Notes about Network Propagation

Yiftach Kolb

May 31, 2020

1 Matrix Algebra Primer

Definition 1.1. A **Transition Matrix** is a real valued non negative square (n^2) matrix A s.t. each of its columns sums up to one:

$$\forall j \sum_i A_{i,j} = 1, \ \forall i, j A_{i,j} \geq 1$$

A acts from the left as a linear mapping: $T(v) := Av, \forall v \in R^n$. In the following script we might interchangeably and indistinguishably use T (the mapping) or A the matrix.

A transition is **positive**, designated by $A > 0$ if all its entries are positive.

A transition is **regular** if for some $k >$ A^k is positive.

Definition 1.2. A State is a non-negative vector $v \in R^n$ s.t $\sum_i v_i = 1$.

Remark 1. If v is a state then and A is a transition as defined in 1.1, Av is also a state because:

$$\sum_i (Av)_i = \sum_j v_j (\sum_k A_{j,k}) = \sum_j v_j \cdot 1 = 1$$

Lemma 1.1. *If T is a transition, then there is a state u such that $Tu = u$.*

Proof. Because the columns of A all sum to 1, the columns of $A - I$ all sum to 0. Therefore $(1, 1, \ldots, 1)$ is an (left) eigenvector of the rows with eigenvalue 1. Therefore there is also some real (right) column eigenvector with eigenvalue 1. (Also it follows from the Brouer fixed point theorem because T maps the l_1 sphere to itself).

Let $u \in R^n$ be such vector: $Au = u$. Let $u = u^+ + u^-$ such that $u^- = \min(u_i, 0)$ and the other one defined similarly for the non-negative components.

Because A is non-negative $A(u^+) \geq (Au)^+ \geq 0$, and $A(u^-) \leq (Au)^- \leq 0$.

From A being non-negative and $(Au)^+ + (Au)^- = Au = u = u^+ + u^-$ And also $(Au)^+ = (u)^+$, so we must have $Au^+ \geq (Au)^+ = u^+$ (component wise). But if we had a strict inequality we would get: $\|A(u^+/\|u^+\|_1)\|_1 > 1$ which is a contradiction to A being a transition matrix.

Then $Au^+ = u^+, Au^- = u^-$ and one of them must be non-zero. It follows that A has a non-negative eigenvector with eigenvalue 1 (one of $u^+, -u^-$ which is non-zero). If we l_1-normalize that eigenvector it becomes a state. □

Lemma 1.2. *If A is strictly positive (entries are all positive), then it has exactly one eigenvector with eigenvalue 1.*

Proof. We already know that there exists at least one such eigenvector. Let u, v s.t $Au = u, Av = v$. We can assume these are real vectors because A has only real entries. Therefore we can choose u, v to be states as we may do have proven above.

Then let $w = u - v$. So $Aw = A(u - v) = u - v = w$. And $\sum_i w_i = 1 - 1 = 0$ by choice of u, v.

Like before $Aw^+ = w^+, Aw^- = w^-$ and because $w \neq 0$ but sums up to 0, both $w^+, -w^- > 0$. Because w^- is non zero exactly on entries where w^+ is zero and vice versa, each of them must have at least one 0 entry (and one none zero). But because A is strictly positive and w^+ is non-negative, Aw^+ must have ONLY positive entries, contradicting $Aw^+ = w^+$. It follows then that $u - v = 0$ is the only possibility, hence the eigenvector with 1 as eigenvalue is unique. □

□

So far we have seen than if T is a transition it has a stationary state. If it is positive (and we could extend it to regular) this stationary state is unique.

Now we want to find out more about the spectrum of T in the positive (regular) case. We want all the other eigenvalues to be strictly less than 1, because then we could easily conclude that the long term distribution always converges to the stationary.

So in fact the Perron-Frobenius theorem guaranties that the other eigenvalues are indeed strictly less than 1. But we can also try to show that directly below.

Definition 1.3. let $u \in \mathbb{C}^n$, then we define: $\tilde{u} = (|u_i|)_{i \in [n]}$, meaning the entires of $\tilde{u}$ are the absolute values of the entries of u. It follows then that both have the same l_k norms.

