Multiplex stochastic bloc model for social networks

Avner Bar-Hen,

Pierre Barbillon, Sophie Donnet

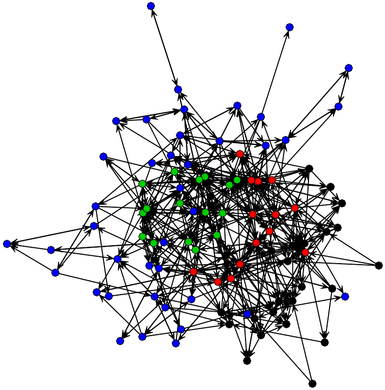
October 10, 2014

Catching up with big fish in the big pond?

Catching up with big fish in the big pond? Multi-level network analysis through linked design
Emmanuel Lazega, Marie-Thérèse Jourda, Lise Mounier, Rafaël Stofer
Social Networks, **30:2**, 159-176 (2008)

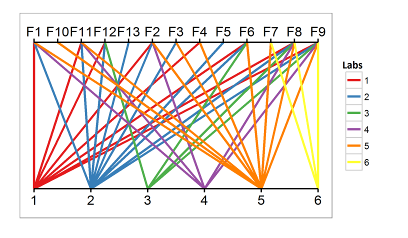
- "Elite", of French cancer researchers at the end of 1990s
- Among the 168 researchers, 128 persons (76%) accepted an interview
- Description of researchers : age, speciality, laboratory, performance, status
- Description of lab : city, # researchers,
- Inter-individual connections
- Inter-lab connections

Researchers' network



Few remarks on the data

- Hierarchical: Labs/Researchers
- Graph
- Sparse
- Missing data
- Not bipartite graph



Fish/Pond

Small/big fish : Indegree centrality

Small/big pond

- indegree centrality in inter-organizational networks
- outdegree (indicating the potential resources to which its director declares having access)
- size

A laboratory is a "big pond" if its values were above the median for at least two of these criteria.

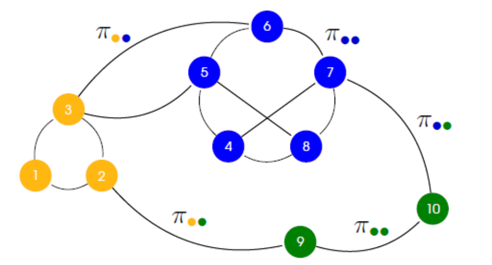
Big Fish in a Big Pond : researcher's indegree centrality must be higher than 5.2, that of the laboratory higher than 2.75 ; the laboratory's outdegree must be higher than 2 and its size higher than 26 researchers.

Stochastic Block Model (Nowicki and Snijders, 2001)

Z_i = q: vertex i belongs to class q (Q classes). Ties are independent given the class memberships.

$$X_{ij}|\{Z_i = q, Z_j = l\} \sim \mathcal{B}(\pi_{ql})$$

Mixture model for graphs: $X_{ij} \sim \Sigma_{q=1}^{Q} \Sigma_{l=1}^{Q} \alpha_q \alpha_l \mathcal{B}(\pi_{ql})$



Computations are performed by **wmixnet**

estimation of α's, π's and latent group (probability of appartenance).

SBM without covariates on Researcher network

- 4 groups selected (group $\neq$ community).
- Study on fully observed data $\Rightarrow$ 95 researchers (from 76 labs).
- Groups are not clearly linked to specialities.

	G 1	G 2	G 3	G 4	Sums
Small Fish	8	0	4	33	45
Big Fish	16	12	17	5	50
Sums	24	12	21	38	95

$$\widehat{\boldsymbol{\pi}} = \begin{pmatrix} 0.08 & 0.08 & 0.01 & 0.00 \\ 0.21 & 0.32 & 0.17 & 0.06 \\ 0.01 & 0.08 & 0.36 & 0.05 \\ 0.01 & 0.02 & 0.04 & 0.04 \end{pmatrix}$$

