Tests for Large Tree-structured Data

Karthik Bharath

School of Mathematical Sciences

University of Nottingham

IN THIS TALK..

Tree: acyclic graph with a ‘root’ which can be embedded on a plane.

In this talk..

Tree: acyclic graph with a ‘root’ which can be embedded on a plane.

- A simple model for binary trees using Poisson process;
- Extend to general classes of large trees based on the Continuum Random Tree;
- Goodness-of-fit tests;
- Application to detecting brain tumor heterogeneity from images (Time permitting).

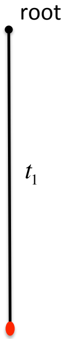
Motivation: model for binary trees

- Consider a non-homogeneous Poisson process with rate $\lambda(t) = t$.
- Let $t_1, t_2, \ldots,$ be inter-event times.

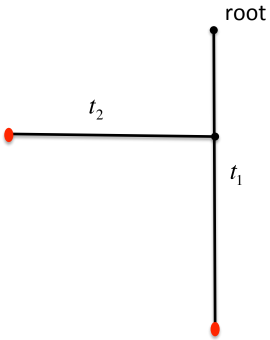
MOTIVATION: MODEL FOR BINARY TREES

- Consider a non-homogeneous Poisson process with rate $\lambda(t) = t$.

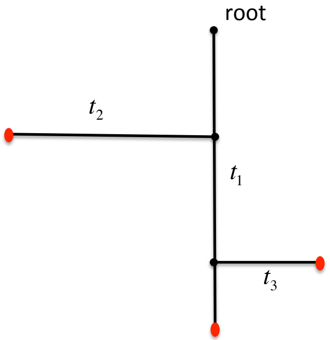
- Let $t_1, t_2, \ldots,$ be inter-event times.



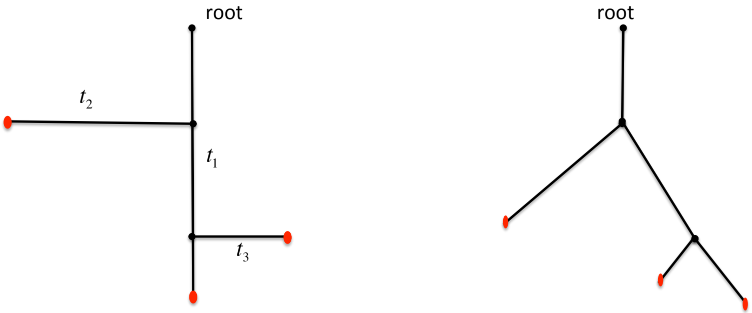
Motivation: model for binary trees



Motivation: model for binary trees



MOTIVATION: MODEL FOR BINARY TREES



$t_1 + t_2 + t_3$ = total path length of binary tree

MOTIVATION: MODEL FOR BINARY TREES

With n inter-event times, a binary tree $\tau(n)$ with n leaves or terminal nodes, $2n$ vertices and $2n - 1$ edges is constructed.

Motivation: model for binary trees

With n inter-event times, a binary tree $\tau(n)$ with n leaves or terminal nodes, $2n$ vertices and $2n-1$ edges is constructed.

Proposition

From the properties of the Poisson process with rate t, $\tau(n)$ can be given the density

$$f(\tau(n)) = \left[\prod_{i=1}^{n-1} \frac{1}{2i-1}\right]^{-1} \frac{1}{2^{n-1}} s e^{-s^2/2}, \quad s = \sum_{i=1}^{2n-1} t_i.$$

- $f(\cdot)$ is exchangeable with respect to the branch lengths.
- $f(\cdot)$is independent of the “shape" of the tree.
- $f(\cdot)$ is ‘consistent’: removal of a leaf from $\tau(n)$ results in a tree a with density $f(\tau(n-1))$.

Proposition

Suppose $\tau(k)$ is a binary tree with k leaves and branch lengths $x_1, \ldots, x_{2k-1}$ generated from the non-homogeneous Poisson model. Then the total path length $\sum_{i=1}^{2k-1} x_i$ follows a Gamma distribution with shape k and scale 2.

GoF test for binary trees: one-sample

Suppose $\boldsymbol{\tau}(\boldsymbol{n}) = (\tau(n_1), \ldots, \tau(n_p))$ is an independent sample of binary trees from π_τ.