Remark 2. If $u \in \mathbb{C}^n$ and T a transition then componentwise:

$$\widetilde{Tu} = (\sum_j \widetilde{T_{i,j}u_j})_i = (|\sum_j T_{i,j}u_j|)_i \leq (\sum_j T_{i,j}|u_j|)_i = T(\tilde{u})$$

If T is positive, and if for some j u_j is complex or non-positive than the inequality is strict.

But since $\|\widetilde{Tu}\| = \|Tu\|$ we have then: $\|Tu\| \leq \|T\tilde{u}\|$ with equality only in case that $u = \tilde{u}$ if T is positive.

Theorem 1.1. *1. Let T be a positive (regular) transition. Then 1 is the greatest eigenvalue of T and it has one unique eigenvector which is also non-negative, so there exists a unique stationary state.*

2. All the other eigenvalues have absolute value strictly less than 1.

3. For every state u, $T^k u$ converges to the stationary state π. In particular the columns of T converge to π.

Proof. 1. We already know.

2. Suppose that transition T has an eigenvalue $\lambda \neq 1$ such that $|\lambda| \geq 1$. Let $u \in \mathbb{C}^n$ be a corresponding eigenvector $Tu = \lambda u$ and choose it so that $\|u\|_1 = \|\tilde{u}\|_1 = 1$.

then

$$\|T\tilde{u}\| \geq \|Tu\| = \|\lambda u\| = \|u\| = \|\tilde{u}\|$$

If T is positive, as explained above, then equality holds just in the case where $u = \tilde{u}$. If the inequality is strict, and we choose the l_1 norm, we get a contradiction because $\tilde{u}$ is a state and therefore $T\tilde{u}$ is also a state. So this means that we must have equality, and therefore we must have $u = \tilde{u}$ and as we have already proven before T has a unique stationary state if it's positive. □

3. Let u be the unique stationary state of T, and let w be any state. Then we can write $w = \alpha u + v$ such that v is spanned by the other eigenvectors of w (whose eigenvalues are all some $|\lambda_i| < \lambda < 1$

If we explicitly write $v = \sum_i \beta_i s_i$ where the s_i are the other eigenvectors (with smaller than 1 eigenvalues), then:

$$\|T^k v\| = \|\sum_i \beta_i s_i\| \leq \lambda^k \sum_i |\beta_i| \to_{k\to\infty} 0$$

So we get a rather weird result, that $T^k w = T^k(\alpha u + v) = \alpha u + T^k v \to \alpha u$

But w is a state and T^k is a positive transition for all k (for all bigger k if T is just regular). So $T^k w$ is also a transition, so αu as the limit of a continuos map in a compact set also belongs to the positive part of the l_1 sphere, so it must also be a transition, which means that we must have taken $\alpha = 1$ (is that weird or did I make a mistake somewhere?), and so $T^k w \to u$ □

I think all of the above can easily be extended to regular transitions ....

□

2 More on Matrices, Graphs and Stochastics

A directed graph can be uniquely represented by its adjacency matrix. $A_{i,j} = 1$ if and only if there is a directed edge from i to j (if we want to use it for transitioning on columns as done above). It's possible to assign different edge weights rather than only 1 and 0. If the graph is undirected each edge would be counted in both directions and the matrix is symmetric.

Then to turn A into a transition, normalize each column by dividing it with its out going rank, so let $D_{i,i} = \text{outrank}(i)^{-1}$, AD is the transition matrix of this graph (because right-multiplying by D normalizes each column by its rank).

A graph is strongly connected (meaning there is a directed path form any edge to any edge) iff the adjacency matrix is irreducible. A matrix is Irreducible by definition if it is a similar by permutations matrices to a block upper triangular matrix.

If a graph has a reducible adjacency matrix it cannot be strongly connected. If we start with such matrix and replace all its non-zero entries by one (or any positive number), we could never fill up all the entries by exponentiating the matrix, whereas with any other non-negative matrix we could populated all the entries with positive numbers if we exponentiate sufficiently high.

Intuitively it is easy to see why

If we know the graph is strongly connected, if we take its transition matrix and exponent it, eventually all the entries will be non 0 because of the strong connectivity, every node can be reached from every node, so such matrix cannot be similar to a triangular matrix, hence it is irreducible.