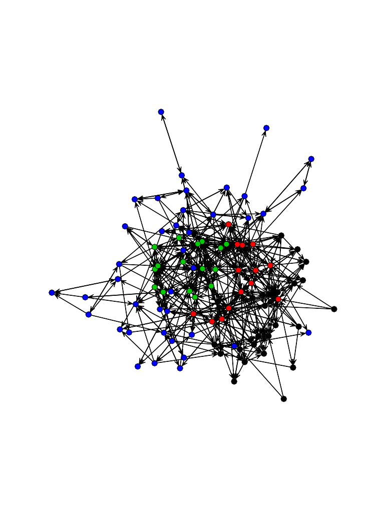


Figure: Researcher-Network with 4 groups

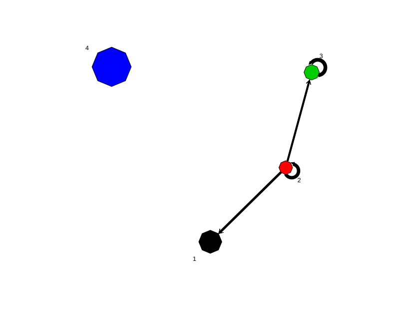


Figure: Meta-Researcher-network (edge displayed if $\pi \geq 0.1$)

Stochastic Block Model with covariates

$Z_i = q$: vertex i belongs to class q (Q classes)

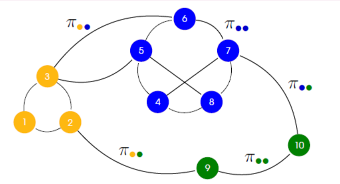
$$X_{ij}|\{Z_i = q, Z_j = l\} \sim \mathscr{B}(\pi_{ql})$$

Covariates can be included, in that case :

$$X_{ij}|\{Z_i = q, Z_j = l, V_{ij}\} \sim \mathscr{B}(g(\mu_{ql} + {\beta_{ql}}^T V_{ij}))$$

where $g(x) = (1 + \exp(-x))^{-1}$ and β may depend on q and l.

Mixture model for graphs: $X_{ij} \sim \Sigma_{q=1}^{Q} \Sigma_{l=1}^{Q} \alpha_q \alpha_l \mathscr{B}(g(\mu_{ql} + {\beta_{ql}}^T V_{ij}))$



In SBM, Ties are independent given the class memberships.

SBM with covariates on Researcher network

- with 5 covariates (describing edge or vertices: sender → receiver):
  1. Same speciality (0/1),
  2. Status of the sender (1/0: Director/not director),
  3. Status of the receiver (1/0: Director/not director),
  4. Lab relation (0/1): sender's lab → receiver's lab,
  5. Lab relation (0/1): receiver's lab → sender's lab.
- Lab network level is taken into account thanks to the covariates !
- Performance is not selected as a relevant covariate to explain edges.
- β does not depend on groups.

$$X_{ij}|\{Z_i = q, Z_j = l, V_{ij}\} \sim \mathscr{B}(g(\mu_{ql} + \beta^T V_{ij}))$$

SBM with covariates on Researcher network: results

- 3 groups selected

	Spe	Stat S	Stat R	Lab S->R	Lab R->S
$\widehat{\boldsymbol{\beta}} =$	(1.16	0.33	0.27	1.75	1.71)

- Lab effect is larger than Status effect
- no directional effect

		G 1	G 2	G 3	Sums
Small Pond	Small Fish	16	0	5	21
	Big Fish	3	2	11	16
Big Pond	Small Fish	24	0	2	26
	Big Fish	12	8	12	32
	Sums	55	10	30	95

Groups obtained in SBM with covariates are quite different from groups in SBM without covariates.

		G 1	G 2	G 3	Sums
Small Pond	Small Fish	16	0	5	21
	Big Fish	3	2	11	16
Big Pond	Small Fish	24	0	2	26
	Big Fish	12	8	12	32
	Sums	55	10	30	95

"Average connections"

$$\begin{pmatrix} 0.03 & 0.06 & 0.02 \\ 0.04 & 0.53 & 0.06 \\ 0.02 & 0.05 & 0.19 \end{pmatrix}$$

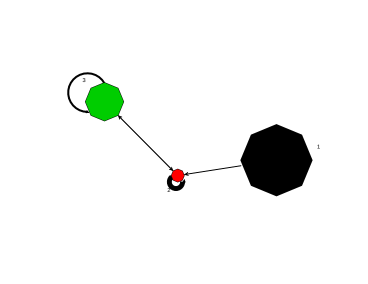


Figure: Meta-Researcher-network (edge displayed if $\pi \geq 0.05$)

SBM conclusions

- 3 researchers groups selected
  (sizes : 55 - 10 - 30).
- Relevant covariates to explain edges:
  - same speciality,
  - Status (R & S),
  - Lab links.
- Performance is NOT relevant.
- Posterior proba to be in a given group are mostly close to 99%.

