{HH} = \frac{1}{n}\boldsymbol{H}^T\boldsymbol{H},$
- Standard formulas: $\mathbb{E}[H|X] = \Sigma_{HX}\Sigma_{XX}^{-1}X$ and $\text{var}(H|X) = \Sigma_{HH} - \Sigma_{HX}\Sigma_{XX}^{-1}\Sigma_{XH}$.
- This gives $\mathbb{E}[S_{HX}|\boldsymbol{X}] = \Sigma_{HX}\Sigma_{XX}^{-1}S_{XX}$ and

$$\mathbb{E}[S_{HH}|\boldsymbol{X}] = \Sigma_{HH} - \Sigma_{HX}\Sigma_{XX}^{-1}\Sigma_{XH} + \Sigma_{HX}\Sigma_{XX}^{-1}S_{XX}\Sigma_{XX}^{-1}\Sigma_{XH}.$$

The M-step

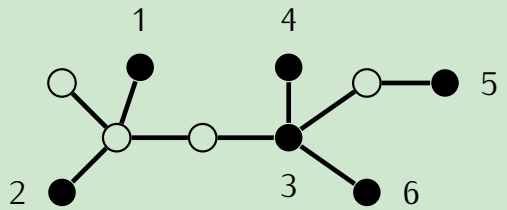
- Here we take the full sample covariance matrix estimated in the E-step and use the Chow-Liu algorithm

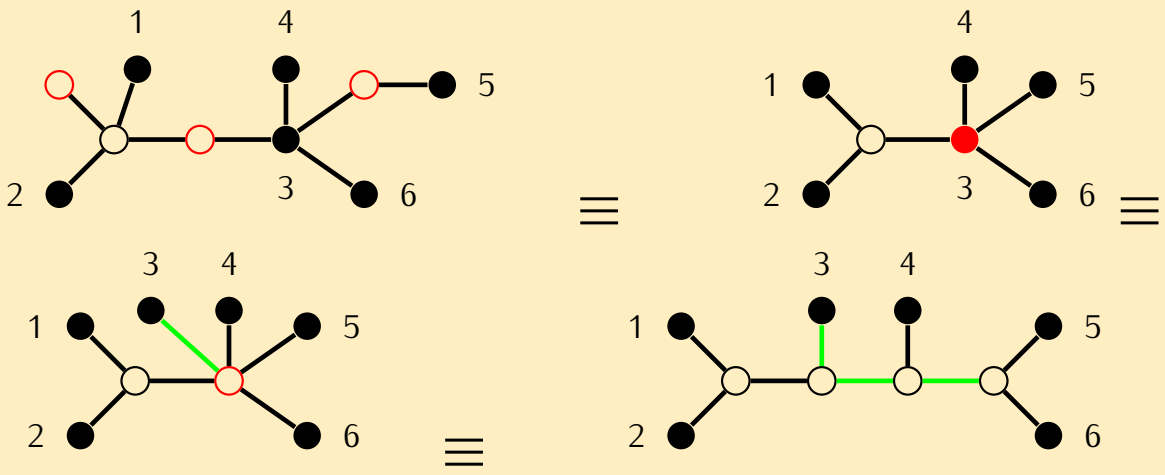
- **Problem**: the Chow-Liu algorithm does not distinguish hidden nodes from observed nodes so it can output a tree with hidden leaves and inner nodes that are observed (in fact it often does in practice).

- **Proposition**: For every tree given as an output of the Chow-Liu algorithm, there exists a binary phylogenetic tree with exactly the same (observed) likelihood.

Example: equal likelihood tree

- If we initialize with a binary phylogenetic tree, then the number of hidden nodes is $m-2$, S is a $(2m-2) \times (2m-2)$ matrix.
- If $m = 6$, then $2m - 2 = 10$.
- Suppose that the M-step reported the tree on the right.

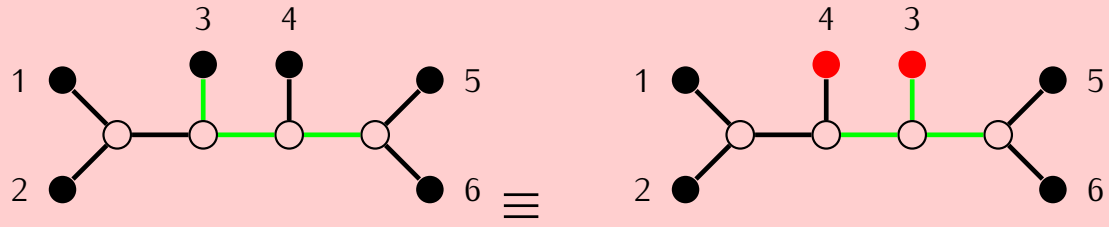




- here — is an edge whose transition matrix is the identity

Equal likelihood tree

- The tree obtained in the previous step is by no means unique.