Other direction: if the graph is not strongly connected we want to show its adjacency matrix is reducible. We take a node i that has a minimal number of reachable nodes. If i is a sink then we switch it's name to 0. We assume by induction that the rest of the nodes have some permutations that results in a triangular matrix and then we trivially extend it with the sink node that we took out as the new first column. If i is not a sink by minimality there must be a cycle of minimally connected nodes including i, so we reduce all of them to one representative which must be a sink of the reduced graph. Now build the block triangular matrix on the smaller graph (so induction hypothesis), then extend it again which should be easy because the removed nodes form a block that connects only to itself (columns with 1 only between said indices, 0 otherwise) so we can put them as the first indices of the matrix $\dots$. This also confirms that irreducible matrix is equivalent to regular matrix as defined in the first section.

It is possible to turn a reducible transition into an irreducible by way of random restart as explained in the article Cowen et al. [1].

If we start with the adjacency matrix and replace each 0 with some small ϵ representing a light weight edge, then normalize the rows, we would get the same thing. Another way to see it: take two transition matrices P and Q, then any convex combination of them, namely $\alpha Q + (1-\alpha)P$ for $\alpha \in [0,1]$ is also a transition because the rows clearly still all sum to 1.

Remark 3. Regarding permutation matrices and similarity. There is a natural isomorphism between the permutation group S_n and the parmutation matrices of $n \times n$ size: $\pi \mapsto (e_{\pi(1)}, \dots, e_{\pi(n)})$. If P is a permutation matrix with a corresponding permutation π, and A any (n square) matrix, then PA is be the matrix obtained by permuting the rows of A according to the permutation π. AP is the result of permuting the columns by the permutation π^{-1}, why is that? see below for the explanation but if you just consider that $P = \prod \Theta_i$ is the product of permutation matrices that each corresponds to a 2-cycle permutations (I think these are called transpositions), it becomes clear.

Now when we are dealing with adjacency matrices, we want to rename the indices, thereby rearranging the adjacency matrix to be block triangular. This ’rearrangement’ means exactly multiplying it from left and right by some PAP^{-1}, and from here arises this similarity condition. If we recall every permutation π is a composition of 2-cycles, and 2-cycles are their own inverse. So when we switch indices i and j what we actually do is switching row i with jm then permuting column i with j. Then, we permute indices i' and j', so the row permutation will stack up from the left, the column permutation will stack from the right, and thus arises this PAP^{-1} type of matrix from the original adjacency matrix.

Lemma 2.1. *If s is a state (column), T is a transition, and $\bar{1}$ is the matrix with all 1's, then $\bar{1}s = (1,1,\dots,1)^t = \mathbf{1}$ and $\bar{1}T = \bar{1}$.*

Proof. Every element of the product matrix is a sum of a column of T, hence 1. s is the same it's like taking just one column of T.

We therefore see that, using the notations of the article [1] Since $\bar{1} \cdot P(t) = \bar{1}$, using $(\frac{(1-\alpha)}{N}\bar{1} + \alpha W)$ as the transition matrix (where W is the transition originated from the normalized adjacency matrix of the original graph), we get:

$$P(t+1) = (\frac{(1-\alpha)}{N}\bar{1} + \alpha W) \cdot P(t) = \frac{(1-\alpha)}{N}\mathbf{1} + \alpha P(t)$$

Also, instead of taking the uniform distribution, let s be some state, and let $S = (s, s, \dots, s)$ a matrix whose columns are all s. Then if p is any state, $Sp = s$. So we can replace $\mathbf{1}/N$ with S in the above remarks, so our random restart distribution can be chosen arbitrarily as long as it is has no 0 entries.

□

Remark 4. As we have seen for T irreducible, and any state s, $T^k s \to \pi$ where π is the unique stationary state.

π can be approximated by itetrating this sequence until sufficent accuracy has been reached.

There is also a direct solution for the random walk with restart which goies as follows: let $p > 0$ be a state (no 0 entries), let W be its column-normalized adjacency matrix, and let $P = (p, \dots, p)$. as we have seen above, $T = (1-\alpha)P + \alpha W$ is an irreducible transition matrix therefore π exists and unique.

