L4: Variational inference and mixed-membership models

Caterina De Bacco

May 2, 2025

1. Introduction

In the previous lecture we learned about the variational inference and how it can be used to learn posterior distributions. We saw an example in the gaussian mixture model. Here we see how we can apply to learn mixed-membership communities in networked datasets.

2. Bayesian inference with Variational Inference

We now show an alternative method for parameter inference in mixed-membership models for community detection on network using VI.
In particular, this allows to add priors and calculate full posterior distributions (before we obtained point-estimates).

Recall the likelihood of the model:

$$P(A; u, v, C) = \prod_{ij} \frac{\left(\sum_{k,q} u_{ik} v_{jq} c_{kq}\right)^{A_{ij}}}{A_{ij}!} e^{-\sum_{k,q} u_{ik} v_{jq} c_{kq}} \quad . \tag{1}$$

For simplicity, we assume diagonal affinity matrix C. In this case, we do not need to incorporate the entries c_{kq} 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:

PROPOSITION 1. *The 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:

$$P(z_{ijk}|\theta) = \text{Pois}(z_{ijk}; u_{ik} v_{jk}) \quad , \tag{2}$$

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 joint likelihood:

$$\begin{aligned} P(A, z|\theta) &= P(A|z)P(z|\theta) && (3) \\ &= \prod_{i,j} \delta\left(\sum_k z_{ijk} - A_{ij}\right) \prod_{k=1}^{K} \text{Pois}(z_{ijk}; u_{ik}v_{jk}) \quad , && (4) \end{aligned}$$

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

We then set gamma priors:

$$\begin{aligned} P(u_{ik}|a, b) &\propto u_{ik}^{a}\, e^{-bu_{ik}} && (5) \\ P(v_{ik}|c, d) &\propto v_{ik}^{c}\, e^{-dv_{ik}} \quad . && (6) \end{aligned}$$

Putting all together:

$$P(A, z, u, v) \propto \prod_{i,j} \delta\left(\sum_k z_{ijk} - A_{ij}\right) \prod_{k=1}^{K} \text{Pois}(z_{ijk}; u_{ik}v_{jk}) \prod_{ik} u_{ik}^{a}\, e^{-bu_{ik}} \prod_{ik} v_{ik}^{c}\, e^{-dv_{ik}} \quad , \tag{7}$$

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:

$$q(u, v, z) = \prod_{ik} q(u_{ik}; \alpha_{ik}^{shp}, \alpha_{ik}^{rte}) q(v_{ik}; \beta_{ik}^{shp}, \beta_{ik}^{rte}) \prod_{ij} q(z_{ij}; \phi_{ij}) \quad , \tag{8}$$

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}:

$$\begin{aligned} P(u_{ik}|A, z, v, u_{\setminus ik}) &\propto u_{ik}^{a}\, e^{-bu_{ik}} \prod_j e^{-u_{ik}v_{jk}} u_{ik}^{z_{ijk}} && (9) \\ &= u_{ik}^{a+\sum_j z_{ijk}}\, e^{-(b+\sum_j v_{jk})u_{ik}} \sim \text{Gam}\left(u_{ik}; a + \sum_j z_{ijk}, b + \sum_j v_{jk}\right) && (10) \end{aligned}$$

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:

$$\rho_i = \mathbb{E}_{q(y)} \kappa_i(y) \, , \tag{11}$$

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 $\mathrm{Gam}(\alpha, \beta)$ are $(\alpha - 1, -\beta)$; for a Multinomial $\mathrm{Mult}(n, [\log p_1, \dots, \log p_K])$ and a Categorical distribution $Cat([p_1, \dots, p_K])$ are $(\log p_1, \dots, \log p_K)$.

Hence, we can conclude that the optimal posterior is:

$$\begin{aligned}
q(u_{ik}; \alpha^{shp}_{ik}, \alpha^{rte}_{ik}) &= \mathrm{Gam}\left(u_{ik}; \alpha^{shp}_{ik}, \alpha^{rte}_{ik}\right) && (12)\\
\alpha^{shp}_{ik} &= a + \sum_j \mathbb{E}_q\left[z_{ijk}\right] && (13)\\
\alpha^{rte}_{ik} &= b + \sum_j \mathbb{E}_q\left[v_{jk}\right] = b + \sum_j \frac{\beta^{shp}_{jk}}{\beta^{rte}_{jk}} \quad . && (14)
\end{aligned}$$

