Poisson matrix factorization

Caterina De Bacco

1 Poisson matrix factorization: the problem

We have seen the stochastic block model (SBM) as a foundational method to cluster nodes in a network into communities based on their pairwise interactions.

Beyond enforcing stochastic equivalence and conditional independence between edges, the SBM assumes that nodes can belong to only one group. In many scenarios this assumption is too restrictive. For instance, in social networks, it is fair to assume that people can belong to more than one group, and with different intensities. In other words, we want a model for mixed-membership, or overlapping communities.

Objective: cluster people in multiple groups based on their pattern of interaction.

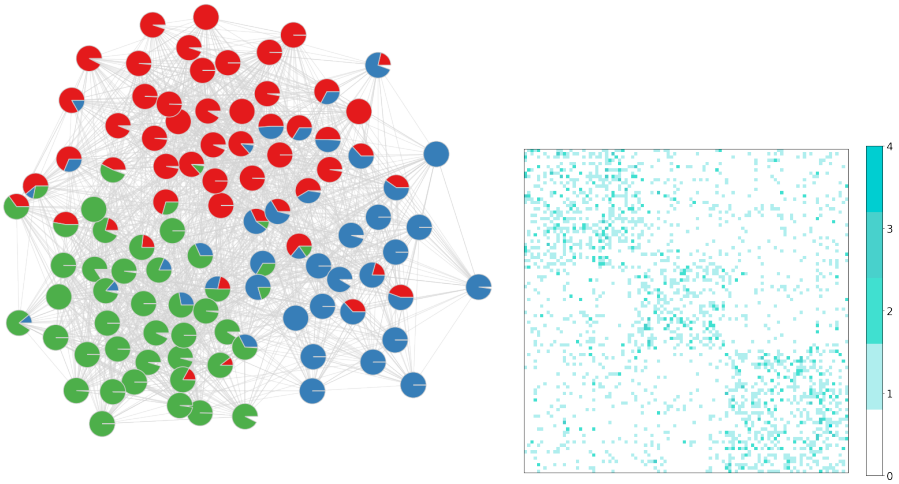

Figure 1: Example of a mixed-membership graph and its adjacency matrix.

Imposing mixed-membership can be obtained by defining the membership $u_i \in \mathbb{R}_{>0}$ of node i as K-dimensional vector of positive entries where there can be more than one non-zero entry (as opposed to the SBM where, using one-hot encoding, only one entry is equal to 1 and all the others 0). Furthermore, we may assume that the more two nodes belong to the same multiple groups, the higher their probability of interaction. Mathematically, this can be obtained by considering the dot products of the membership vectors, so that now stochastic equivalence is generalized as:

$$Pr(A_{ij} = k) = f(u_i \cdot u_j) \quad . \tag{1}$$

In order to allow directed networks, i.e. the probability of an edge (i, j) may be different than that of the opposite edge (j, i), we should consider the existence of two types of membership, out-going u_i

and in-coming v_j. We can still keep a $K \times K$ affinity matrix C, that allows to tune different topology structures, e.g. assortative, disassortative etc... Putting all together, we assume the likelihood:

$$P(A;u,v,C) = \prod_{ij} \frac{\left(\sum_{k,q} u_{ik} v_{jq} c_{kq}\right)^{A_{ij}}}{A_{ij}!} e^{-\sum_{k,q} u_{ik} v_{jq} c_{kq}} \quad . \tag{2}$$

This is an example of a matrix factorization, in particular a Poisson matrix factorization (PMF). In fact, denoting with U and V the $N \times K$ membership matrices of rows u_i and v_i respectively, we obtain that the expected value of the adjacency matrix is:

$$\mathbb{E}[A] = UCV^T \quad , \tag{3}$$

which can be related to a linear algebra problem of finding a low-rank approximation of a matrix $A \approx UV^T$.

1.1 Maximum likelihood and EM

We start by showing the approach using MLE estimation. We fit the model to an observed network by maximizing this probability with respect to the parameters, or equivalently (and more conveniently) maximizing its logarithm. Defining $\lambda_{ij} = \sum_{k,q} u_{ik} v_{jq} c_{kq}$, yields:

$$\mathscr{L}(u,v,C) = \log P(A;u,v,C) = \sum_{ij} A_{ij} \log \lambda_{ij} - \sum_{ij} \lambda_{ij} \quad , \tag{4}$$

where we omitted constants not depending on the parameters.
Obs1: Notice that this time we can take derivatives w.r.t. the parameters, as these are now real-valued quantities.
Obs2: In addition, it is not evident how to set up a Monte Carlo scheme for mixed-membership cases, as for hard-communities one was switching the colors of nodes at each step.