- we can decide between the two based on some other distance-based argument

- Even a naive implementation works pretty well and very fast for $m \leq 500$.

Tree identifiability

- Suppose that $p \in M(\mathcal{T}, d)$. Recall: for any two leaves $i, j \in [m]$

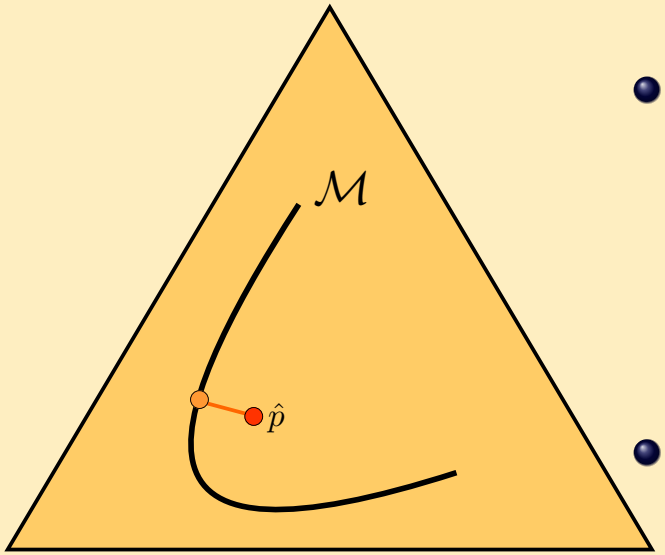
$$u_{ij} := \frac{\det p_{ij}}{\sqrt{\det(\mathrm{diag}(p_i)) \det(\mathrm{diag}(p_i))}}$$

  - $|u_{ij}| = \prod_{e \in \overline{ij}} \rho_e$, where $\rho_e = \sqrt{|\det \theta_e|} \in [0, 1]$.

- If all u_{ij} are nonzero, $d_{ij} := -\log |u_{ij}| > 0$ forms a $\mathcal{T}$-metric
- Buneman: $(\mathcal{T}, (d_e))$ can be uniquely identified from (d_{ij})

- Given some data, the task is to find the best tree
  - From sample proportions $\hat{p}$ compute sample versions of u_{ij} and d_{ij}.
  - Use standard algorithms (least squares, neighbor joining) to learn the best underlying tree.

Phylogenetic invariants

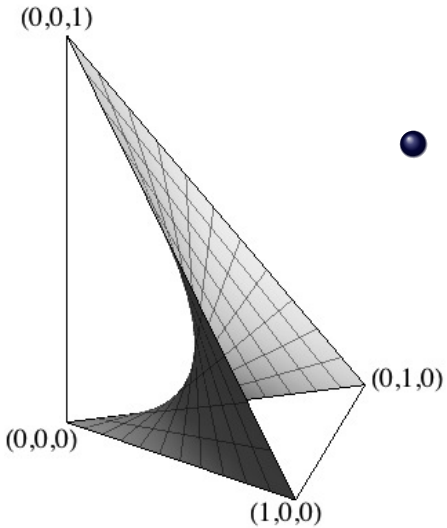


- Another method is the method of phylogenetic invariants that uses some geometric information to choose the best tree model explaining the data.
- We introduce some basic ideas behind this and discuss the method.

Geometric viewpoint

- $X \in \mathcal{X} := \{1, \ldots, k\}$ with distribution $\mathbb{P}(X = i) = p_i$ for $i \in \mathcal{X}$
- probability simplex: $\Delta_k = \{\boldsymbol{p} \in \mathbb{R}^k : p_i \geq 0, \sum_{i=1}^k p_i = 1\}$

