RECONSTRUCTING PHYLOGENETIC TREES WITH MISSING LEAFS USING CONSENSUS OPTIMIZATION

1. Introduction

The problem statement is as follows. Let S be a series of species, $S = 1, \ldots, |S|$ each with several individuals $N_S = 1, \ldots, n_s$. For every individual we are provided with a phylogenetic tree: we define i, j to be the indices of the leafs of the tree. Assuming that leafs are missing in some of the trees, we would like to infer the missing distances so as to use them in our analysis. In the literature it is common for missing leafs to be disregarded in any and every tree. Another approach would be to estimate the missing distances by calculating the average of all distances in trees where the missing leafs are present.

Consensus optimization is an approach to tackling large-scale, non-decomposable mathematical programs. The main idea is that instead of solving one very big and convoluted problem, a decomposition scheme is devised and then k subproblems (of smaller size) are solved. However, seeing as the k subproblems need to agree on exactly one optimal solution (or, reach consensus), the initial solution obtained by the subproblems is bound to be inconsistent. Hence, an iterative "negotiation" is initiated by the subproblems; in our case this happens in the form of an Augmented Lagrangian term in each objective function.

2. Approach

First, let us define the following parameters and variables. We say that $s, t \in S$ correspond to different species s and t, respectively. Furthermore, we say that $m, n \in N_s$ are two individuals (m and n) belonging to species $s \in S$. Last, with i, j we refer to the leafs of a phylogenetic tree. Leafs that are both present result in a readily available distance d_{ijn} for some individual n. By x_{ijn} we refer to the (missing or present) distance variable between leafs i and j in individual n. Hence, the overall problem we are trying to solve is, in optimization form:

(1)
$$QP : \min \sum_{s\in S} \sum_{t\in S:s\neq t} \sum_{n\in N_s} \sum_{m\in N_t} \sum_{i<j} (x_{ijn} - x_{ijm})^2$$

(2) $$s.t.\ x_{ijn} = d_{ijn}, \qquad \forall \text{ available } d_{ijn}, \forall i,j,n \in N_s, \forall s \in S$$

(3) $$0 \leq x_{ijn} \leq \overline{x}, \qquad \forall i,j,n,$$

(4) $$(x_{ijn} - x_{ijm})^2 \leq \epsilon, \qquad \forall s \in S, \forall n,m \in N_s.$$

The objective function captures the (Euclidean) distance between any two individuals belonging to different species and aims to minimize it. Then, the constraints are in order setting the distances of the known (present leafs), ensuring nonnegativity and maximum value of all unknown distances, and, last but not least, ensuring that any distance between leafs of individuals belonging to the same species s can never exceed a threshold value ϵ.

As this problem can grow to be very large, we propose a decomposition scheme as follows. For each species $s \in S$, we define the following subproblem:

(5)
$$SP_s : \min \sum_{t\in S:s\neq t} \sum_{n\in N_s} \sum_{m\in N_t} \sum_{i<j} (x_{ijn} - x_{ijm})^2$$

(6) $$s.t.\ x_{ijn} = d_{ijn}, \qquad \forall \text{ available } d_{ijn}, \forall i,j,n \in N_s, \forall s \in S$$

(7) $$0 \leq x_{ijn} \leq \overline{x}, \qquad \forall i,j,n,$$

(8) $$(x_{ijn} - x_{ijm})^2 \leq \epsilon, \qquad \forall i,j,n,m \in N_s.$$

Note that the above subproblem only incorporates information that pertain to species s in the original problem (QP). The problem now lies with the fact that all $|S|$ subproblems need to reach consensus for the common subvectors of x that they contain. To do that, we create $|S|$ copies of each variable and obtain the formulation:

(9)
$$SP_s : \min \sum_{t\in S:s\neq t} \sum_{n\in N_s} \sum_{m\in N_t} \sum_{i<j} (x^{(s)}_{ijn} - x^{(s)}_{ijm})^2$$

(10) $$s.t.\ x^{(s)}_{ijn} = d_{ijn}, \qquad \forall \text{ available } d_{ijn}, \forall i,j,n \in N_s, \forall s \in S$$

(11) $$0 \leq x^{(s)}_{ijn} \leq \overline{x}, \qquad \forall i,j,n,$$

(12) $$(x^{(s)}_{ijn} - x^{(s)}_{ijm})^2 \leq \epsilon, \qquad \forall i,j,n,m \in N_s,$$

(13) $$x^{(s)}_{ijn} = x^{(t)}_{ijn}, \qquad \forall i,j,n, \forall s \neq t.$$

The last constraint equates all different copies of the variables. As easily seen, this will lead often to infeasible solutions. Hence, we dualize the coupling constraint and add it to the objective function as an Augmented Lagrangian term in the objective function. In the next formulation, ρ can be viewed as the quadratic penalty factor and λ as the respective vector of the Lagrange multipliers.

$$SP_s : \min \sum_{t\in S:s\neq t} \sum_{n\in N_s} \sum_{m\in N_t} \sum_{i<j} (x_{ijn}^{(s)} - x_{ijm}^{(s)})^2 + \sum_{t_1\neq s} \lambda_{st_1} \sum_{t\in S} \sum_{n\in N_t} (x_{ijn}^{(s)} - x_{ijn}^{(t)}) + \frac{\rho^2}{2} \sum_{t_1\neq s} \lambda_{st_1} \sum_{t\in S} \sum_{n\in N_t} (x_{ijn}^{(s)} - x_{ijn}^{(t)})^2 \tag{14}$$

$$s.t.\ x_{ijn}^{(s)} = d_{ijn},\ \forall\ available\ d_{ijn}, \forall i,j,n \in N_s, \forall s \in S \tag{15}$$

$$0 \leq x_{ijn}^{(s)} \leq \overline{x},\ \forall i,j,n, \tag{16}$$

$$(x_{ijn}^{(s)} - x_{ijm}^{(s)})^2 \leq \epsilon,\ \forall i,j,n,m \in N_s. \tag{17}$$

Updating the Lagrange multipliers can be performed using (18).

$$\lambda_{st} \leftarrow \lambda_{st} + \rho(x^{(s)} - x^{(t)}) \tag{18}$$

Last, the penalty factor can be updated using a non-decreasing function of the iteration. As an example, $\rho \leftarrow \rho \cdot \alpha$ or $\rho \leftarrow \rho^{\alpha}$ can be used.