If we directly derive this expression by one of the parameters, e.g. u_{ik}, would lead to ugly terms like the logarithms of a sum. Instead, we use **Jensen's inequality**:

$$\log\left(\sum_k x_k\right) \geq \sum_k q_k \log \frac{x_k}{q_k} \quad , \tag{5}$$

where q_k are any probabilities satisfying $\sum_k q_k = 1$ and x_k are any set of positive numbers. The exact equality can always be achieved imposing:

$$q_k = \frac{x_k}{\sum_k x_k} \quad . \tag{6}$$

Applying this to $\mathscr{L}(u,v,C)$ we get:

$$\mathscr{L}(u,v,C) \geq \sum_{ijkq} A_{ij} q_{ijkq} \log \frac{u_{ik} v_{jq} c_{kq}}{q_{ijkq}} - \sum_{ijkq} u_{ik} v_{jq} c_{kq} := L(u,v,C,q) \quad . \tag{7}$$

The exact equality is obtained when:

$$q_{ijkq} = \frac{u_{ik} v_{jq} c_{kq}}{\sum_{kq} u_{ik} v_{jq} c_{kq}} \, , \tag{8}$$

hence the double maximization of $L(u,v,C,q)$ w.r.t. q and $\theta = (u,v,C)$ is equivalent to maximize $\mathscr{L}(u,v,C)$ w.r.t. θ. Therefore we now proceed in maximizing the more convenient L.

Obs3: maximizing L is more convenient because it contains only products inside the logarithms, not sums.
Obs4: the distribution q plays a role of a variational distribution and nicely integrates into a EM-algorithms as we saw in a past lecture. In particular, this is the distribution updated in the E-step.
Obs5: there are in principle many entries to update for determining q, as this is a $N \times N \times K \times K$, so a $N^2 \times K^2$ complexity. However, the q_{ijkq} terms only enter L through a multiplication by the A_{ij}. This means that we only need to calculate the q_{ijkq} such that $A_{ij} > 0$. In real networks, which are often sparse, this means only accounting for non-zero entries, which are equal to the number of edges $M \sim N$. This means that the complexity is linear in N, not N^2!
Obs6: the value of q_{ijkq} has a simple physical interpretation: it is proportional to the probability that an edge between i and j exists because of i having color k and j color q.

Now we can proceed differentiating w.r.t. θ. For instance, taking the derivative w.r.t. u_{ik}.

$$\frac{\partial L}{\partial u_{ik}} = -\sum_{jq} v_{jq} c_{kq} + \frac{1}{u_{ik}} \sum_{jq} A_{ij} q_{ijkq} \quad , \tag{9}$$

setting this to zero leads to:

$$u_{ik} = \frac{\sum_{jq} A_{ij} q_{ijkq}}{\sum_{jq} v_{jq} c_{kq}} \quad . \tag{10}$$

Similarly for the other parameters:

$$v_{jq} = \frac{\sum_{ik} A_{ij} q_{ijkq}}{\sum_{ik} u_{ik} c_{kq}} \tag{11}$$

$$c_{kq} = \frac{\sum_{ij} A_{ij} q_{ijkq}}{\sum_{ij} u_{ik} v_{jq}} \quad . \tag{12}$$

Maximizing the log likelihood is now simply a matter of simultaneously solving Eq. (8) and Eqs. (10) - (12), which can be done iteratively by choosing a random set of initial values and alternating back and forth between the two equations. This is the EM algorithm.

Good references for this method are Ball *et al.* (2011) and De Bacco *et al.* (2017); Contisciani *et al.* (2020) for further generalizations for multilayer networks and covariates.

1.2 Constraints and regularization

We did not impose any constraints on the parameters. However, we implicitly imposed that u, v, C must be positive (we cannot have negative parameters for the Poisson). In fact, because the updates are all multiplicative, i.e. of the type:

$$u_{ik}^{(t+1)} = \frac{\sum_{jq} A_{ij} q_{ijkq}^{(t)}}{\sum_{jq} v_{jq}^{(t)} c_{kq}^{(t)}} = u_{ik}^{(t)} \times \frac{\sum_{jq} A_{ij} v_{jq}^{(t)} c_{kq}^{(t)}}{\sum_{jq} v_{jq}^{(t)} c_{kq}^{(t)}} \quad , \tag{13}$$

where we made explicit the iteration step with the superscripts and unpacked q_{ijkq}. Hence, if we initialize $u_{ik}^{(0)} \geq 0$ and similarly for the other parameters, we will always obtain positive parameters.