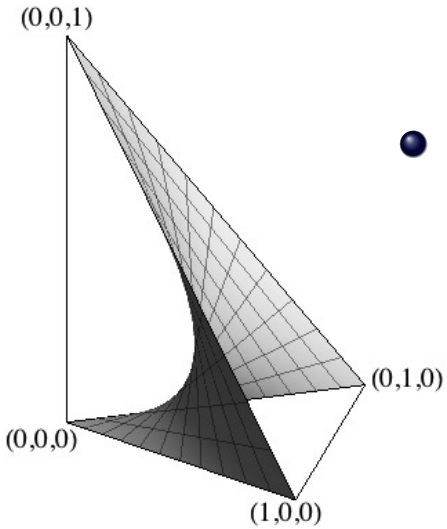
- statistical model on $\mathcal{X}$: a family of probability distributions on $\mathcal{X}$
  - equivalently: a family $\boldsymbol{M}$ of points in Δ_k
- parametric model given as an image of a map $\Theta \to \Delta_k$



- **Example**: Let $X, Y \in \{0, 1\}$. We have $X \perp\!\!\!\perp Y$ if and only if $p_{ij} = p_{i+}p_{+j}$ for all $i, j \in \{0, 1\}$, or equivalently

$$p_{00}p_{11} - p_{10}p_{01} = 0.$$

Phylogenetic invariants: basic idea



- **Example**: Let $X, Y \in \{0,1\}$. We have $X \perp\!\!\!\perp Y$ if and only if $p_{ij} = p_{i+}p_{+j}$ for all $i, j \in \{0,1\}$, or equivalently

$$p_{00}p_{11} - p_{10}p_{01} = 0.$$

- Given a random sample of size n let $\hat{p}$ be the sample proportions.
- If the true data generating distribution q satisfies $X \perp\!\!\!\perp Y[q]$ then for large n we have $\hat{p}_{11}\hat{p}_{00} - \hat{p}_{01}\hat{p}_{10} \approx 0$.
- We can use this fact to test whether $X \perp\!\!\!\perp Y$

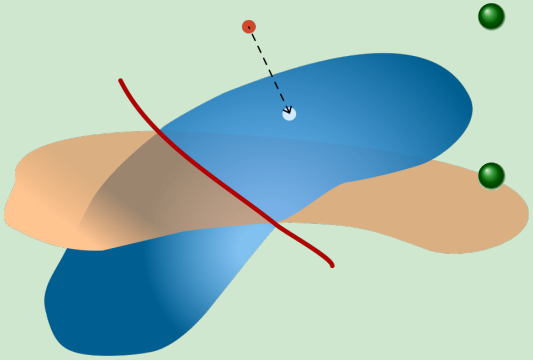
Semialgebraic sets

- A simple semialgebraic set is a subset of $\mathbb{R}^d$ described by polynomial equations and inequalities.
- A semialgebraic set is a subset of $\mathbb{R}^d$ given as a finite union of simple semialgebraic sets.
- **Theorem [Tarski, Seidenberg]**: The image of a semialgebraic set under a polynomial map is semialgebraic.

- $M(\mathcal{T}, \mathcal{Y})$ is given as the image of a polynomial parametrization. The parameter space is a product of simplices and so semialgebraic. It follows that $M(\mathcal{T}, \mathcal{Y}) \subseteq \Delta_{|\mathcal{X}|-1}$ is semialgebraic.

Phylogenetic invariants: application

- The study of defining equations (phylogenetic invariants) was proposed independently by Joseph Felsenstein, James Cavender, and James Lake in 1980's.



- Suppose we have a collection of competing latent tree models.
- We use (some) algebraic constraints defining these models to select the best model.
  - no parameter estimation is needed
  - the method is consistent

- There are several problems with this procedure:
  - there are many invariants and some are very sensitive
  - by ignoring inequalities we lose some information
  - the statistical theory is underdeveloped

Lecture 4: Likelihood geometry and model identifiability

Piotr Zwiernik

University of Genoa

Algebraic Statistics 2015
Genova
June 11, 2015

UNIVERSITÀ DEGLI STUDI
DI GENOVA

Short overview

- Lecture 1: Trees, tree metrics and tree spaces
- Lecture 2: Latent tree graphical models
- Lecture 3: Tree inference and estimation
- **Lecture 4: Likelihood geometry and model identifiability**