Multiplex: model

$\forall (i,j) \in \{1,\dots n\}^2, i \neq j, \forall (\delta_{ij}, \delta'_{ij}) \in \{0,1\}^2$,
δ_{ij}: researcher's connections
δ'_{ij}: lab's connections

$$\mathbb{P}(X_{ij} = \delta_{ij}, X'_{ij} = \delta'_{ij}) \quad = \quad \pi_{\delta_{ij}\delta'_{ij}} = \pi_{11}^{\delta_{ij}\delta'_{ij}} \pi_{01}^{(1-\delta_{ij})\delta'_{ij}} \pi_{10}^{\delta_{ij}(1-\delta'_{ij})} \pi_{00}^{(1-\delta_{ij})(1-\delta'_{ij})}$$

where $\pi_{00} + \pi_{11} + \pi_{01} + \pi_{10} = 1$.

$$\mathbb{P}(X_{ij} = \delta_{ij}) = (\pi_{10} + \pi_{11})^{\delta_{ij}} (\pi_{00} + \pi_{01})^{(1-\delta_{ij})}.$$

$$\mathbb{P}(X_{ij} = \delta_{ij} | X'_{ij} = \delta'_{ij}) = \left(\frac{\pi_{1,\delta'_{ij}}}{\pi_{1,\delta'_{ij}} + \pi_{0,\delta'_{ij}}} \right)^{\delta_{ij}} \left(\frac{\pi_{0,\delta'_{ij}}}{\pi_{1,\delta'_{ij}} + \pi_{0,\delta'_{ij}}} \right)^{(1-\delta_{ij})}$$

Components of (X_{ij}, X'_{ij}) are independent if and only if

$\log\left(\frac{\pi_{00}\pi_{11}}{\pi_{10}\pi_{01}}\right) = 0 \Leftrightarrow \pi_{00}\pi_{11} = \pi_{10}\pi_{01}$.

Maximum likelihood estimates:

$$\widehat{\pi}_{00} = \frac{S_{00}}{n(n-1)}, \quad \widehat{\pi}_{01} = \frac{S_{01}}{n(n-1)}, \quad \widehat{\pi}_{10} = \frac{S_{10}}{n(n-1)}, \quad \widehat{\pi}_{11} = \frac{S_{11}}{n(n-1)}.$$

$$\begin{array}{rclrcl}
S_{00}, & = & \sum_{i,j i\neq j}(1-X_{ij})(1-X'_{ij}), & S_{01} & = & \sum_{i,j i\neq j}(1-X_{ij})X'_{ij}, \\
S_{10} & = & \sum_{i,j i\neq j}X_{ij}(1-X'_{ij}), & S_{11} & = & \sum_{i,j i\neq j}X_{ij}X'_{ij}.
\end{array}$$

Multiplex stochastic block model

$$\mathbb{P}(X_{ij} = \delta_{ij}, X'_{ij} = \delta'_{ij}) \quad = \quad \pi_{\delta_{ij}\delta'_{ij}} = \pi_{11}^{\delta_{ij}\delta'_{ij}} \pi_{01}^{(1-\delta_{ij})\delta'_{ij}} \pi_{10}^{\delta_{ij}(1-\delta'_{ij})} \pi_{00}^{(1-\delta_{ij})(1-\delta'_{ij})}$$

$$\begin{aligned}
\mathbb{P}(X_{ij} = \delta_{ij}, X'_{ij} = \delta'_{ij} | Z_i = q, Z_j = l) &= \pi^{ql}_{\delta_{ij},\delta'_{ij}} \\
&= (\pi^{ql}_{11})^{\delta_{ij}\delta'_{ij}} (\pi^{ql}_{01})^{(1-\delta_{ij})\delta'_{ij}} (\pi^{ql}_{10})^{\delta_{ij}(1-\delta'_{ij})} (\pi^{ql}_{00})^{(1-\delta_{ij})} \\
P(Z_i = q) &= \alpha_q
\end{aligned}$$

where $\sum_{q=1}^{Q} \alpha_q = 1$ and $\forall (q, l) \in \{1, \dots Q\}^2$, $\pi^{ql}_{00} + \pi^{ql}_{11} + \pi^{ql}_{01} + \pi^{ql}_{10} = 1$.