Similarly, we have:

$$\begin{aligned}
q(v_{ik}; \beta^{shp}_{ik}, \beta^{rte}_{ik}) &= \mathrm{Gam}\left(v_{ik}; \beta^{shp}_{ik}, \beta^{rte}_{ik}\right) && (15)\\
\beta^{shp}_{jk} &= c + \sum_i \mathbb{E}_q\left[z_{ijk}\right] && (16)\\
\beta^{rte}_{jk} &= d + \sum_i \mathbb{E}_q\left[u_{ik}\right] = d + \sum_j \frac{\alpha^{shp}_{ik}}{\alpha^{rte}_{ik}} \quad , && (17)
\end{aligned}$$

where we used the fact that the mean of a Gamma distribution of shape and rate parameters α and β is α/β.

REMARK 1. *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:

$$\begin{aligned}
P(z_{ij}|A, u, v) &= \frac{\delta\left(\sum_k z_{ijk} - A_{ij}\right) \prod_{k=1}^K \frac{e^{-u_{ik}v_{jk}}(u_{ik}v_{jk})^{z_{ijk}}}{z_{ijk}!}}{P(A|\theta)} && (18)\\
&= \frac{\delta\left(\sum_k z_{ijk} - A_{ij}\right) \prod_{k=1}^K \frac{e^{-u_{ik}v_{jk}}(u_{ik}v_{jk})^{z_{ijk}}}{z_{ijk}!}}{\frac{e^{-\sum_k u_{ik}v_{jk}}\left(\sum_k u_{ik}v_{jk}\right)^{A_{ij}}}{A_{ij}!}} && (19)\\
&\propto \delta\left(\sum_k z_{ijk} - A_{ij}\right) \frac{A_{ij}!}{\prod_k z_{ijk}!} \prod_k \left(u_{ik}v_{jk}\right)^{z_{ijk}} \sim \mathrm{Mult}\left(z_{ij}; (u_{i1}v_{j1}, \dots, u_{iK}v_{jK})\right) && (20)
\end{aligned}$$

Hence we have the the variational posterior:

$$q(z_{ij}; \phi_{ij}) = \mathrm{Mult}\left(z_{ij}; \phi_{ij} = (\phi_{ij1}, \dots, \phi_{ijK})\right) \quad . \qquad (21)$$

Using Eq. (11) we get:

$$\begin{aligned}\log \phi_{ijk} &= \mathbb{E}_q\left[\log u_{ik}\right] + \mathbb{E}_q\left[\log v_{jk}\right] \\ &= \Psi(\alpha^{shp}_{ik}) - \log \alpha^{rte}_{ik} + \Psi(\beta^{shp}_{jk}) - \log \beta^{rte}_{jk}\ , \end{aligned} \tag{22}$$

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

$$\mathbb{E}_q\left[z_{ijk}\right] = A_{ij}\phi_{ijk} \quad . \tag{23}$$

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).

The model outputs posterior estimates. We can use them to assess uncertainty. For instance, in a trade network we find various types of uncertainty, based on the node (country), as shown in fig. 1.

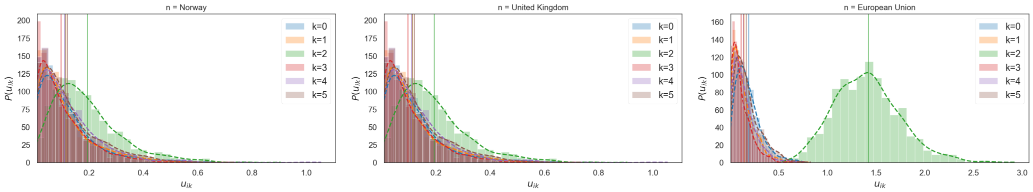

FIGURE 1. Posterior estimates on a trade network. We illustrate the estimates for the membership u_i for various nodes. Some nodes have more mixed-membership, with overlapping posteriors. While, for instance, European union is more clearly placed in one communities, with a more distinct posterior for entry $k = 0$. Vertical lines are expected values $\mathbb{E}[u_{ik}] = \alpha^{shp}_{ik}/\alpha^{rte}_{ik}$.

3. Comparing EM and VI in mixed-membership models.

Now that we have inference updates for both types of models, we can compare them. For this, we can run numerical analysis and use model selection tools to assess performance. We show an example result for a trade network in figs. 2 and 3.

3.1. 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.