The model identifiability

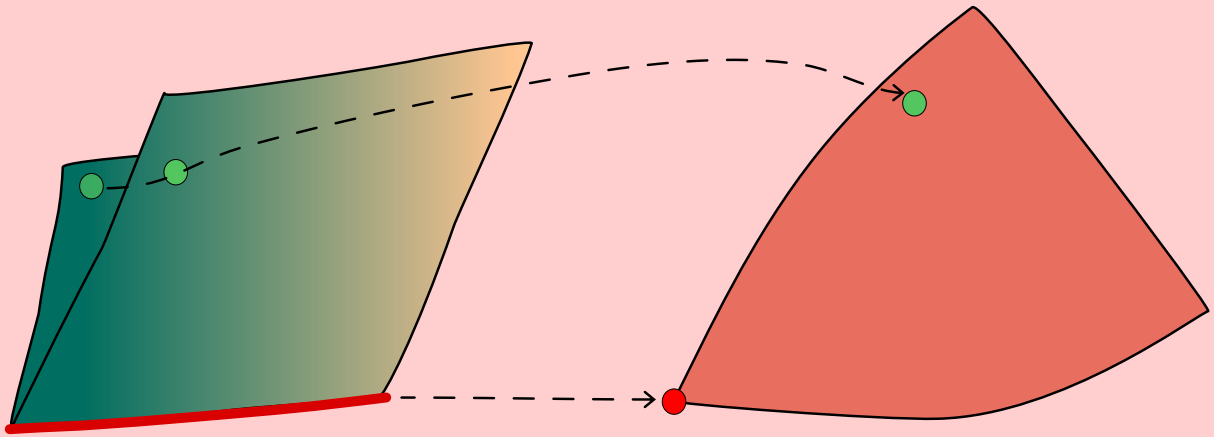
- We say that a parametric model $(P_\theta)_{\theta \in \Theta}$ is identifiable if $P_\theta = P_{\theta'}$ implies $\theta = \theta'$
  - otherwise, even with infinite data, we cannot learn the parameter

- This definition is too restrictive in general for models with hidden variables
  - label swapping problem
  - special parameter values correspond to degenerate cases

Generic model identifiability

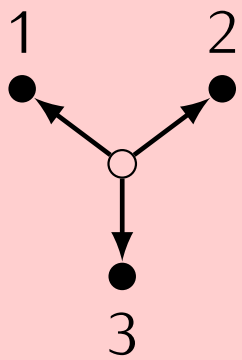
- A parametric model is given by a parametrization $\theta \mapsto p_\theta$.
- Such a model is identifiable if the parametrization is one-to-one.

- **Definition**: We say that a parametric model $(P_\theta)_{\theta \in \Theta}$ is generically identifiable if the parametrization is finite-to-one for almost all distributions in the model.



Simple examples

- Model: $1 \bullet \leftarrow \overset{H}{\circ} \rightarrow \bullet 2$, where X_1, X_2, H binary
- the parameter space has dimension 5, the model dimension is ≤ 3 so there is no identifiability



- Model: $X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp X_3 | H$, where X_1, X_2, X_3 any discrete and H binary
- This model is generically identifiable; the parametrization is generically two-to-one.
  - switch rows of θ_h, $\theta_{1|h}$, $\theta_{2|h}$, $\theta_{3|h}$.
- There is an infinite number of parameter vectors that map to any distribution in the model satisfying $X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp X_3$.

Example: the Gaussian tripod

$\mathcal{T}$ the tripod tree. Suppose that $\Sigma \in M(\mathcal{T})$ with $\rho_{ij} \geq 0$.

- First note that precisely one zero correlation is impossible.

We have three cases:

(i) Correlations non-zero: $\rho_1 := \sqrt{\frac{\rho_{12}\rho_{13}}{\rho_{23}}}$, $\rho_2 := \sqrt{\frac{\rho_{12}\rho_{23}}{\rho_{13}}}$, $\rho_3 := \sqrt{\frac{\rho_{13}\rho_{23}}{\rho_{12}}}$, then $\rho_i\rho_j = \rho_{ij}$ and $\rho_i \in [0,1]$.

(ii) Two correlations are zero: say $\rho_{12} \neq 0$ then $\rho_3 := 0$ and ρ_1, ρ_2 any such that $\rho_1\rho_2 = \rho_{12}$.

(iii) All are zero: three cases, e.g. $\rho_1 = \rho_2 = 0$ and ρ_3 arbitrary.

Kruskal's theorem

- Suppose X_1, X_2, X_3, H discrete with d_1, d_2, d_3, r values.
- Using Kruskal's theorem for 3-way contingency tables the following sufficient condition for generic identifiability can be given:
  - **Theorem**: The tripod model is generically identifiable, provided

$$\min(r, d_1) + \min(r, d_2) + \min(r, d_3) \quad \geq \quad 2r + 2.$$

Identifiability for star trees

- The basic idea is to realize a more general model as a submodel of the tripod tree model.