Multiplex stochastic block model

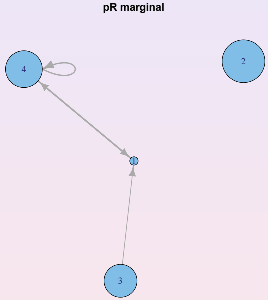
- Introduction of covariates:

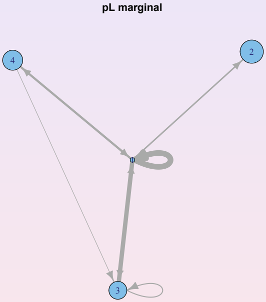
$$\operatorname{logit} \pi^{ql}_{\delta_{ij},\delta'_{ij}} = \mu^{ql}_{\delta^1_{ij}\dots\delta^K_{ij}} + (\beta^{ql}_{\delta^1_{ij}\dots\delta^K_{ij}})^\top \mathbf{y}_{ij}$$

  but the number of parameters drastically increases
- For n large or large number of groups (Q), likelihood is not tractable: variational EM to maximize likelihood (≈ like SBM) Explanation
- Multivariate (K >2) Bernouilli.
- Number of groups chosen with penalized likelihood. Penalization:

$$-\frac{1}{2}\left\{Q^2(2^K-1)\log(Kn(n-1)) + (Q-1)\log n\right\}.$$

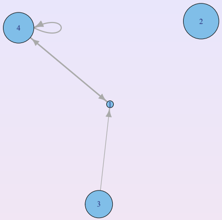
Application to social network: marginals



