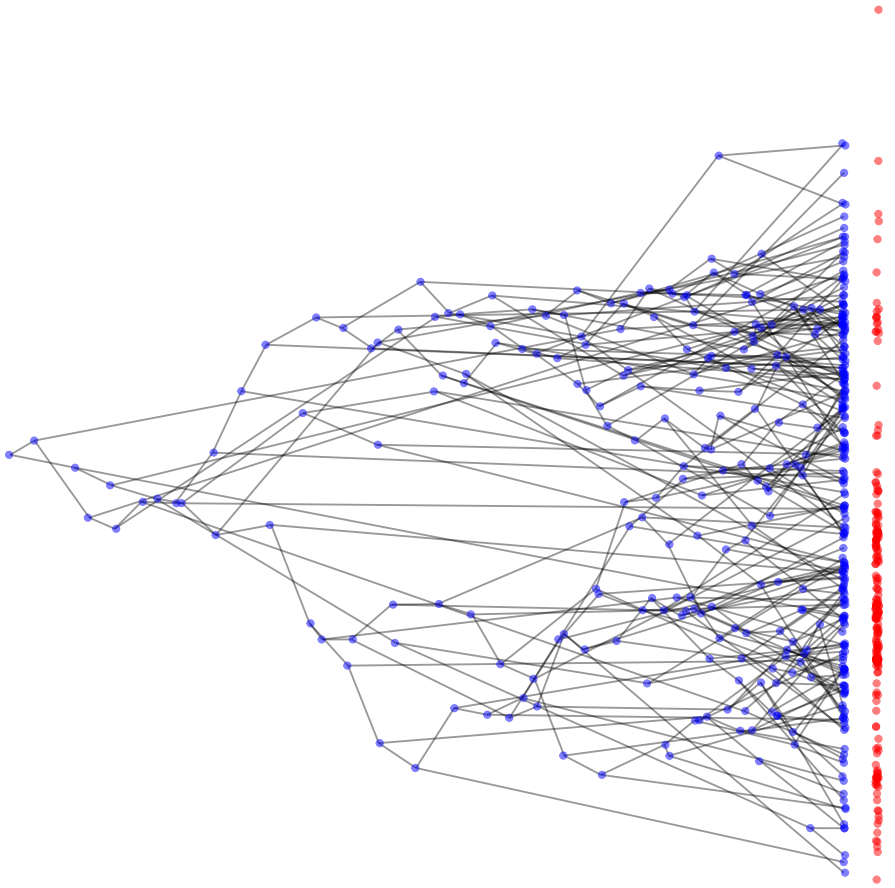
GAUSSIAN TREE MODELS OF GENE LENGTH EVOLUTION

Exam project for the course *Models for Complex Systems*



1. Introduction

A gene is a segment of DNA in the chromosomes of organims. For the purpose of this project, a gene can be thought of as a sequence of letters `ATGGCCTAGGTT...` from the four-letter DNA alphabet {`A,C,G,T`}. Some genes are quite short – only a few hundred letters – while other genes are very long, and some can be hundreds of thousands of letters long.

During evolution, genes within an organism can mutate, and these mutations are then inherited by their offspring. Moreover, populations of organisms divide in a process called speciation and branch out into different species. For the purpose of this project, we can think of all organisms from a species as having the exact same genes each with the exact same DNA-sequence across organisms, though in reality there may be small differences between organisms.

For species that have survived until today, we can sequence genomes and compare the genes we find in one species to genes from other species. A gene from one species will often correspond to very similar genes in other species, though some genes are only found in a subset of present day species. Similarity can be due to shared ancestry during evolution, where the similar genes across the species have evolved from a common ancestor. But at some point back in time, the lineages that let to species A and species B branched out from the common ancestor, and mutations in the genes have happened independently since then. For that reason, even genes that share a common ancestor may be different.