Theorem

Consider the critical function

$$\phi(\mathbf{n}, \alpha) = \begin{cases} 1 & \text{if } \sum_{i=1}^{p} s_i > \chi_{1-\alpha, 2\sum_{i=1}^{p} n_i}; \\ 0 & \text{otherwise,} \end{cases}$$

where s_i is the sum of the branch lengths of $\tau(n_i)$ and $\chi_{\alpha,b}$ denotes the αth percentile of a Chi-square distribution with b degrees of freedom. For the hypotheses $H_0 : \pi_\tau = f$ against $H_1 : \pi_\tau \neq f$, where f is the density from the non-homogeneous Poisson model, $E_{H_0}\phi(\mathbf{n}, \alpha) = \alpha$.

GoF test for binary trees: two-sample

Suppose $\boldsymbol{\tau}(\boldsymbol{n}) = (\tau(n_1), \ldots, \tau(n_p))$ and $\boldsymbol{\eta}(\boldsymbol{m}) = (\eta(m_1), \ldots, \eta(m_q))$ are independent samples of binary trees from π_τ and π_η respectively.

Theorem

Let r_j denote the sum of the branch lengths of $\eta(m_j)$, and without loss of generality assume that $\sum_{i=1}^p s_i > \sum_{j=1}^q r_j$. Then, the critical function

$$\psi(\mathbf{n}, \mathbf{m}, \alpha) = \begin{cases} 1 & \text{if } \frac{\sum_{i=1}^p s_i}{\sum_{j=1}^q r_j} > \left(\frac{\sum_{i=1}^p n_i}{\sum_{j=1}^q m_j}\right) F_{1-\alpha, 2\sum_{i=1}^p n_i, 2\sum_{j=1}^q m_j}; \\ 0 & \text{otherwise,} \end{cases}$$

where $F_{\alpha,a,b}$ is the αth percentile of an F distribution with a and b degrees of freedom, for testing $H_0 : \pi_\tau = \pi_\eta = f$, is such that $E_{H_0}\psi(\mathbf{n}, \mathbf{m}, \alpha) = \alpha$.

Non-binary trees

- Can the model and tests for binary trees be extended to other types of trees?

Non-binary trees

- Can the model and tests for binary trees be extended to other types of trees?

Yes, in an asymptotic sense for 'large' trees with some modifications.

Motivation

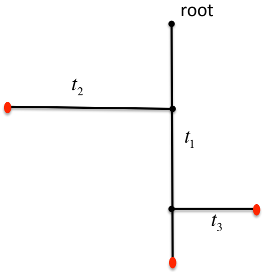
" Given a sequence of discrete combinatorial random structures of size n, does there exists a continuous structure representing their $n \to \infty$ limit? If so, then for many questions about the size-n object one can obtain the $n \to \infty$ limit by simply asking the same question of the limit structure. This is the *weak convergence paradigm*. "

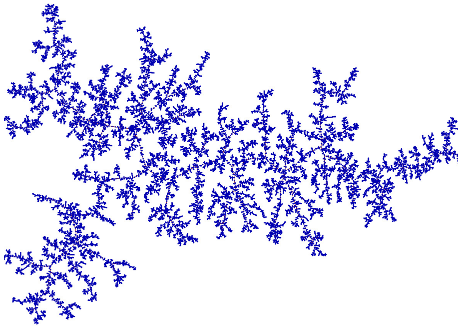
– *David Aldous*.

For random trees, the continuous structure is termed the Continuum Random Tree (CRT).

Continuum Random Tree (CRT)

As $n \to \infty$,





(http://www.normalesup.org/ kortchem/english.html)

What is the CRT?

- It can thought of as the closure of the union of binary trees $\cup_n \tau(n)$.
- From a stochastic process perspective, its ‘finite-dimensional distributions’ are the the binary trees studied earlier!

What is the CRT?

- It can thought of as the closure of the union of binary trees $\cup_n \tau(n)$.
- From a stochastic process perspective, its 'finite-dimensional distributions' are the the binary trees studied earlier!

It is realized as:

- the weak limit, as vertices grow without bound, of Galton–Watson process conditioned on total progeny;
- an object constructed from the dense set of local minima of a standard Brownian excursion process.

General tree models with the CRT

Representation of tree

- We will use a more intuitive representation of a tree τ_n with n vertices and $n-1$ edges:

$$\tau_n = (\mathcal{V}(\tau_n), \mathcal{E}(\tau_n)),$$

  where $\mathcal{V}(\tau_n) = (\text{root}, v_1, \ldots, v_{n-1})$ is the vertex-set and $\mathcal{E}(\tau_n) = (e_1, \ldots, e_{n-1})$ is the edge-set.