- **Theorem**(Allman,Matias,Rhodes): Consider the star tree model $M(\mathcal{T}, \mathcal{Y})$ where $|\mathcal{X}_i| = d_i$ and $|\mathcal{H}| = r$. Suppose that there exists a tripartition of the labeling set $[m]$ into three sets A_1, A_2, A_3 such that if $\kappa_i = \prod_{j \in A_i} d_j$ then

$$\min(r, \kappa_1) + \min(r, \kappa_2) + \min(r, \kappa_3) \quad \geq \quad 2r + 2.$$

  Then the model is generically identifiable up to label swapping.

Allman, Matias, Rhodes, Identifiability of Parameters in Latent Structure Models with Many Observed Variables, Annals of Statistics, 2009.

Identifiability for general Markov models

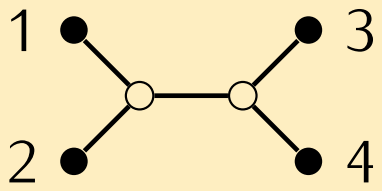
- **Theorem (Chang)**: Let $\mathcal{T}$ be a semi-labeled tree. The corresponding general Markov model $M(\mathcal{T}, d)$ is generically identifiable up to label swapping of the latent variables.
- If $d = 2$, we have explicit formulas for the parameters and we understand all special fibers of the parametrization.

- **Theorem**: The Gaussian latent tree model on a semi-labeled tree $\mathcal{T}$ is generically identifiable up to sign of the latent variables.
- In this case can explicitly give the inverse map from the model to the parameter space.

Chang, *Full Reconstruction of Markov Models on Evolutionary Trees: Identifiability and Consistency*, 1996.

Zwiernik, Smith, *Tree-cumulants and the geometry of binary tree models*, 2012.

Formulas for parameters: Gaussian case



Check that:

- $\rho_0^2 = \frac{\rho_{13}\rho_{24}}{\rho_{12}\rho_{34}} = \frac{\rho_{14}\rho_{23}}{\rho_{12}\rho_{34}}$
- $\rho_1^2 = \frac{\rho_{12}\rho_{13}}{\rho_{23}} = \frac{\rho_{12}\rho_{14}}{\rho_{24}}$

- Suppose $\rho_{12} = 1/6$, $\rho_{13} = 1/60$, $\rho_{14} = 1/90$, $\rho_{23} = 1/40$, $\rho_{24} = 1/60$, $\rho_{34} = 1/24$.
- Then $(\rho_0^2, \rho_1^2, \rho_2^2, \rho_3^2, \rho_4^2) = (1/25, 1/9, 1/4, 1/16, 1/36)$.
- We have four possible solutions $s \cdot (1/5, 1/3, 1/2, 1/4, 1/6)$, where s is one of:

$$\{(+,+,+,+,+), (-,-,-,+,+), (-,+,+,-,-), (+,-,-,-,-,)\}$$

- Identical formulas can be derived for general $M(\mathcal{T})$.

Constrained multinomial likelihood

- Let $\theta \mapsto p_\theta$ be a parametric model $\boldsymbol{M}$ over $\mathcal{X}$, $\boldsymbol{M} \subset \Delta_{\mathcal{X}}$.
- Fix data $u = (u(x))_{x \in \mathcal{X}}$; the likelihood function