We can also add further constraints, for instance the normalization $\sum_k u_{ik} = 1$. This could be enforced using Lagrange multipliers. It would lead to similar multiplicative updates but with different denominators.

Alternatively, one can take a Bayesian approach and impose a prior on the parameters to impose a regularization. For instance, if we want the entries of u and v to be small so that nodes do not belong

to many communities, i.e. sparsity constraints, then we can assume an exponential prior:

$$P(u_{ik}|a_{ik}) = a_{ik}e^{-a_{ik}u_{ik}} \quad (14)$$
$$P(v_{ik}|b_{ik}) = b_{ik}e^{-b_{ik}v_{ik}} \quad , \quad (15)$$

where a, b are hyper-priors given in input. One should then use MAP estimate instead of MLE and maximize the log-posterior:

$$L(u,v,C,q) - \sum_{i,k} a_{ik}u_{ik} - \sum_{i,k} b_{ik}v_{ik} \quad . \quad (16)$$

Notice that if we assume uniform $a_{ik} = a$ and $b_{ik} = b$, then we get a L1-regularization:

$$L(u,v,C,q) - a\sum_{i,k} u_{ik} - b\sum_{i,k} v_{ik} \quad . \quad (17)$$

A similar result can be obtained with a gamma prior.
Exercise: prove this.

1.3 Bayesian inference with Variational Inference

We now show an alternative method for parameter inference, based on VI.
In particular, this allows to add priors and calculate full posterior distributions (before we obtained point-estimates). For simplicity, we assume diagonal affinity matrix C. In this case, we do not need to incorporate the entries c_k in the model explicitly, as they will be automatically included inside u_{ik} and v_{jk} (we can always multiply by a constant).
To proceed, we need the useful property of Poisson distributions: *a sum of Poisson-distributed random variables is also a Poisson-distributed random variable, with parameter the sum of the parameters.*
We use this to extend the Poisson likelihood using auxiliary variables z_{ijk} such that:

$$P(z_{ijk}|\theta) = \text{Pois}(z_{ijk}; u_{ik}v_{jk}) \quad , \quad (18)$$

and they should obey the constraint $\sum_k z_{ijk} = A_{ij}$. This can be imposed "probabilistically" with a delta distribution $P(A|z) = \delta\left(\sum_k z_{ijk} - A_{ij}\right)$. With this, we have the extended likelihood:

$$P(A,z|\theta) = P(A|z)P(z|\theta) \quad , \quad (19)$$

whose marginal $P(A|\theta)$ i.e., the distribution on A *not* conditioned on z (so we integrate z out) is equivalent to the original likelihood.
Hence now we have the joint likelihood:

$$P(A,z|\theta) = \prod_{i,j} \delta\left(\sum_k z_{ijk} - A_{ij}\right) \prod_{k=1}^{K} \text{Pois}(z_{ijk}; u_{ik}v_{jk}) \quad . \quad (20)$$

We then set gamma priors:

$$P(u_{ik}|a,b) \propto u_{ik}^a e^{-bu_{ik}} \quad (21)$$
$$P(v_{ik}|c,d) \propto v_{ik}^c e^{-dv_{ik}} \quad . \quad (22)$$

Putting all together:

$$P(A,z,u,v) \propto \prod_{i,j} \delta\left(\sum_k z_{ijk} - A_{ij}\right) \prod_{k=1}^{K} \text{Pois}(z_{ijk}; u_{ik}v_{jk}) \prod_{ik} u_{ik}^a e^{-bu_{ik}} \prod_{ik} v_{ik}^c e^{-dv_{ik}} \quad , \quad (23)$$

where the proportionality is neglecting constants depending on the hyper-priors.
We are looking for the posterior $P(u,v,z|A)$ but this is in general intractable (the denominator involves

difficult marginalization). Hence, we adopt VI and a mean-field family of variational distributions:

$$q(u,v,z) = \prod_{ik} q(u_{ik};\alpha_{ik}^{shp},\alpha_{ik}^{rte}) q(v_{ik};\beta_{ik}^{shp},\beta_{ik}^{rte}) \prod_{ij} q(z_{ij};\phi_{ij}) \quad , \tag{24}$$

where $\Theta = (\alpha_{ik}^{shp},\alpha_{ik}^{rte},\beta_{ik}^{shp},\beta^{rte},\phi_{ijk})$ are the variational parameters that we need to find. We use what learned in a previous lecture, i.e. we write the complete conditional of each individual parameter and see how this looks like. This ensures the maximization of the ELBO for this problem.
For instance, focusing on u_{ik}:

$$P(u_{ik}|A,z,v,u_{\setminus ik}) \propto u_{ik}^{a} e^{-bu_{ik}} \prod_j e^{-u_{ik}v_{jk}} u_{ik}^{z_{ijk}} \tag{25}$$

$$= u_{ik}^{a+\sum_j z_{ijk}} e^{-(b+\sum_j v_{jk})u_{ik}} \sim \text{Gam}\left(u_{ik}; a+\sum_j z_{ijk}, b+\sum_j v_{jk}\right) \tag{26}$$

