The Random Subgraph Model for the Analysis of an Ecclesiastical Network in Merovingian Gaul

Charles Bouveyron

Laboratoire MAP5, UMR CNRS 8145
Université Paris Descartes

This is a joint work with
Y. Jernite, P. Latouche, P. Rivera, L. Jegou & S. Lamassé

Outline

Introduction

The analysis of networks:

- is a recent but increasingly important field in statistical learning,
- with applications in domains ranging from biology to history:
  - biology: analysis of gene regulation processes,
  - social sciences: analysis of political blogs,
  - history: visualization of medieval social networks.

Two main problems are currently well addressed:

- visualization of the networks,
- clustering of the network nodes.

Introduction

The analysis of networks:

- is a recent but increasingly important field in statistical learning,
- with applications in domains ranging from biology to history:
  - biology: analysis of gene regulation processes,
  - social sciences: analysis of political blogs,
  - history: visualization of medieval social networks.

Two main problems are currently well addressed:

- visualization of the networks,
- clustering of the network nodes.

Network comparison:

- is a still emerging problem is statistical learning,
- which is mainly addressed using graph structure comparison,
- but limited to binary networks.

Introduction

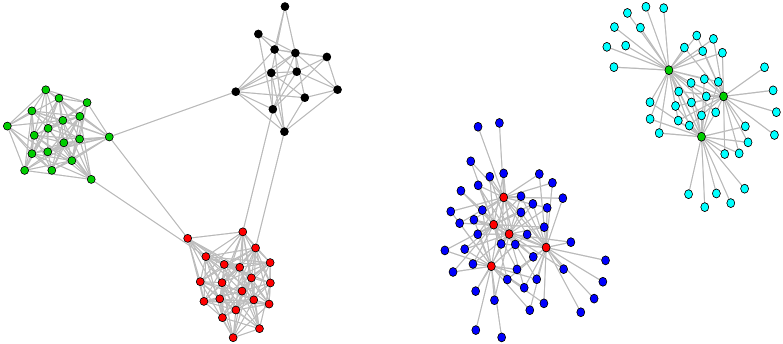


Figure : Clustering of network nodes: communities (left) *vs.* structures with hubs (right).

Introduction

Key works in probabilistic models:

- stochastic block model (SBM) by Nowicki and Snijders (2001),
- latent space model by Hoff, Handcock and Raftery (2002),
- latent cluster model by Handcock, Raftery and Tantrum (2007),
- mixed membership SBM (MMSBM) by Airoldi et al. (2008),
- mixture of experts for LCM by Gormley and Murphy (2010),
- MMSBM for dynamic networks by Xing et al. (2010),
- overlapping SBM (OSBM) by Latouche et al. (2011).

A good overview is given in:

- M. Salter-Townshend, A. White, I. Gollini and T. B. Murphy, "Review of Statistical Network Analysis: Models, Algorithms, and Software", Statistical Analysis and Data Mining, Vol. 5(4), pp. 243–264, 2012.

Introduction: the historical problem

Our colleagues from the LAMOP team were interested in answering the following question:

Was the Church organized in the same way within the different kingdoms in Merovingian Gaul?

Introduction: the historical problem

Our colleagues from the LAMOP team were interested in answering the following question:

Was the Church organized in the same way within the different kingdoms in Merovingian Gaul?

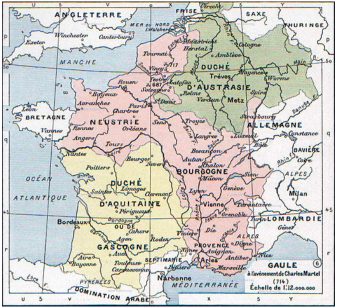
To this end, they have build a relational database:

- from written acts of ecclesiastical councils that took place in Gaul during the 6th century (480-614),
- those acts report who attended (bishops, kings, dukes, priests, monks, ...) and what questions (regarding Church, faith, ...) were discussed,
- they also allowed to characterize the type of relationship between the individuals,
- it took 18 months to build the database.

Introduction: the historical problem

The database contains:

- 1331 individuals (mostly clergymen) who participated to ecclesiastical councils in Gaul between 480 and 614,
- 4 types of relationships between individuals have been identified (positive, negative, variable or neutral),
- each individual belongs to one of the 5 regions of Gaul:
  - 3 kingdoms: Austrasia, Burgundy and Neustria,
  - 2 provinces: Aquitaine and Provence.



- additional information is also available: *social positions*, family relationships, birth and death dates, hold offices, councils dates, ...