Genes in different species that share a common ancestor are called *orthologous*. In the project we will focus entirely on the lengths of orthologous genes defined as their number of DNA-letters. We will let

$$X_1, \ldots, X_n$$

denote these lengths for species $1, \ldots, n$, and we will build a Bayesian network model of them called a Gaussian Tree Model. In this project, orthologous genes have a unique orthId, and we may refer to *a gene* from hereon to mean the collection of orthologous genes from the different species that all share the same orthId.

Figure 1 shows the distributions of lengths of the genes in the data set with orthId `1CQBX`, `1CQJ6` and `1CTI9` from 204 vertebrates. In the data file for this project you will find the lengths for ten different genes including these three. Not all genes are found in all species, with e.g. `1D0EM` only in 188 of the species.

With the Gaussian Tree Model we will get a tool to explore data of this form and a systematic way to trace the evolutionary history of present day observations of gene lengths.

2. A generative model

The generative model we will consider is a Bayesian network, whose graph is a directed rooted binary tree. The root is denoted Z_0, all other non-leaf nodes are denoted $Z_1, \ldots, Z_{n-2}$ and the leaves are $X_1, \ldots, X_n$.

There is a *time* associated with every edge. For an edge into the leaf X_i we denote the time from its parent by t_i, and for an edge into Z_i we denote the time by $\tilde{t}_i$. The CPDs are *linear* Gaussian and given by $P(Z_0) = \mathcal{N}(\alpha_0, \sigma_0^2)$ and

$$P(Z_i \mid \mathrm{Pa}_{Z_i} = z) = \mathcal{N}(\alpha\tilde{t}_i + \beta z; \sigma^2\tilde{t}_i)$$
$$P(X_i \mid \mathrm{Pa}_{X_i} = z) = \mathcal{N}(\alpha t_i + \beta z; \sigma^2 t_i)$$

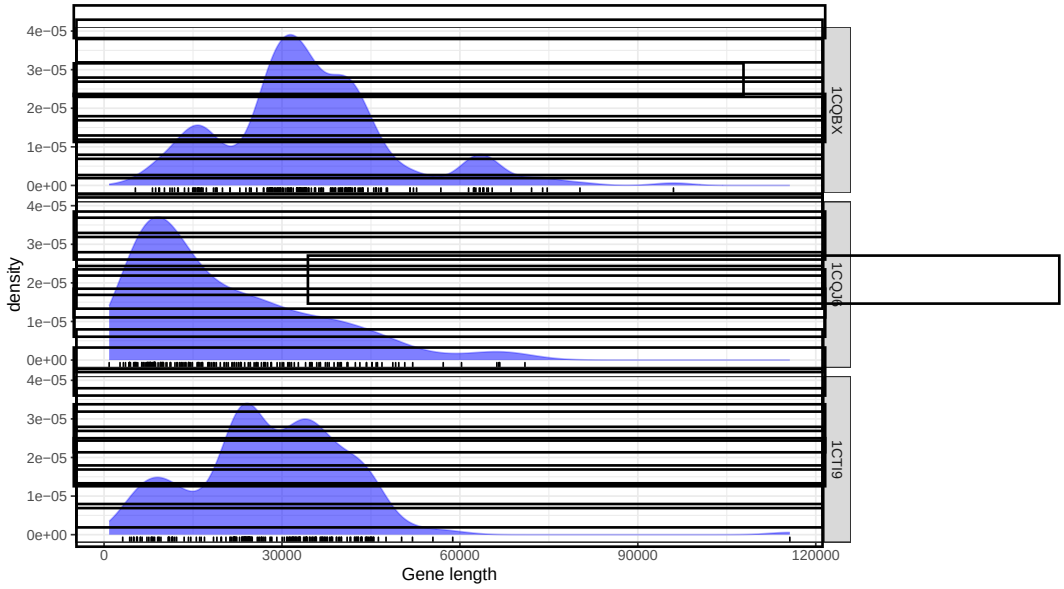


FIGURE 1. Density and rug plots of gene lengths for three genes (three different orthIds) across the different species.

for the non-root nodes. The parameters of the model are $\alpha_0, \alpha, \beta \in \mathbb{R}$ and $\sigma_0^2, \sigma^2 > 0$.