Now we use a fact learned in a previous lecture about VI. When all the complete conditionals are in the exponential family, we can use the result Blei *et al.* (2017) that the natural parameters ρ_i of the variational family satisfy:

$$\rho_i = \mathbb{E}_{q(y)} \kappa_i(y) \,, \tag{27}$$

where y is the parameter from the original posterior and $\kappa_i(y)$ is the natural parameter of the complete conditional. The expectation is with respect to the variational distribution $q(y)$. The natural parameters for a Gamma distribution $\text{Gam}(\alpha,\beta)$ are $(\alpha-1,-\beta)$; for a Multinomial $\text{Mult}(n,[\log p_1,\ldots,\log p_K])$ and a Categorical distribution $Cat([p_1,\ldots,p_K])$ are $(\log p_1,\ldots,\log p_K)$.

Hence, we can conclude that the optimal posterior is:

$$q(u_{ik};\alpha_{ik}^{shp},\alpha_{ik}^{rte}) = \text{Gam}\left(u_{ik};\alpha_{ik}^{shp},\alpha_{ik}^{rte}\right) \tag{28}$$

$$\alpha_{ik}^{shp} = a + \sum_j \mathbb{E}_q\left[z_{ijk}\right] \tag{29}$$

$$\alpha_{ik}^{rte} = b + \sum_j \mathbb{E}_q\left[v_{jk}\right] = b + \sum_j \frac{\beta_{jk}^{shp}}{\beta_{jk}^{rte}} \quad . \tag{30}$$

Similarly, we have:

$$q(v_{ik};\beta_{ik}^{shp},\beta_{ik}^{rte}) = \text{Gam}\left(v_{ik};\beta_{ik}^{shp},\beta_{ik}^{rte}\right) \tag{31}$$

$$\beta_{jk}^{shp} = c + \sum_i \mathbb{E}_q\left[z_{ijk}\right] \tag{32}$$

$$\beta_{jk}^{rte} = d + \sum_i \mathbb{E}_q\left[u_{ik}\right] = d + \sum_j \frac{\alpha_{ik}^{shp}}{\alpha_{ik}^{rte}} \quad , \tag{33}$$

where we used the fact that the mean of a Gamma distribution of shape and rate parameters α and β is α/β.
Obs: the parameters of u_{ik} are influenced by those of the v_{jk} (and vice-versa), but not by other u_{jk}.
Now, we need to update the auxiliary z_{ij}. We proceed similarly, but now we have to account for the constraint from the delta distribution:

$$P(z_{ij}|A,u,v) = \frac{\delta\left(\sum_k z_{ijk} - A_{ij}\right) \prod_{k=1}^{K} \frac{e^{-u_{ik}v_{jk}} \left(u_{ik}v_{jk}\right)^{z_{ijk}}}{z_{ijk}!}}{P(A|\theta)} \tag{34}$$

$$= \frac{\delta\left(\sum_k z_{ijk} - A_{ij}\right) \prod_{k=1}^{K} \frac{e^{-u_{ik}v_{jk}} \left(u_{ik}v_{jk}\right)^{z_{ijk}}}{z_{ijk}!}}{\frac{e^{-\sum_k u_{ik}v_{jk}} \left(\sum_k u_{ik}v_{jk}\right)^{A_{ij}}}{A_{ij}!}} \tag{35}$$

$$\propto \delta\left(\sum_k z_{ijk} - A_{ij}\right) \frac{A_{ij}!}{\prod_k z_{ijk}!} \prod_k \left(u_{ik}v_{jk}\right)^{z_{ijk}} \sim \text{Mult}\left(z_{ij};(u_{i1}v_{j1},\ldots,u_{iK}v_{jK})\right) \tag{36}$$

Hence we have the the variational posterior:

$$q(z_{ij};\phi_{ij}) \quad = \quad \text{Mult}\big(z_{ij};\phi_{ij}=(\phi_{ij1},\ldots,\phi_{ijK})\big) \quad . \tag{37}$$