Introduction: the historical problem

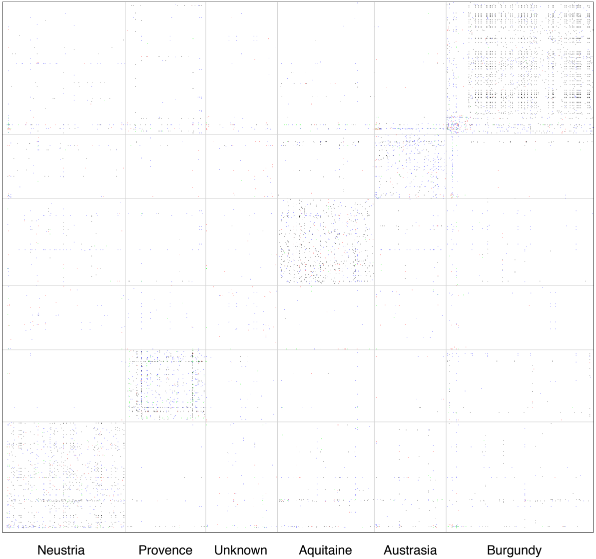


Figure : Adjacency matrix of the ecclesiastical network (sorted by regions).

Introduction

Expected difficulties:

- existing approaches can not analyze networks with categorical edges and a partition into subgraphs,
- comparison of subgraphs has, up to our knowledge, not been addressed in this context,
- a “source effect” is expected due to the overrepresentation of some places (Neustria through “Ten History Book” of Gregory of Tours) or individuals (hagiographies).

Introduction

Expected difficulties:

- existing approaches can not analyze networks with categorical edges and a partition into subgraphs,
- comparison of subgraphs has, up to our knowledge, not been addressed in this context,
- a "source effect" is expected due to the overrepresentation of some places (Neustria through "Ten History Book" of Gregory of Tours) or individuals (hagiographies).

Our approach:

- we consider directed networks with typed (categorical) edges and for which a partition into subgraphs is known,
- we base our comparison on the cluster organization of the subgraphs,
- we propose an extension of SBM which takes into account typed edges and subgraphs,
- subgraph comparison is possible afterward using model parameters.

Outline

The stochastic block model (SBM)

The SBM (Nowicki and Snijders, 2001) model assumes that the network (represented by its adjacency matrix X) is generated as follows:

- each node i is associated with an (unobserved) group among K according to:

  $$Z_i \sim \mathcal{M}(\alpha),$$

  where $\alpha \in [0,1]^K$ and $\sum_{k=1}^{K} \alpha_k = 1$,

The stochastic block model (SBM)

The SBM (Nowicki and Snijders, 2001) model assumes that the network (represented by its adjacency matrix X) is generated as follows:

- each node i is associated with an (unobserved) group among K according to:

  $$Z_i \sim \mathcal{M}(\alpha),$$

  where $\alpha \in [0,1]^K$ and $\sum_{k=1}^K \alpha_k = 1$,

- then, each edge X_{ij} is drawn according to:

  $$X_{ij} | Z_{ik} Z_{jl} = 1 \sim \mathcal{B}(\pi_{kl}),$$

  where $\pi_{kl} \in [0,1]$.

The stochastic block model (SBM)

The SBM (Nowicki and Snijders, 2001) model assumes that the network (represented by its adjacency matrix X) is generated as follows:

- each node i is associated with an (unobserved) group among K according to:

$$Z_i \sim \mathcal{M}(\alpha),$$

  where $\alpha \in [0,1]^K$ and $\sum_{k=1}^K \alpha_k = 1$,

- then, each edge X_{ij} is drawn according to:

$$X_{ij} | Z_{ik} Z_{jl} = 1 \sim \mathcal{B}(\pi_{kl}),$$

  where $\pi_{kl} \in [0,1]$.

- this model is therefore a mixture model:

$$X_{ij} \sim \sum_{k=1}^{K} \sum_{\ell=1}^{K} \alpha_k \alpha_\ell \mathcal{B}(\pi_{kl}).$$

The stochastic block model (SBM)

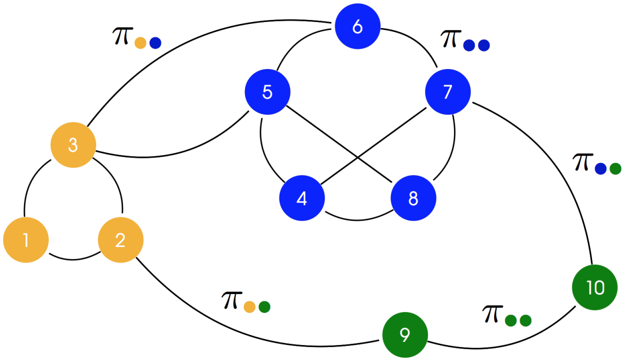


Table : A SBM network.