- In other words, τ_n is a point in $\mathcal{T}_n \times \mathbb{R}_+^{n-1}$ where $\mathcal{T}_n$ is the set of all combinatorial trees with n vertices.

Conditioned Galton-Watson tree models

- Suppose ξ is non-negative integer-valued r.v. with distribution $(\pi_k : k \geq 0)$.
- Construct a tree τ recursively starting with root and giving each node a number of children that is an independent copy of ξ. This induces a distribution on τ:

$$P(\tau = t) = \prod_{v \in \mathcal{V}(t)} \pi_{o(v,t)},$$

where $o(v, t)$ is the out-degree or the number of children of vertex v in tree t.

Conditioned Galton-Watson tree models

- Suppose ξ is non-negative integer-valued r.v. with distribution $(\pi_k : k \geq 0)$.
- Construct a tree τ recursively starting with root and giving each node a number of children that is an independent copy of ξ. This induces a distribution on τ:

$$P(\tau = t) = \prod_{v \in \mathcal{V}(t)} \pi_{o(v,t)},$$

  where $o(v, t)$ is the out-degree or the number of children of vertex v in tree t.
- Conditioned Galton-Watson (CGW) trees are family trees of Galton-processes conditioned on total progeny. The distribution of a CGW tree τ_n conditioned on n vertices is then

$$P(\tau_n = t) \propto \prod_{v \in \mathcal{V}(t)} \pi_{o(v,t)} \quad \text{on } \{t : \text{ cardinality of } \mathcal{V}(t) = n\}.$$

Conditioned Galton-Watson tree models

- π is referred to as the offspring distribution.
- CGW represent a broad class of trees which can serve as probability models.
  1. *Plane trees* (ordered trees): CGW trees with offspring distribution given by a Geometric distribution with success probability $1/2$, and $\sigma^2 = 2$;
  2. *Binary trees*: CGW trees with vertices containing 0,1 or 2 children with a Binomial distribution with 2 trials and success probability 1/2, and $\sigma^2 = 1/2$;
  3. *Strict binary trees*: CGW trees with vertices containing either 0 or 2 children with equal probability 1/2, and $\sigma^2 = 1$;
  4. *Unary-binary trees*: CGW trees with vertices containing 0, 1 or 2 children each with probability 1/3, and $\sigma^2 = 2/3$;
  5. *m-ary trees*: CGW trees with vertices containing $0, 1, \ldots, m$ for $m > 3$ children with distribution given by a Binomial with m trials and success probability $1/m$, and $\sigma^2 = \frac{m-1}{m}$.

Distribution of CRT: LCA trees

For a CGW tree $\tau_n = (\mathcal{V}(\tau_n), \mathcal{E}(\tau_n))$, consider the Least Common Ancestor (LCA) subtree spanned by a randomly chosen subset B of the leaves, including the root.

Distribution of CRT: LCA subtrees (Aldous 1993)

For a CGW tree $\tau_n = (\mathcal{V}(\tau_n), \mathcal{E}(\tau_n))$, suppose $|B|$ is $k < n$ and denote the LCA tree spanned by B as $LCA(\tau_n, B)$. Then

As $n \to \infty$, for a fixed k, $LCA(\tau_n, B)$ "converges weakly" to a binary tree, $\mathcal{L}(k)$, obtained via the non-homogeneous Poisson process model.

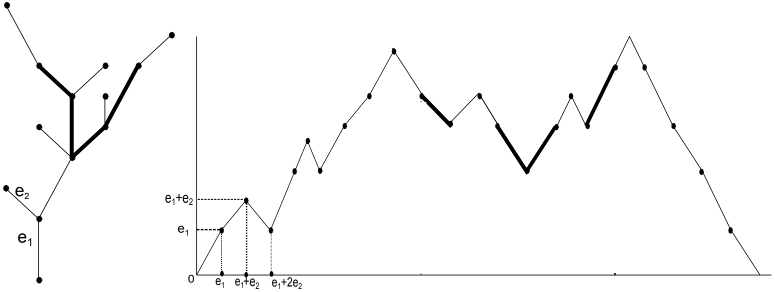
- $\mathcal{L}(k)$ can be viewed as "finite-dimensional projection" of CRT.
- Key point: Distribution of CRT is completely specified by distribution of $\{\mathcal{L}(k), k \geq 1\}$.

DISTRIBUTION OF CRT: DYCK PATH REPRESENTATION

For a CGW tree with n vertices , define a continuous function $H_n : [0, 2n] \to \mathbb{R}_{\geq 0}$ such that

$$H_n(s) = d(\text{root}, v),$$

where v is the vertex obtained during the depth-first walk such that the sum of the edges traversed till v is s.