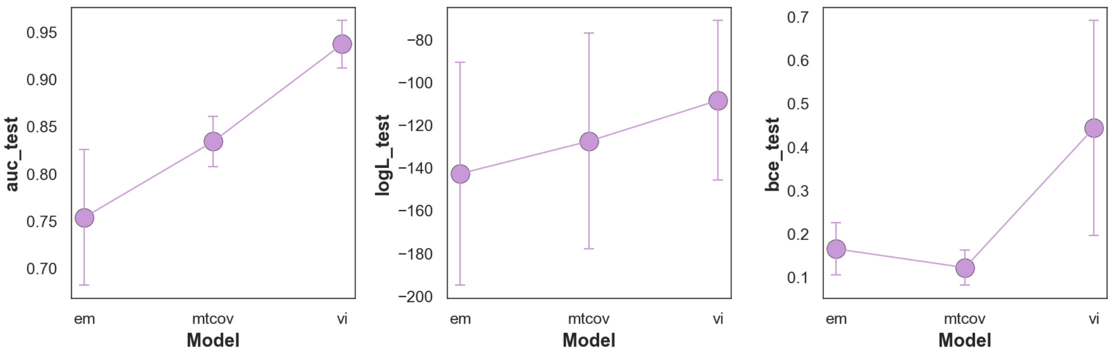

FIGURE 2. Model selection with cross-validation in a WTO trade network. We illustrate the performance on 5-fold link prediction for various algorithms. Markers and error bars are mean and standard deviations over 5 folds.

3.2. 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.

4. Multilayer networks

So far we focused on single layer networks, where edges are of one type. In several contexts, networks nodes connect via multiple types of different edges. For instance, in a social support network people interact via borrowing items, asking for advice, exchanging goods or worshiping together. Each of these types of edges can be represented with a network, and all the networks together as a multilayer network. Here we focus on the case where the set of nodes is the same for all layers, instead the set of edges is different for each layer.

To start, the input data is distinct from the ones used so far. We have now a 3-way tensor A of entries $A_{ij}^{(\alpha)}$, representing the weight of an interaction $i \to j$ of type α. There are L types of interactions, or layers. Alternatively, we can represent each layer with an adjacency matrix $A^{(\alpha)}$, of dimension $N \times N$.

Similarly to what discussed for node attributes and extra information, we can ask the following question.

> *Question*: What makes a multilayer network different from a simple collection of single layer networks?

> *Question*: What information should be shared across layer? What should be layer-specific?

The key here again is that there should be some variable shared across layers, otherwise each layer is independent and we have a simple collection of single layers. Hence, the first assumption we can make is that there are parameters $\Theta = (\theta, \left\{\theta^{(\alpha)}\right\}_{\alpha})$, some are shared