The stochastic block model (SBM)

Inference of the SBM model (maximum likelihood):

- log-likelihood:

$$\log p(X|\alpha,\Pi) = \log\left\{\sum_{Z} p(X,Z|\alpha,\Pi)\right\},$$

$\hookrightarrow$ K^N terms!

The stochastic block model (SBM)

Inference of the SBM model (maximum likelihood):

- log-likelihood:

$$\log p(X|\alpha, \Pi) = \log \left\{ \sum_{Z} p(X, Z|\alpha, \Pi) \right\},$$

$\hookrightarrow$ K^N terms!

- Expectation Maximization (EM) algorithm requires the knowledge of $p(Z|X, \alpha, \Pi)$,
- Problem: $p(Z|X, \alpha, \Pi)$ is not tractable (no conditional independence)!

The stochastic block model (SBM)

Inference of the SBM model (maximum likelihood):

- log-likelihood:

$$\log p(X|\alpha,\Pi) = \log\left\{\sum_{Z} p(X,Z|\alpha,\Pi)\right\},$$

$\hookrightarrow$ K^N terms!

- Expectation Maximization (EM) algorithm requires the knowledge of $p(Z|X,\alpha,\Pi)$,
- Problem: $p(Z|X,\alpha,\Pi)$ is not tractable (no conditional independence)!

Solutions:

- Variational EM (Daudin et al., 2008) + ICL (Biernacki et al., 2003),
- Variational Bayes EM + $ILvb$ criterion (Latouche et al., 2012).

Outline

The random subgraph model (RSM)

Before the maths, an example of an RSM network:

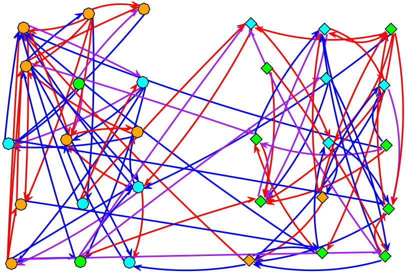


Figure : Example of an RSM network.

We observe:

- the partition of the network into $S = 2$ subgraphs (node form),
- the presence A_{ij} of directed edges between the N nodes,
- the type $X_{ij} \in \{1, ..., C\}$ of the edges ($C = 3$, edge color).

The random subgraph model (RSM)

Before the maths, an example of an RSM network:

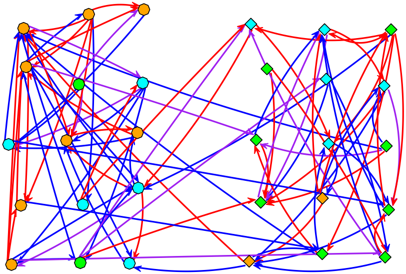


Figure : Example of an RSM network.

We observe:

- the partition of the network into $S = 2$ subgraphs (node form),
- the presence A_{ij} of directed edges between the N nodes,
- the type $X_{ij} \in \{1, ..., C\}$ of the edges ($C = 3$, edge color).

We search:

- a partition of the node into $K = 3$ groups (node color),
- which overlap with the partition into subgraphs.

The random subgraph model (RSM)

The network (represented by its adjacency matrix X) is assumed to be generated as follows:

- the presence of an edge between nodes i and j is such that:

$$A_{ij} \sim \mathcal{B}(\gamma_{s_i s_j})$$

where $s_i \in \{1, ..., S\}$ indicates the (observed) subgraph of node i,

The random subgraph model (RSM)

The network (represented by its adjacency matrix X) is assumed to be generated as follows:

- the presence of an edge between nodes i and j is such that:

$$A_{ij} \sim \mathcal{B}(\gamma_{s_i s_j})$$

  where $s_i \in \{1, ..., S\}$ indicates the (observed) subgraph of node i,

- each node i is as well associated with an (unobserved) group among K according to:

$$Z_i \sim \mathcal{M}(\alpha_{s_i})$$

  where $\alpha_s \in [0, 1]^K$ and $\sum_{k=1}^{K} \alpha_{sk} = 1$,

The random subgraph model (RSM)