Aldous' remarkable result

Theorem

Let τ_n be a CGW tree with offspring distribution with mean 1 *and variance $\sigma^2 \in (0, \infty)$. Let $H_n(k), 0 \le k \le 2n$ be the Dyck path associated with τ_n. Then, as $n \to \infty$,*

$$\left\{ \frac{1}{\sqrt{n}} H_n([2nt]), 0 \le t \le 1 \right\} \Rightarrow \left\{ \frac{2}{\sigma} B_t^{ex} : 0 \le t \le 1 \right\}$$

where B^{ex} is the standard Brownian excursion.

Aldous' even more remarkable result

Theorem

The Brownian excursion which arises as the limit of the normalized Dyck path of a conditioned Galton-Watson tree is the "Dyck path" of the CRT.

DISTANCE FROM ROOT OF RANDOMLY CHOSEN VERTEX

PROPOSITION

Let U be uniform on $[0, 1]$. Consider the functional $B^{ex}(U)$. For a fixed σ^2, the class of distributions induced through $U \mapsto \frac{2}{\sigma} B^{ex}(U)$ characterises the law of $\frac{2}{\sigma} B^{ex}$, with $\frac{2}{\sigma} B^{ex}(U)$ following a Rayleigh distribution with scale $1/\sigma$.

Distance from root of randomly chosen vertex

Proposition

Let U be uniform on $[0,1]$. Consider the functional $B^{ex}(U)$. For a fixed σ^2, the class of distributions induced through $U \mapsto \frac{2}{\sigma}B^{ex}(U)$ characterises the law of $\frac{2}{\sigma}B^{ex}$, with $\frac{2}{\sigma}B^{ex}(U)$ following a Rayleigh distribution with scale $1/\sigma$.

Proposition

On an ordered conditioned Galton–Watson tree τ_n with offspring variance σ^2, suppose v is a vertex chosen according to a uniform distribution on $\mathcal{V}(\tau_n)$. Then, the random variable

$$n^{-1/2}d(root, v) \xrightarrow{d} W,$$

as $n \to \infty$, where W is a Rayleigh distributed random variable with scale $1/\sigma$.

GoF tests for CGW trees

LCA-based tests:

- For each tree in a sample, choose a subset of leaves at random, and construct LCA trees—each LCA tree will be a strict binary tree;
- From a non-homogeneous Poisson process with rate $\lambda(t) = \sigma^2 t$, obtain a parametric class of densities seen earlier for binary trees;
- With a consistent estimator for σ^2, the χ^2 and F GoF tests for binary trees are valid, as the number of vertices in each tree grow without bound.

GoF tests for CGW trees

LCA-based tests:

- For each tree in a sample, choose a subset of leaves at random, and construct LCA trees—each LCA tree will be a strict binary tree;
- From a non-homogeneous Poisson process with rate $\lambda(t) = \sigma^2 t$, obtain a parametric class of densities seen earlier for binary trees;
- With a consistent estimator for σ^2, the χ^2 and F GoF tests for binary trees are valid, as the number of vertices in each tree grow without bound.

Dyck path-based tests:

- For each tree in a sample, choose a vertex at random, and record its normalized distance from the root;
- Consider a σ^2-parameterized class of Rayleigh densities, and construct a consistent estimator of σ^2;
- Noting that if X is Rayleigh distributed with scale $1/b$, then X^2 is Chi-square distributed with 2 degrees of freedom scaled by $1/b^2$, we get similar GoF tests.

Performance of LCA-based test

One-sample

Distribution	$n = 10$		$n = 100$		$n = 1000$	
	χ^2	perm	χ^2	perm	χ^2	perm
Geo(0.5)	0.09	0.15	0.05	0.08	0.03	0.09
Bin(0.5)	0.13	0.08	0.04	0.03	0.01	0.01
Bin(0.35)	0.10	0.16	0.12	0.07	0.06	0.08
GW-Bin(2,0.5)	0.78	0.91	0.91	0.97	0.99	1.00
Phylo.bd	0.81	0.83	0.89	0.91	0.98	0.94
Phylo.coal	0.26	0.37	0.14	0.21	0.11	0.08

Application: Detecting tumor heterogeneity with MRI

Joint work with collaborators at M.D. Anderson Cancer Center, Houston.

Tumour heterogeneity

- Variety of genetic, cellular and molecular mutations occur during the course of tumour development or during the course of a treatment.
- Classification of tumours using clustering of pixel intensities is popular.
- Typically histograms are compared with few parameters.