$$L(u; \theta) = \prod_{x \in \mathcal{X}} p_\theta(x)^{u(x)}.$$

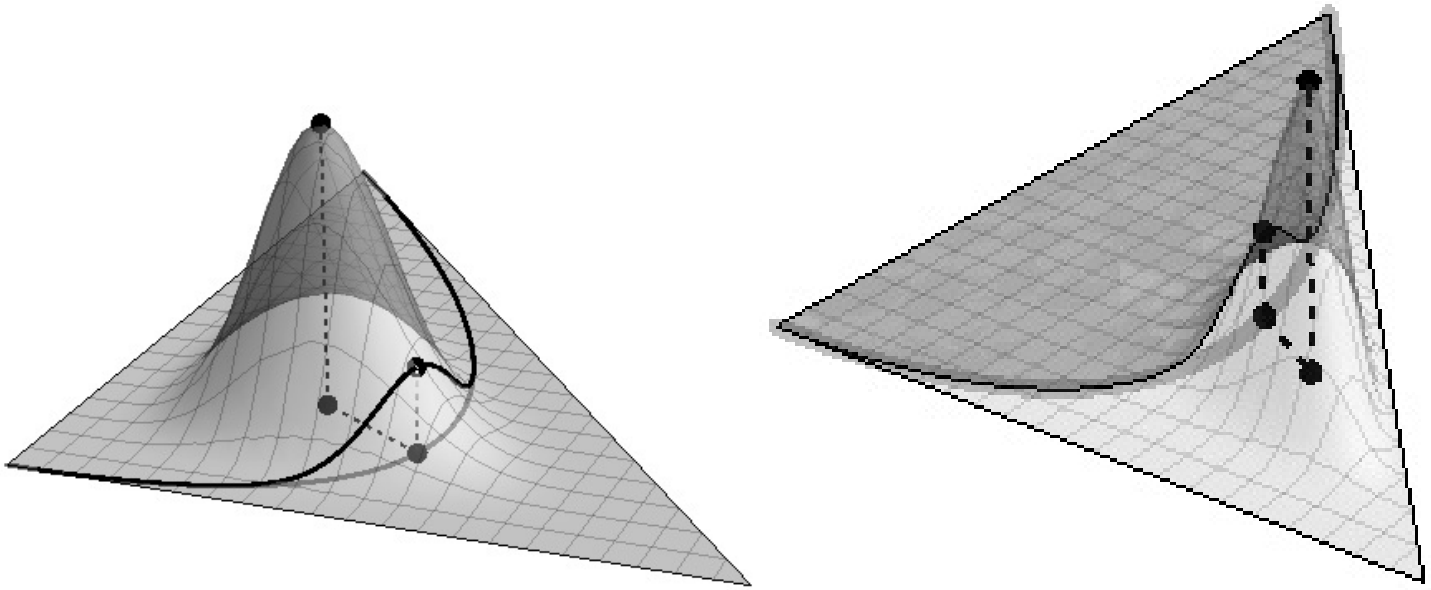
- Multinomial likelihood $L_m(u; p) = \prod_{x \in \mathcal{X}} p(x)^{u(x)}$, $p \in \Delta_{\mathcal{X}}$.

- Instead of maximizing $L(u; \theta)$ we can maximize the multinomial likelihood constrained to $p \in \boldsymbol{M}$.

- This gives a good insight into the likelihood geometry for latent tree models because $L_m(u; p)$ is strictly concave with a unique maximizer $\hat{p}(x) = u(x)/n$ as long as u has only positive entries.

Some examples

- Consider the model Bin$(2, \theta)$ and its mixture.



- In general the situation is much more complicated

The Gaussian tripod tree model

- **Proposition**: A covariance matrix Σ lies in the Gaussian tripod tree model if and only if $K = \Sigma^{-1}$ satisfies $k_{12}k_{13}k_{23} \leq 0$.
- The Gaussian likelihood function is strictly concave when expressed in K.

- Recall: the boundary corresponds to $\overset{i}{\bullet} - \overset{j}{\bullet} - \overset{k}{\bullet}$
- Maximizing the likelihood function over the boundary is straightforward. For example, over $\overset{1}{\bullet} - \overset{2}{\bullet} - \overset{3}{\bullet}$ we have

$$\rho^*_{12} = \hat{\rho}_{12}, \quad \rho^*_{23} = \hat{\rho}_{23}, \quad \rho^*_{13} = \hat{\rho}_{12}\hat{\rho}_{23}.$$

- Maximizing over the interior is also easy
  - Σ^* exists if and only if the sample covariance matrix S lies in the model ($\Sigma^* = S$).

Binary tripod model

- **Theorem**: Let $\mathcal{T}$ be the tripod phylogenetic tree. A distribution p lies in $M(\mathcal{T}, 2)$ if and only if (up to the action of $Z_2 \times Z_2 \times Z_2$)

$$\begin{aligned}
p_{000}p_{111} &\geq p_{001}p_{110} & p_{000}p_{111} &\geq p_{010}p_{101} & p_{000}p_{111} &\geq p_{100}p_{011} \\
p_{001}p_{111} &\geq p_{011}p_{101} & p_{010}p_{111} &\geq p_{011}p_{110} & p_{100}p_{111} &\geq p_{101}p_{110} \\
p_{000}p_{011} &\geq p_{001}p_{010} & p_{000}p_{101} &\geq p_{001}p_{100} & p_{000}p_{110} &\geq p_{010}p_{100}
\end{aligned}$$

- In particular, there are no equations and the model has dimension 7.