The network (represented by its adjacency matrix X) is assumed to be generated as follows:

- the presence of an edge between nodes i and j is such that:

$$A_{ij} \sim \mathcal{B}(\gamma_{s_i s_j})$$

  where $s_i \in \{1, ..., S\}$ indicates the (observed) subgraph of node i,

- each node i is as well associated with an (unobserved) group among K according to:

$$Z_i \sim \mathcal{M}(\alpha_{s_i})$$

  where $\alpha_s \in [0, 1]^K$ and $\sum_{k=1}^{K} \alpha_{sk} = 1$,

- each edge X_{ij} can be finally of C different (observed) types and such that:

$$X_{ij} | A_{ij} Z_{ik} Z_{jl} = 1 \sim \mathcal{M}(\Pi_{kl})$$

  where $\Pi_{kl} \in [0, 1]^C$ and $\sum_{c=1}^{C} \Pi_{klc} = 1$.

The random subgraph model (RSM)

Notations	Description
$\mathbf{X}$	Adjacency matrix. $X_{ij} \in \{0, \dots, C\}$ indicates the edge type
$\mathbf{A}$	Binary matrix. $A_{ij} = 1$ indicates the presence of an edge
$\mathbf{Z}$	Binary matrix. $Z_{ik} = 1$ indicates that i belongs to cluster k
N	Number of vertices in the network
K	Number of latent clusters
S	Number of subgraphs
C	Number of edge types
$\boldsymbol{\alpha}$	α_{sk} is the proportion of cluster k in subgraph s
$\mathbf{\Pi}$	Π_{klc} is the probability of having an edge of type c between vertices of clusters k and l
$\boldsymbol{\gamma}$	γ_{rs} probability of having an edge between vertices of subgraphs r and s

Table : Summary of the notations.

The random subgraph model (RSM)

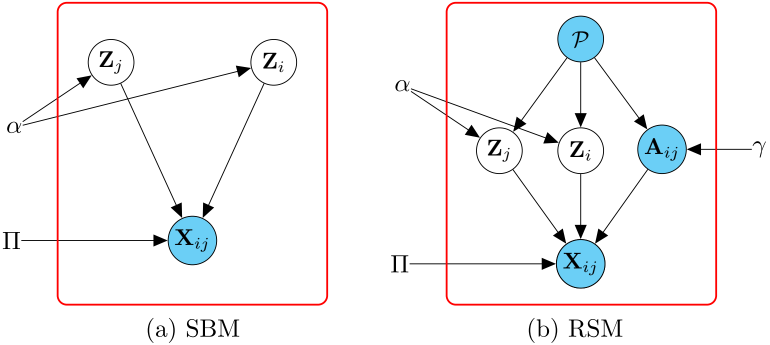


Figure : SBM model *vs.* RSM model.

The random subgraph model (RSM)

Remark 1:

- the RSM model separates the roles of the known partition and the latent clusters,
- this was motivated by historical assumptions on the creation of relationships during the 6th century,
- indeed, the possibilities of connection were preponderant over the type of connection and mainly dependent on the geography.

The random subgraph model (RSM)

Remark 1:

- the RSM model separates the roles of the known partition and the latent clusters,
- this was motivated by historical assumptions on the creation of relationships during the 6th century,
- indeed, the possibilities of connection were preponderant over the type of connection and mainly dependent on the geography.

Remark 2:

- an alternative approach would consist in allowing X_{ij} to directly depend on both the latent clusters and the partition,
- however, this would dramatically increase the number of model parameters ($K^2S^2(C+1)+SK$ instead of S^2+K^2C+SK),
- if $S=6$, $K=6$ and $C=4$, then the alternative approach has 6 516 parameters while RSM has only 216.

The random subgraph model (RSM)

We consider a Bayesian framework:

- the previous model is fully defined by its joint distribution:

$$p(X, A, Z|\alpha, \gamma, \Pi) = p(X|A, Z, \Pi)p(A|\gamma)p(Z|\alpha),$$

- which we complete with conjuguate prior distributions for model parameters:
  - the prior distribution for α is:

$$p(\gamma_{rs}) = Beta(a_{rs}, b_{rs}),$$

  - the prior distribution for γ is:

$$p(\alpha_s) = Dir(\chi_s),$$

  - the prior distribution for Π is:

$$p(\Pi_{kl}) = Dir(\Xi_{kl}).$$

The random subgraph model (RSM)

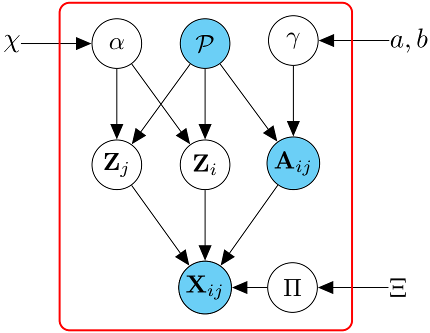


Figure : A graphical representation of the RSM model.