Tumour heterogeneity

- Variety of genetic, cellular and molecular mutations occur during the course of tumour development or during the course of a treatment.
- Classification of tumours using clustering of pixel intensities is popular.
- Typically histograms are compared with few parameters.

Our approach: Capture heterogeneity through variations in pixel intensities via hierarchical relationships between groups of pixels.

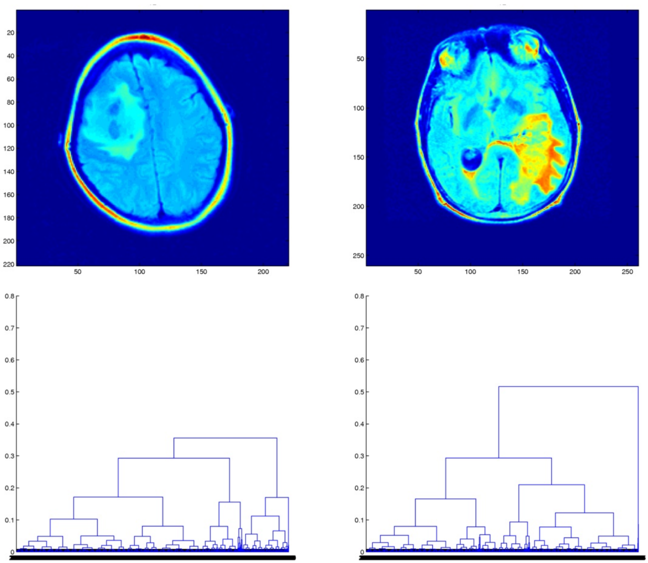
Dataset

- Images of 82 patients with histologically confirmed GBM and molecular data from The Cancer Genome Atlas (TCGA) database (`https://www.cancerimagingarchive.net/`).
- T1-post and T2-FLAIR images were registered spatially followed by intensity bias correction using Medical Image Processing Analysis and Visualization software (v 6.0).13.
- The tumor region was segmented semi-automatically in 3D using the Medical Image Interaction Toolkit (`MITK.org`) with in-plane resolution of 1mm $\times$ 1mm.

Dataset

- Images of 82 patients with histologically confirmed GBM and molecular data from The Cancer Genome Atlas (TCGA) database (`https://www.cancerimagingarchive.net/`).
- T1-post and T2-FLAIR images were registered spatially followed by intensity bias correction using Medical Image Processing Analysis and Visualization software (v 6.0).13.
- The tumor region was segmented semi-automatically in 3D using the Medical Image Interaction Toolkit (`MITK.org`) with in-plane resolution of 1mm $\times$ 1mm.
- T1 and T2 intensities from the segmented regions were grouped with agglomerative hierarchical clustering to obtain dendrograms of image intensities.

DATASET



20
40
60
80
100
120
140
160
180
200
220
50
100
150
200
50
100
150
200
250
50
100
150
200
250
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

DENDROGRAMS FROM HIERARCHICAL CLUSTERING

These dendrograms are ultrametric trees, and do not correspond to CGW trees. In fact, they can be generated by a coalescent process used in phylogenetics (recall poor power against phylo.coal).

Dendrograms from Hierarchical Clustering

These dendrograms are ultrametric trees, and do not correspond to CGW trees. In fact, they can be generated by a coalescent process used in phylogenetics (recall poor power against phylo.coal).

However, there is a surprising connection: if the leaves of an ultrametric binary tree, $\tau(n)$ with n leaves, arising from a hierarchical clustering method are exchangeable in the sense that the distribution of the the leaves is invariant to permutation, then the LCA-trees converge in distribution, as $n \to \infty$, to the family $\mathcal{L}(k)$ of subtrees which characterise the CRT. (Ph.D. dissertation of Chris Haulk [2012])

Two-sample test to detect heterogeneity

- Using the survival times, we created two groups of patients: those with survival times of utmost 12 months and those exceeding 12 months.
- The 12-month cut-off corresponded to a certain genetic classification— this was based on recommendations by neuroscientists.
- Differences in groups was detected by LCA-based test at 1% significance level.
- Naive Bayes classifier with the likelihood from LCA trees, provided 69% classification accuracy.

Details available in a paper on `Arxiv`: *Statistical Tests for Large Tree-structured Data*.

Details available in a paper on `Arxiv`: *Statistical Tests for Large Tree-structured Data*.

“If you can’t convince them, confuse them."

—*Harry S. Truman*.