Let $p_0 = p, p_k := Tp_{k-1} = p_0(1-\alpha) + \alpha p_{k-1}$ Then $p_k \to \pi$ and We have the relation:

$$I\pi = \lim(p_k) = (1-\alpha)p_0 + \alpha \lim(Wp_{k-1}) = (1-\alpha)Ip_0 + \alpha W\pi$$

and rearragement gives:

$$(I - \alpha W)\pi = (1-\alpha)p_0$$

Now because W is a transition and $0 < \alpha < 1$, for any v such that $\|v\|_1 = 1$, then $\tilde{v}$ is a state and it holds that:

$$\|\alpha W v\|_1 \le \alpha \|W\tilde{v}\| = \alpha < 1$$

This guaranties that $I - \alpha W$ is invertible and the direct solution is:

$$\pi = (1-\alpha)(I - \alpha W)^{-1} p_0$$

3 The Markov Random Fields Methd

We use almost the same terminology of Deng et al. [2] so I won't repeat everything here.

We have a PPI and partial annotaion which gives the potential function $U(x)$ which actually depends on the parameter $\Theta = (\alpha, \beta, \gamma)$ And we want to maximize:

$$P(X|\Theta) = \frac{1}{Z(\Theta)} \exp(-U(x)) \tag{1}$$

To eliminate $Z(\Theta)$ from the equation we are looking at:

$$P(X_i = 1|X_{[-i]}, \Theta) = \frac{\exp(\alpha + (\beta - 1)M_0^i + (\gamma - \beta)M_1^i)}{1 + \exp(\alpha + (\beta - 1)M_0^i + (\gamma - \beta)M_1^i)} \tag{2}$$

If we think of every node labeled with 1 as being occupied by a traveling agent, who in the next iteration is going to visit some other (i.e not allowed to stay in the same place) node according to some probability, then $P(X_i = 1|X_{[-i]}, \Theta)$ is the probability That node i will be visited in the next iteration. In the propagation model we look at the neighbours of i that have a visitor in them, and we look for each of these how many other neighbours it has to calculate the visit probability of i, whereas in equation (2) we don't explicitly look at neighbours of neighbours but rather consider direct interaction between i and all its neighbours.

In the propagation model without restart, we have (of course with the resulting value being capped at a maximum of 1):

$$P(X_i = 1|X_{[-i]}) = \sum_{j \in \text{Nei}(i)} \frac{1}{\text{Nei}(j)} \tag{3}$$

Here we assume that the visiting agent will choose one of the neighbours at equal probability. We can also formulate this similarly to the random restart case. Also perhaps it would make sense, to consider instead of $X_{[-i]}$, a label assignment of all the nodes including i, then update X_i in the next iteration by the probability that it would be visited in the next iteration. And in this case we can add restart an reformulate (3) as:

$$P(X_i = 1|X) = (1 - \alpha) \sum_{j \in \text{Nei}(i)} \frac{1}{\text{Nei}(j)} + \alpha \sum_{j : X_j = 1} \frac{1}{\text{Nei}(j)} \tag{4}$$

$P(X_i = 1|X)$ here means that giving a labeling X to all the nodes, we calculate the probabiliyt that X_i will be labeled 1 in the next iteration which is the probabilty that at least one of the agents will choose to visit it.

I believe if we use gibbs sampling based algorithm with on (4) to find the π the stationary distribution but I am not sure if it is mor efficient than the iterative method.

3..1 Estimation of the parametes

The authors of Deng et al. [2] used the quasi-likelihood approach and logistic regression.

If I understand this correctly, they take the subnetwork of annotated proteins (of a specific function), that gives a binary vector $X = (X_i)_{i=1\ldots m}$ and each X_i is treated as an observation that is independent from the other observations, then they find the paramers that best fit the logistic model to the sample distribution.

3..2 Estimating the probabvilty

In the article, for a given a function annotaion, π is the probability that a protein has that a protein has that function annotation (disregarding the information from the PPI network). This π is used to assign random values to the missing data.

An Idea: Maybe we can use use the pagerank (and a cutoff) for the initial assignment instead?

4 Reference

References

[1] Lenore Cowen et al. "Network propagation: a universal amplifier of genetic associations". In: *Nature Reviews Genetics* 18.9 (2017), p. 551.

[2] Minghua Deng et al. "Prediction of protein function using protein-protein interaction data". In: *Proceedings. IEEE Computer Society Bioinformatics Conference*. IEEE. 2002, pp. 197–206.

[3] Israel Nathan Herstein and David J Winter. *Matrix theory and linear algebra*. Macmillan Publishing Company, 1989.

[4] Carl D Meyer. *Matrix analysis and applied linear algebra*. Vol. 71. Siam, 2000.