- The boundary is described by points where some of these inequalities become equalities. However $p_{\bullet}p_{\bullet} = p_{\bullet}p_{\bullet}$ is a linear equation in $\log p_{\bullet}$ and so the boundary consists of log-linear models.

Allman, Rhodes, Sturmfels, Zwiernik, *Tensors of nonnegative rank two*, 2015.

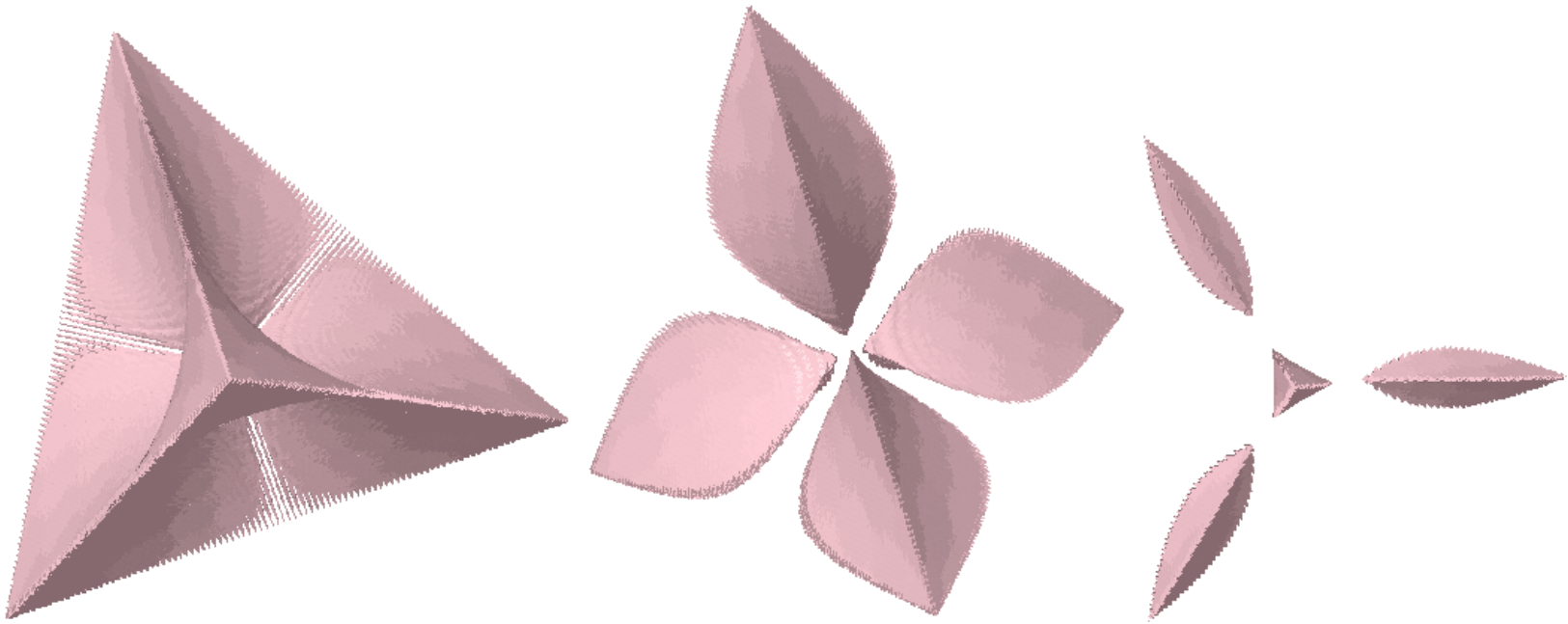
Closed form MLE procedure

- **Theorem**: There is a procedure to get the exact maximum likelihood estimator over the model $M(\mathcal{T}, 2)$, where $\mathcal{T}$ is the phylogenetic tripod tree.