Outline

Model inference

Due to the Bayesian framework introduces above:

- we aim at estimating the posterior distribution $p(Z, \alpha, \gamma, \Pi | X, A)$, which in turn will allow us to compute MAP estimates of Z and (α, γ, Π),
- as expected, this distribution is not tractable and approximate inference procedures are required,
- the use of MCMC methods is obviously an option but MCMC methods have a poor scaling with sample sizes.

Model inference

Due to the Bayesian framework introduces above:

- we aim at estimating the posterior distribution $p(Z, \alpha, \gamma, \Pi | X, A)$, which in turn will allow us to compute MAP estimates of Z and (α, γ, Π),
- as expected, this distribution is not tractable and approximate inference procedures are required,
- the use of MCMC methods is obviously an option but MCMC methods have a poor scaling with sample sizes.

We chose to use variational approaches:

- because they allow to deal with large networks ($N > 1000$),
- recent theoretical results (Celisse et al., 2012; Mariadassou and Matias, 2013) gave new insights about convergence properties of variational approaches in this context.

The VBEM algorithm

We aim at estimating the posterior distribution $p(Z,\theta|X)$:

- we use the decomposition of the marginal log-likelihood:

$$\log(p(X)) = \mathcal{L}(q(Z,\theta)) + KL(q(Z,\theta)||p(Z,\theta|X)),$$

  where:
  - $\mathcal{L}(q(Z,\theta)) = \sum_Z \int_\theta q(Z,\theta)\log(p(X,Z,\theta)/q(Z,\theta))d\theta$ is a lower bound of the log-likelihood,
  - $KL(q(Z,\theta)||p(Z,\theta|X)) = -\sum_Z \int_\theta q(Z,\theta)\log(p(Z,\theta|X)/q(Z,\theta))d\theta$ is the KL divergence between $q(Z,\theta)$ and $p(Z,\theta|X)$.

- we also assume that q factorizes over Z and θ:

$$q(Z,\theta) = \prod_i q_i(Z_i)q_\theta(\theta).$$

The VBEM algorithm

We aim at estimating the posterior distribution $p(Z,\theta|X)$:

- we use the decomposition of the marginal log-likelihood:

$$\log(p(X)) = \mathcal{L}(q(Z,\theta)) + KL(q(Z,\theta)||p(Z,\theta|X)),$$

  where:
  - $\mathcal{L}(q(Z,\theta)) = \sum_Z \int_\theta q(Z,\theta)\log(p(X,Z,\theta)/q(Z,\theta))d\theta$ is a lower bound of the log-likelihood,
  - $KL(q(Z,\theta)||p(Z,\theta|X)) = -\sum_Z \int_\theta q(Z,\theta)\log(p(Z,\theta|X)/q(Z,\theta))d\theta$ is the KL divergence between $q(Z,\theta)$ and $p(Z,\theta|X)$.

- we also assume that q factorizes over Z and θ:

$$q(Z,\theta) = \prod_i q_i(Z_i) q_\theta(\theta).$$

The VBEM algorithm:

- VB-E step: $q_\theta(\theta)$ is fixed and $\mathcal{L}$ is maximized over the q_i
  $\Rightarrow \log q_j^*(Z_j) = E_{i\neq j,\theta}[\log p(X,Z,\theta)] + c$
- VB-M step: all $q_i(Z_i)$ are now fixed and $\mathcal{L}$ is maximized over q_θ
  $\Rightarrow \log q_\theta^*(\theta) = E_Z[\log p(X,Z,\theta)] + c$

The VBEM algorithm for RSM

Variational Bayesian inference in our case:

- we aim at approximating the posterior distribution $p(Z, \alpha, \gamma, \Pi | X, A)$
- we therefore search the approximation $q(Z, \alpha, \gamma, \Pi)$ which maximizes $\mathcal{L}(q)$ where:

$$\log p(X, A) = \mathcal{L}(q) + KL(q || p(.|X, A)),$$

- and q is assumed to factorize as follows:

$$q(Z, \alpha, \gamma, \Pi) = \prod q(Z_i) \prod q(\alpha_s) \prod q(\gamma_{st}) \prod q(\Pi_{kl}).$$

The VBEM algorithm for RSM:

- E step: compute the update parameter τ_i for $q(Z_i)$,
- M step: compute the update parameters χ, γ, Ξ for respectively $q(\alpha_s)$, $q(\gamma_{st})$ and $q(\Pi_{kl})$.