The model above with $\beta = 1$ and $\alpha = 0$ is called a random walk or random drift model. With these parameters, gene lengths simply fluctuate randomly up or down with a variance that is proportional to the time. If $\alpha \neq 0$ there is a systematic drift (this could be evolutionary preasure). A model with $\beta \neq 1$ means that genes either grow or shrink systematically in proportion to their lengths.

For part I of the project you need to:

- Implement forward simulation from the Gaussian Bayesian network using the graph for the 204 vertebrates provided in the data file.
- Illustrate the implementation by generating example data and present them visually.
- Fit a linear regression model of Z_0 given $X_1, \ldots, X_n$ using (lots of) simulated data.

You can for the simulations use $\alpha_0 = 50,000$, $\sigma_0^2 = 5,000$, $\alpha = 0$, $\beta = 1$ and $\sigma^2 = 2,500$. But you are also welcome to test what happens for other choices of parameters.

The model can be generalized by letting the conditional means be

$$\gamma_0 + \alpha t_i + \beta z + \gamma t_i z$$

for two additional parameters $\gamma_0, \gamma \in \mathbb{R}$. This generalization models, via the γ parameter, an *interaction* between time and gene length. You are welcome to also explore this generalization, but it is not a requirement.

3. Inference of hidden nodes

Only the leaf nodes, X_i, are in practice observed, and prediction of the unobserved gene lengths of the ancestors, $Z_0, \ldots, Z_{n-2}$, is an inference problem.

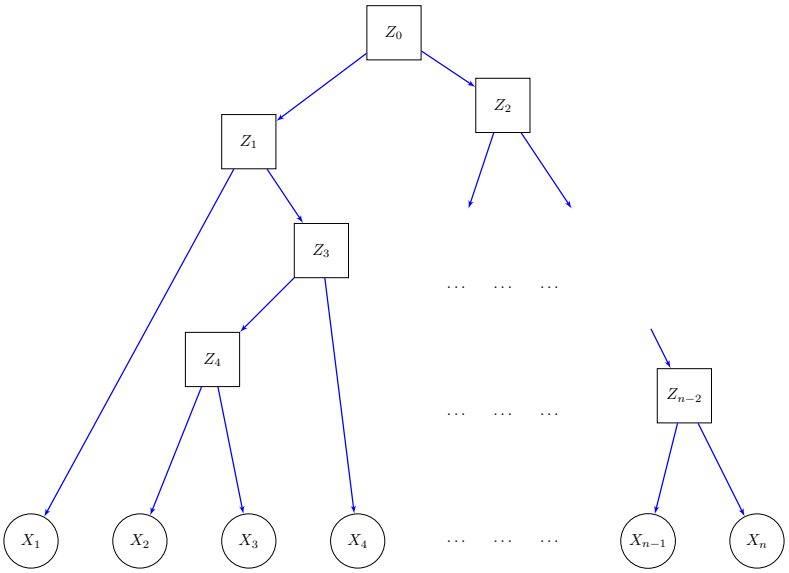


FIGURE 2. An illustration of the Bayesian network structure used in this project. The variables $X_1, \ldots, X_n$ are gene lengths of orthologous genes from n present day species, and the variables are the leaves in a tree where the root, Z_0, and the other non-leaves, $Z_1, \ldots, Z_{n-2}$, are generally unobserved. Each non-leaf variable has precisely two children and equals the gene length in the evolutionary most recent common ancestor of its children.

In part II of the project you need to:

- Implement inference algorithms for computing the conditional distribution of each of the variables $Z_0, \ldots, Z_{n-2}$ given $X_1, \ldots, X_n$.
- Test the inference algorithms using simulated data.
- Apply the inference algorithms on the data in the data file and present the results.