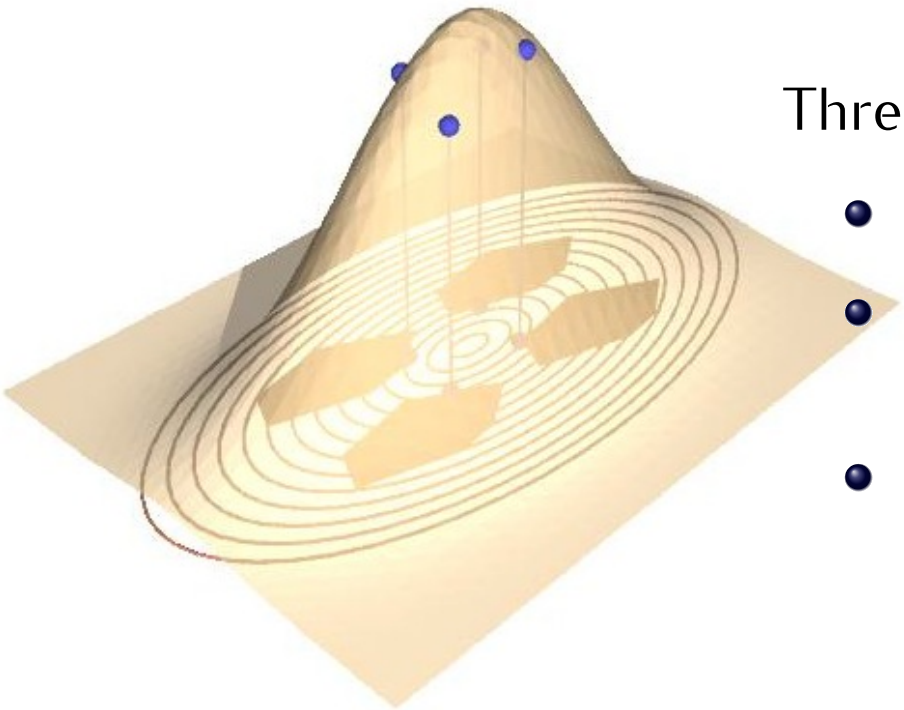
- The maximum over the interior of the model exists if and only if the sample proportions $\hat{p}$ lie in the interior. In this case, the likelihood maximized precisely at $\hat{p}$. Otherwise the maximum lies on the boundary.
- To optimize the likelihood we check smaller dimensional strata.
- In fact, almost all these boundary strata admit a closed form formula for the maximum. The remaining ones require solving a quadratic equations.

Sources of multimodality in the likelihood

- The dimension of the model is 7. Means of the observed nodes are unconstrained, fix all of them to be 1/2.
- We draw three slices of the remaining 4-dimensional set.



Sources of multimodality in the likelihood (2)



Three sources of multimodality:

- Label switching (easy fix).
- each blob can get at least one mode
- blobs are concave, so there may be several modes within a blob

A simple numerical example

- Suppose that a sample of size 10000 has been observed

$$\left[\begin{array}{cc|cc} u_{000} & u_{001} & u_{100} & u_{101} \\ u_{010} & u_{011} & u_{110} & u_{111} \end{array}\right] = \left[\begin{array}{cc|cc} 2069 & 16 & 2242 & 331 \\ 2678 & 863 & 442 & 1359 \end{array}\right].$$

- Use the EM-algortihtm 100 times starting from random parameter values.

- The algorithm found 6 different local maxima

	$\theta_1^{(r)}$	$\theta_{1\|0}^{(1)}$	$\theta_{1\|1}^{(1)}$	$\theta_{1\|0}^{(2)}$	$\theta_{1\|1}^{(2)}$	$\theta_{1\|0}^{(3)}$	$\theta_{1\|1}^{(3)}$
1	0.466	0.337	0.552	1.000	0.000	0.416	0.074
2	0.534	0.552	0.337	0.000	1.000	0.074	0.416
3	0.257	0.361	0.658	0.420	0.865	0.000	1.000
4	0.743	0.658	0.361	0.865	0.420	1.000	0.000
5	0.437	0.000	1.000	0.629	0.412	0.156	0.386
6	0.563	1.000	0.000	0.412	0.629	0.386	0.156

Why this is important

- There may be distant local maxima found by the EM-algorithm with similar value of the likelihood function. This should be part of the whole output of the EM-algorithm.
- Maxima often lie on the boundary of the parameter space
  - Here the usual interpretation of the hidden variable breaks down.
  - This will be a common problem unless variables in the system are highly correlated.
  - Points on the boundary do not correspond to critical points of the likelihood function.
  - A similar problem occurs in the Bayesian framework.

Wang, Zhang, (2006). Severity of Local Maxima for the EM Algorithm.

Zwiernik, Smith, (2011) Implicit inequality constraints in a binary tree model.

Thank you!