Using Eq. (27) we get:

$$\begin{aligned}\log\phi_{ijk} &= \mathbb{E}_q\left[\log u_{ik}\right]+\mathbb{E}_q\left[\log v_{jk}\right] \\ &= \Psi(\alpha_{ik}^{shp})-\log\alpha_{ik}^{rte}+\Psi(\beta_{jk}^{shp})-\log\beta_{jk}^{rte}\,,\end{aligned} \tag{38}$$

where $\Psi(x)$ is the di-gamma function. Here we used $\mathbb{E}[\log x]=\Psi(a)-\log(b)$, valid for Gamma-distributed variables.
Now we can compute the last remaining quantity, using the fact that the mean of a Multinomial-distributed variable z_{ij} of parameters n,ϕ_{ij} is $\mathbb{E}\left[z_{ijk}\right]=n\phi_{ijk}$. Here $n=A_{ij}$, hence

$$\mathbb{E}_q\left[z_{ijk}\right]=A_{ij}\phi_{ijk} \quad . \tag{39}$$

To assess convergence, we then evaluate the lower bound of the variational objective function (ELBO), since this is what we are trying to maximize. This can also be calculated analytically using similar calculations, however it is in general more tedious.

A good reference for this method is Gopalan *et al.* (2015).

1.4 PMF algorithmic updates: EM

The algorithmic updates to optimize the log-likelihood alternate between an E-step updating the variational distributions q_{ijkq} and the M-step updating the parameters u,v,C are shown in Algorithm 1.

Algorithm 1: Expectation-Maximization for Poisson Matrix Factorization

Input: Data A.

Initialize u,v,C randomly.

while *change in L is above some threshold* **do**

E step.

For each pair of nodes such that $A_{ij}>0$, update the variational distributions:

$$q_{ijkq}=\frac{u_{ik}v_{jq}c_{kq}}{\sum_{kq}u_{ik}v_{jq}c_{kq}}$$

M step.

For each node, update the out-going membership parameters:

$$u_{ik}=\frac{\sum_{jq}A_{ij}q_{ijkq}}{\sum_{jq}v_{jq}c_{kq}}$$

update the in-coming membership parameters:

$$v_{jq}=\frac{\sum_{ik}A_{ij}q_{ijkq}}{\sum_{ik}u_{ik}c_{kq}}$$

For each pair k,q, update the affinity matrix parameters:

$$c_{kq}=\frac{\sum_{ij}A_{ij}q_{ijkq}}{\sum_{ij}u_{ik}v_{jq}}$$

end

Output: point-estimates of the parameters (u,v,C).

1.5 PMF algorithmic updates: CAVI

The algorithmic updates to optimize the variational parameters follows a coordinate ascent routine, iteratively optimizing each parameter while holding the others fixed are shown in Algorithm 2.

Algorithm 2: Variational Inference for Poisson Matrix Factorization

Input: Data A.

Initialize α, β, ϕ to the prior with a small random offset.

while *change in ELBO is above some threshold* **do**

For each pair of nodes such that $A_{ij} > 0$, update the multinomial parameters:

$$\phi_{ijk} \propto \exp\left\{\Psi(\alpha_{ik}^{shp}) - \log \alpha_{ik}^{rte} + \Psi(\beta_{jk}^{shp}) - \log \beta_{jk}^{rte}\right\}$$

where the proportionality is such that $\sum_k \phi_{ijk} = 1$.

For each node, update the out-going membership parameters:

$$\begin{aligned} \alpha_{ik}^{shp} &= a + \sum_j A_{ij}\phi_{ijk} \\ \alpha_{ik}^{rte} &= b + \sum_j \frac{\beta_{jk}^{shp}}{\beta_{jk}^{rte}} \end{aligned}$$

update the in-coming membership parameters

$$\begin{aligned} \beta_{jk}^{shp} &= c + \sum_i A_{ij}\phi_{ijk} \\ \beta_{jk}^{rte} &= d + \sum_i \frac{\alpha_{ik}^{shp}}{\alpha_{ik}^{rte}} \end{aligned}$$

end

Output: Variational parameters (α, β, ϕ).

References

B. Ball, B. Karrer, and M. E. Newman, Physical Review E **84**, 036103 (2011).

C. De Bacco, E. A. Power, D. B. Larremore, and C. Moore, Physical Review E **95**, 042317 (2017).

M. Contisciani, E. A. Power, and C. De Bacco, Scientific reports **10**, 1 (2020).

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, Journal of the American Statistical Association **112**, 859 (2017).

P. Gopalan, J. M. Hofman, and D. M. Blei, in *UAI* (2015) pp. 326–335.