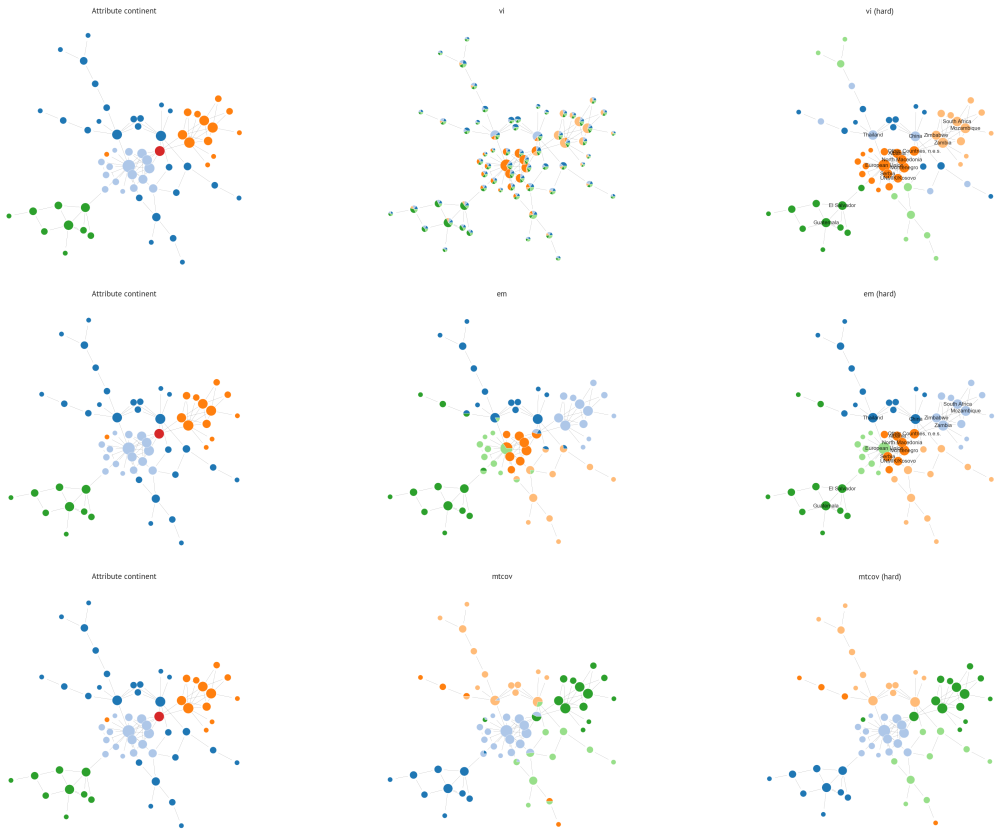

FIGURE 3. Result of various inference algorithms a WTO trade network. In all plots, we used $K = 6$.

and some a layer specific.
Then, we can proceed as usual and further assume that the full joint distribution of the adjacency tensor is factorized, conditioned on the parameters:

$$P(A|\Theta) = \prod_{\alpha} \prod_{ij} P_{\alpha}\left(A_{ij}^{(\alpha)}|\theta, \theta^{(\alpha)}\right) \quad , \tag{24}$$

where the subscript in P_α is to allow different distributions based on the layer. For instance, we can have a layer representing geographical distances, which are real-valued quantities. Another layer may represent observing or not a certain property, which is binary.
In our context of community detection, an example choice is to have communities shared across layers, and then the affinity matrix layer-specific:

$$P(A|\Theta) = \prod_{\alpha} \prod_{ij} \mathrm{Pois}(A_{ij}^{(\alpha)}|\lambda_{ij}^{(\alpha)}) \tag{25}$$

$$\lambda_{ij}^{(\alpha)} = \sum_{k,q} u_{ik}\, v_{jq}\, w_{kq}^{(\alpha)} \quad , \tag{26}$$

so that $\theta = (u, v)$ and $\theta^{(\alpha)} = w^{(\alpha)}$. This is the choice made in Contisciani *et al.* (2025); De Bacco *et al.* (2017); Contisciani *et al.* (2020); Schein *et al.* (2015, 2016). Inference can be done using techniques discussed before, e.g. MLE and EM, VI or Gibbs sampling.

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 $(u, v, C)$.
```

4.1. VI part B: summary

- CAVI updates can be found in closed-form for mixed-membership models of networks
- Choosing between VI, EM etc.. is a model selection task
- Modeling a multilayer network requires asking what type of information we expect layers to share

References

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.

M. Contisciani, M. Hobbhahn, E. A. Power, P. Hennig, and C. De Bacco, PNAS nexus **4**, pgaf005 (2025).

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).

A. Schein, J. Paisley, D. M. Blei, and H. Wallach, in *Proceedings of the 21th ACM SIGKDD International conference on knowledge discovery and data mining* (2015) pp. 1045–1054.

A. Schein, M. Zhou, D. Blei, and H. Wallach, in *International Conference on Machine Learning* (PMLR, 2016) pp. 2810–2819.

Algorithm 2: Variational Inference for Poisson Matrix Factorization

```
Input: Data $A$.
Initialize $\alpha, \beta, \phi$ 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:
        $\phi_{ijk} \propto \exp\left\{\Psi(\alpha^{shp}_{ik}) - \log \alpha^{rte}_{ik} + \Psi(\beta^{shp}_{jk}) - \log \beta^{rte}_{jk}\right\}$
    where the proportionality is such that $\sum_k \phi_{ijk} = 1$.

    For each node, update the out-going membership parameters:
        $\alpha^{shp}_{ik} = a + \sum_j A_{ij}\phi_{ijk}$
        $\alpha^{rte}_{ik} = b + \sum_j \frac{\beta^{shp}_{jk}}{\beta^{rte}_{jk}}$
    update the in-coming membership parameters
        $\beta^{shp}_{jk} = c + \sum_i A_{ij}\phi_{ijk}$
        $\beta^{rte}_{jk} = d + \sum_i \frac{\alpha^{shp}_{ik}}{\alpha^{rte}_{ik}}$
end
Output: Variational parameters $(\alpha, \beta, \phi)$.
```