The conditional distribution of $Z_0, \ldots, Z_{n-2}$ given $X_1, \ldots, X_n$ is Gaussian and can be found by matrix algebra. It is useful to implement this for testing, but you must implement at least one efficient algorithm for computing $Z_i \mid X_1, \ldots, X_n$ for $i = 0, \ldots, n-2$.

You can also test the implementations by comparing results to the results from the linear regression model found using forward simulation in Part I.

4. Learning of the parameters

The objective of this part is to learn the parameters $\alpha_0, \alpha, \beta \in \mathbb{R}$ and $\sigma^2 > 0$ from data. The parameter $\sigma_0^2 > 0$ is still fixed as $\sigma_0^2 = 5,000$. The ultimate goal is to learn the parameters from observing only $X_1, \ldots, X_n$, but the problem is broken down so you first consider learning from a complete observation of all variables.

- Suppose first that all variables, $Z_0, \ldots, Z_{n-2}, X_1, \ldots, X_n$ are observed. Let $\hat{\alpha}_0 = Z_0$ and implement, using linear regression, learning of the three other parameters, α, β, σ^2. Test the implementation using simulated data.

- Proceed to implement learning with only $X_1, \ldots, X_n$ observed. One simple solution is the hard-assignment EM algorithm, which combines the inference algorithm from Part II with the linear regression above. Thus iteratively compute
$$\hat{Z}_i = E(Z_i \mid X_1 = x_1, \ldots, X_n = x_n),$$
and update the parameters using $\hat{Z}_0, \ldots, \hat{Z}_{n-2}, X_1, \ldots, X_n$ as observations. Explore convergence using simulated data and the data from the data file.

Alternatives to the hard-assignment EM algorithm are gradient ascent and the soft-assignment EM algorithm. You are welcome to explore such alternative algorithms, but this is not required.

5. Data

Data for this project comes in the file `proj_Gauss.zip`, which is a zip-file. It contains 2 files, with the file `vert_genes.csv` containing a table with gene lengths of 10 genes across 204 vertebrates, and with the file `tree.csv` containing data on the evolutionary tree of the 204 species.

The `vert_genes.csv` file contains data in a table with four columns with the names `ensembl_id`, `orthId`, `glength` and `species`. Here `ensembl_id` is a unique gene identifier, `orthId` is as explained the identifier of all orthologous genes, `glength` is the gene length and `species` is the name of the species.

The `tree.csv` file contains four columns with the names `Parent`, `Child`, `age_ch`, `t` and `species`. This table is basically an edge list with nodes in the graph labelled 1 to 407 and the columns `Parent` and `Child` containing all edges from a node in the `Parent` column to a node in the `Child` column. The column `t` contains the length of the edge. The root, which has number 407, also appear as a child node but with its parent missing. The `age_ch` column contains the age of the child node (ages are on a relative scale and have no unit), which equals the sum of all edge lengths from that child to any leaf of the tree. Any leaf node has age 0 and leaves are labelled 1 to 204. The `species` column contains the name of the species for any leaf node and is missing otherwise.

6. Postscript

The cover page shows a simulation from the Gaussian Tree Model with the blue points indicating gene lengths (y-axis) and age of node (x-axis). All the blue points aligned vertically to the right correspond to present day species, and the other blue points correspond to common ancestors in the evolutionary tree. The red points to the right are the gene lengths for the orthologous genes with orthId `1CQBX`.

The data originates from the paper: FANCD2 binding identifies conserved fragile sites at large transcribed genes in avian cells, 2018, *NAR 46(3), 1280–1294*, by Pentzold C, Shah SA, Hansen NR, Tallec BL, Seguin-Orlando A, Debatisse M, Lisby M, Oestergaard VH.

Though data was used for a different purpose in the paper, and a different analysis was carried out, the Gaussian Tree Model was inspired by that paper.