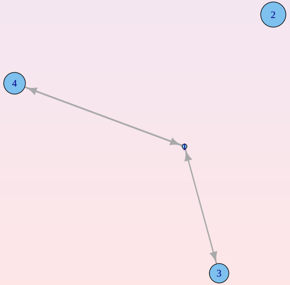


Application to social network: Researchers

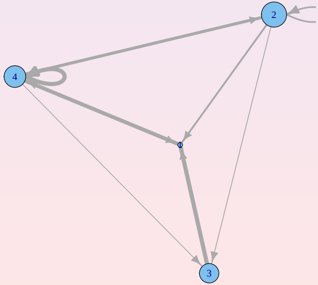
pR marginal



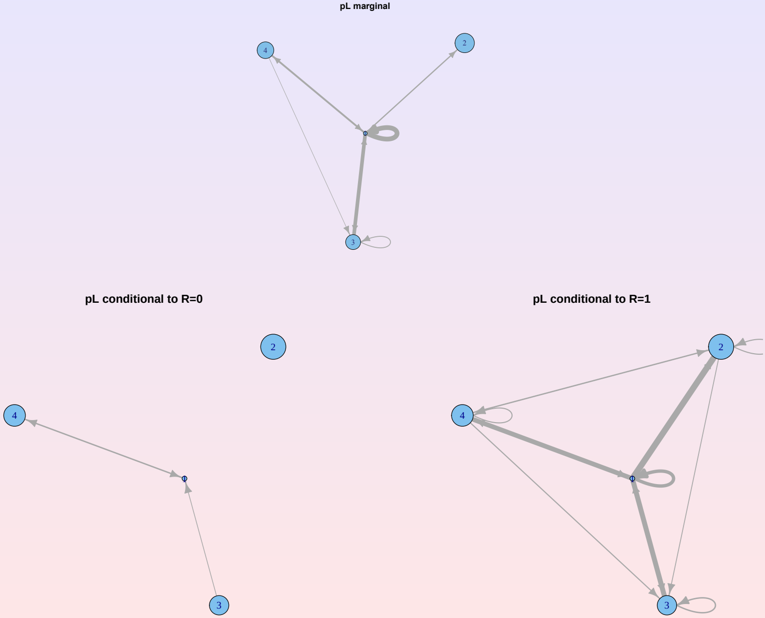
pR conditional to L=0



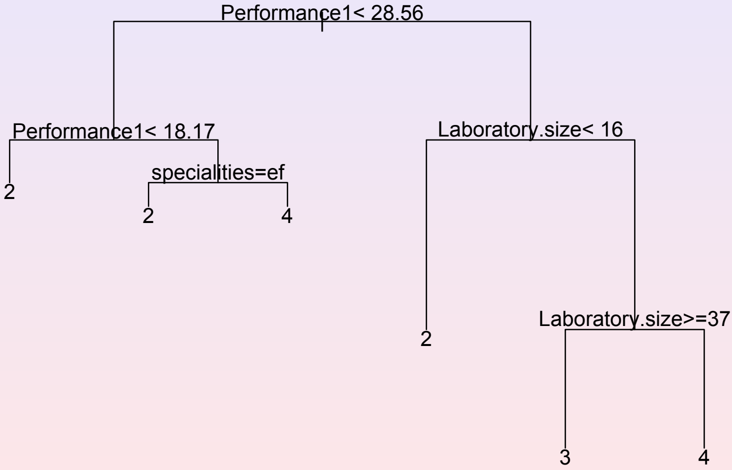
pR conditional to L=1



Application to social network: Labs



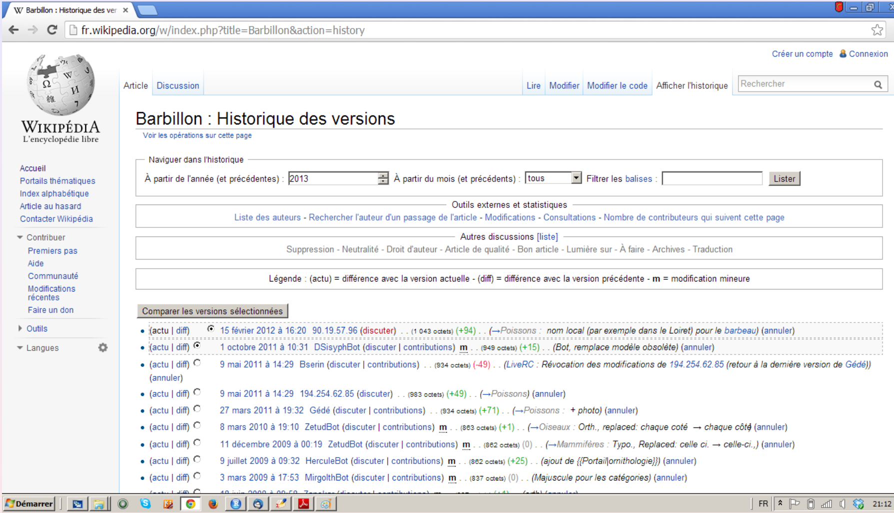
What about the groups?



On going work: Wikipedia

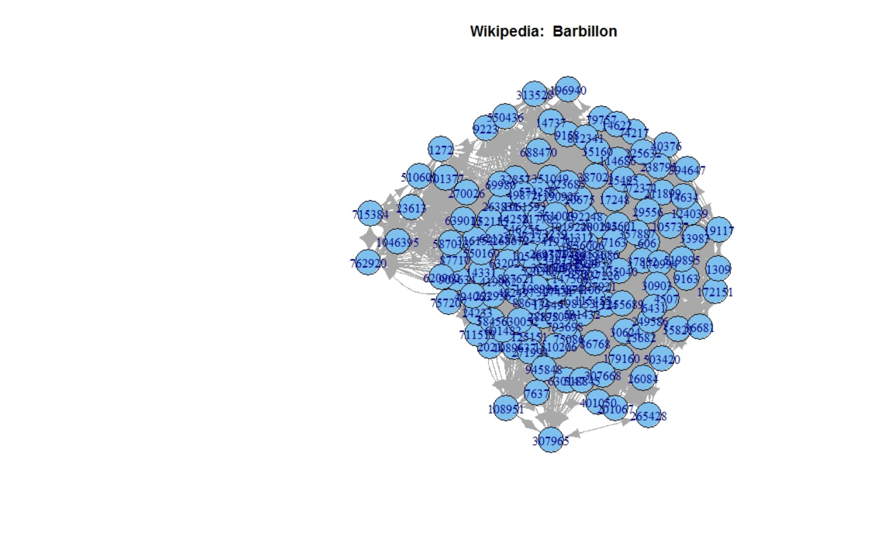
32 links

On going work: Wikipedia



477 users (including 325 that contribute for only one of the 32 articles)