3. Linear Regression

The above optimization problem behaves like a *linear regression* model. Since the objective function that is minimized captures the "distance" (Euclidean or 2-norm) between the leaf distances between a species individual ($n \in N_s$) and other species individuals ($m \in N_t, t \neq s$), the result for a pair of leafs (i,j) is a line of the form:

$$d_{ijn} = \sum_{t\in S\setminus\{s\}} \sum_{m\in N_t} \alpha_m d_{ijm}.$$

Seeing as no individuals of the same species are present in the objective function, they are also not present in the linear regression. However, seeing as there exists a constraint that ensures individuals within the same species are not too far apart (see, for example, constraint (17)), then if the linear regression results in a pair distance that does not satisfy that requirement, the pair distance is updated and modified. The parameters α_m for every individual ends up being equal to $\frac{1}{\sum_{t\in S\setminus\{s\}} |N_t|}$ (hence, producing a weighted average based on the number of individuals in each species). We portray this with an example.

Example. Consider the three species shown in Tables 1, 2, and 3, where pair distances that are equal to -1 imply that either one or both leafs are missing.

TABLE 1. Species 1 distances.

Tree 1	6	8	7	3	4	3
Tree 2	-1	-1	-1	4	4	4
Tree 3	5	-1	7	-1	-1	3
Tree 4	6	7	-1	5	-1	-1

TABLE 2. Species 2 distances.

Tree 1	6	3	2	5	1	4
Tree 2	-1	-1	-1	5	2	3
Tree 3	9	-1	8	-1	-1	2
Tree 4	5	3	-1	5	-1	-1
Tree 5	5	-1	-1	-1	-1	-1

TABLE 3. Species 3 distances.

Tree 1	3	3	7	5	6	7
Tree 2	-1	-1	-1	3	6	6
Tree 3	5	-1	6	-1	-1	8

Solving the model when setting $\epsilon = max\ distance$, we obtain the results in Tables 4, 5, and 6, whereas if we let $\epsilon = M$ (where M is big enough), we have the Tables 7, 8, and 9.

TABLE 4. Species 1 solution.

Tree 1	6	8	7	3	4	3
Tree 2	**5.38**	**6**	**6.08**	4	4	4
Tree 3	5	**6**	7	**4.46**	**4.1**	3
Tree 4	6	7	**6.08**	5	**4.1**	**5**

Let us point out attention first to Species 1, Tree 2, where the biggest difference is observed. In the 2nd pair of leafs, the results with constraint (17) show as optimal distance the value 6, whereas once the constraint is effectively lifted, the optimal distance becomes 4.36. Let us see where this value comes from: if we see the optimal distances when letting ϵ be big enough is the average of the other two species distances for the same pair of leafs. That is, (3+5.26+5.26+3+5.26+3+5.06+5.06)/8=34.9/8=4.36. Hence, under the absence of (17) the optimization problem is akin to calculating an average. On the other

TABLE 5. Species 2 solution.

Tree 1	6	3	2	5	1	4
Tree 2	**5.05**	**5.71**	**6.48**	5	2	3
Tree 3	9	**5.71**	8	**4.12**	**4.6**	2
Tree 4	5	3	**6.48**	5	**4.6**	**5.14**
Tree 5	5	**5.71**	**6.48**	**4.12**	**4.6**	**5.14**

TABLE 6. Species 3 solution.

Tree 1	3	3	7	5	6	7
Tree 2	**5**	**5**	**6.18**	3	6	6
Tree 3	5	**5**	6	**4.41**	**4**	8

TABLE 7. Species 1 solution with big M.

Tree 1	6	8	7	3	4	3
Tree 2	**5.51**	**4.36**	**6.07**	4	4	4
Tree 3	5	**4.36**	7	**4.46**	**4.02**	3
Tree 4	6	7	**6.07**	5	**4.02**	**5.05**

TABLE 8. Species 2 solution with big M.

Tree 1	6	3	2	5	1	4
Tree 2	**5.19**	**5.26**	**6.47**	5	2	3
Tree 3	9	**5.26**	8	**4.12**	**4.52**	2
Tree 4	5	3	**6.47**	5	**4.52**	**5.15**
Tree 5	5	**5.26**	**6.47**	**4.12**	**4.52**	**5.15**

hand, under the presence of (17), the optimization problem can no longer select 4.36 as optimal, since it now is *infeasible*. Hence, it changes the distance value to become feasible by setting it equal to a value that is within the maximum distance present in that particular species (in this example, the maximum distance between pair distances for Species 1 is 2, coming from the 4th pairwise distance of Trees 1 and 4).

A small observation: changing the distance between the 4th pair of leafs for Tree 4 of Species 1 from 5 to 4, results in changing the optimal value for the above distance from 6 to 7, as now the maximum "allowed" distance within the first species is 1.

TABLE 9. Species 3 solution with big M.

Tree 1	3	3	7	5	6	7
Tree 2	**5.86**	**5.06**	**6.17**	3	6	6
Tree 3	5	**5.06**	6	**4.41**	**3.62**	8