On going work: Wikipedia



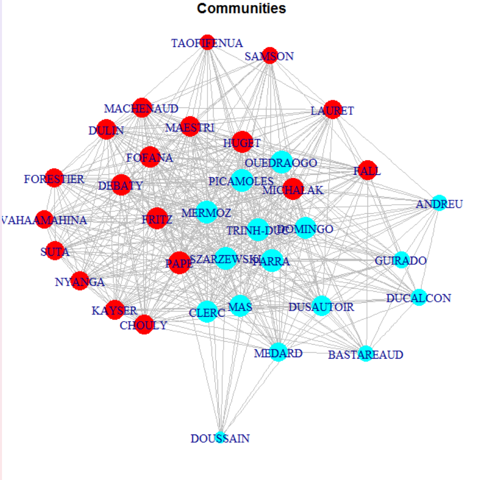
On going work: Collective experience in a rugby team

Is collective experience important for a rugby match

- number of common selection for each pairs of players
- National club for each players
- Opponents, results of the match

On going work: Collective experience in a rugby team

Holders in 2013 (with more than one selection)



Arising questions

- What is the sensitivity of the SBM grouping method if
  - a new edge appears in the graph ?
  - a researcher's lab changes ?
- Are we able to detect the edge that should be set to maximize a criterion such that the mean performance of the researchers, the maximum performance...
- Scalability

Thank you for your attention (and your questions)

Variational EM

Since $X_{ij} | \{Z_i = q, Z_j = l\} \sim \mathscr{B}(\pi_{ql})$ we have

$$X_{ij} \sim \Sigma_{q=1}^{Q} \Sigma_{l=1}^{Q} \alpha_q \alpha_l \mathscr{B}(\pi_{ql})$$

Let $\theta = (\pi_{ij}, \alpha_i)$. We are looking

$$\hat{\theta} = \arg\max_{\theta} \log \mathbb{P}(X, \theta)$$

but $\mathbb{P}(X, \theta) = \sum_Z \mathbb{P}(X, Z, \theta)$ is not tractable.

Classical decomposition(E-M trick)

$$\log \mathbb{P}(X; \theta) = \log \mathbb{P}(X, Z; \theta) - \log \mathbb{P}(Z|X; \theta)$$

$$\mathbb{E}(\log \mathbb{P}(X; \theta)|X) = \log \mathbb{P}(X; \theta) = \mathbb{E}(\log \mathbb{P}(X, Z; \theta)|X) - \mathbb{E}(\log \mathbb{P}(Z|X; \theta)|X)$$

Variational EM

$$\log \mathbb{P}(X;\theta) = \mathbb{E}(\log \mathbb{P}(X,Z;\theta)|X) - \mathbb{E}(\log \mathbb{P}(Z|X;\theta)|X)$$

- E-step: Calculation of $\mathbb{P}(Z|X;\hat{\theta})$ (difficult: forward-backward recursion)
- M-step: $\max_\theta \mathbb{E}(\log \mathbb{P}(X,Z;\theta)|X)$ (similar to MLE)

Variational approximation: replace $\mathbb{P}(Z|X;\theta)$ with approximate distribution $Q(Z)$ ($Q(z)$ within a class of "good" distribution)
For any $Q(z)$

$$\begin{aligned}
\log \mathbb{P}(X;\theta) &\geq \log \mathbb{P}(X;\theta) - KL(Q(Z), P(Z|X)) &(1)\\
&= \mathbb{E}_{Q}(\log \mathbb{P}(X,Z;\theta)) - \mathbb{E}_{Q}(\log Q(Z)) &(2)
\end{aligned}$$

- M-step: $\arg\max_\theta \mathbb{E}_{Q^*}(\log \mathbb{P}(X,Z;\theta))$
- E-step: Replace calculation of $\mathbb{P}(Z|X;\hat{\theta})$ with the search of

$$Q^* = \arg\min_{Q(Z)} KL(Q(Z), P(Z|X))$$

Variational EM

For any $Q(z)$

$$\begin{aligned} \log \mathbb{P}(X;\theta) &\geq \log \mathbb{P}(X;\theta) - KL(Q(Z), P(Z|X)) \qquad (3) \\ &= \mathbb{E}_{Q}(\log \mathbb{P}(X,Z;\theta)) - \mathbb{E}_{Q}(\log Q(Z)) \qquad (4) \end{aligned}$$

$Q(Z)$ within the set of factorisable distributions, ie
$Q(Z_i, Z_j) = Q(Z_i)Q(Z_j)$ (mean field approximation)

Fast to compute but not ML estimates... ▸ Back