The VBEM algorithm for RSM: the M step

The M step of the VBEM algorithm: the VBEM update step for the distributions $q(\alpha_s)$ is:

$$\begin{aligned}\log q^*(\alpha_s) &= E_{Z,\alpha^{\setminus s},\gamma,\Pi}[\log p(X, A, Z, \alpha, \gamma, \Pi)] + c \\ &= \sum_{k=1}^{K} \log(\alpha_{sk}) \left\{ \chi^0_{sk} + \sum_{i=1}^{N} \delta(r_i = s)\tau_{ik} - 1 \right\} + c,\end{aligned}$$

The VBEM algorithm for RSM: the M step

The M step of the VBEM algorithm: the VBEM update step for the distributions $q(\alpha_s)$ is:

$$\begin{aligned}\log q^*(\alpha_s) &= E_{Z,\alpha^{\setminus s},\gamma,\Pi}[\log p(X, A, Z, \alpha, \gamma, \Pi)] + c \\ &= \sum_{k=1}^{K} \log(\alpha_{sk}) \left\{ \chi^0_{sk} + \sum_{i=1}^{N} \delta(r_i = s)\tau_{ik} - 1 \right\} + c,\end{aligned}$$

which is the functional form for a Dirichlet distribution:

$$q(\alpha_s) = \mathrm{Dir}(\alpha_s; \chi_s), \forall s \in \{1, \ldots, S\}$$

where $\chi_{sk} = \chi^0_{sk} + \sum_{i=1}^{N} \delta(r_i = s)\tau_{ik}, \forall k \in \{1, \ldots, K\}$.

The VBEM algorithm for RSM: the M step

The M step of the VBEM algorithm: the VBEM update step for the distributions $q(\alpha_s)$, $q(\gamma_{st})$ and $q(\Pi_{kl})$ are:

- $q(\alpha_s) = \mathrm{Dir}(\alpha_s; \chi_s), \forall s \in \{1, \ldots, S\}$,
- $q(\gamma_{rs}) = \mathrm{Beta}(\gamma_{rs}; a_{rs}, b_{rs}), \forall (r, s) \in \{1, \ldots, S\}^2$,
- $q(\Pi_{kl}) = \mathrm{Dir}(\Pi_{kl}; \Xi_{kl}), \forall (k, l) \in \{1, \ldots, K\}^2$,

where:

- $\chi_{sk} = \chi^0_{sk} + \sum_{i=1}^{N} \delta(r_i = s)\tau_{ik}, \forall k \in \{1, \ldots, K\}$,
- $a_{rs} = a^0_{rs} + \sum_{r_i=r, r_j=s}(A_{ij}),\ b_{rs} = b^0_{rs} + \sum_{r_i=r, r_j=s}(1 - A_{ij})$,
- $\Xi_{klc} = \Xi^0_{klc} + \sum_{i \neq j}^{N} \delta(X_{ij} = c)\tau_{ik}\tau_{jl}, \forall c \in \{1, \ldots, C\}$.

The VBEM algorithm for RSM: the E step

The E step of the VBEM algorithm: the VBEM update step for the distribution $q(Z_i)$ is given by:

$$\log q^*(Z_i) = E_{Z^{\setminus i},\alpha,\gamma,\Pi}[\log p(X, A, Z, \alpha, \gamma, \Pi)] + c$$

which implies that

$$q(Z_i) = \mathcal{M}(Z_i; 1, \tau_i),\ \forall i = 1, ..., N$$

where

$$\begin{aligned}
\tau_{ik} \propto \exp & \left(\psi(\chi_{r_i,k}) - \psi(\sum_{l=1}^{K} \chi_{r_i,l}) \right) \\
& + \exp \left\{ \sum_{j \neq i}^{N} \sum_{c=1}^{C} \sum_{l=1}^{K} \delta(X_{ij} = c) \tau_{jl} \left(\psi(\Xi_{klc}) - \psi(\sum_{u=1}^{C} \Xi_{klu}) \right) \right\} \\
& + \exp \left\{ \sum_{j \neq i}^{N} \sum_{c=1}^{C} \sum_{l=1}^{K} \delta(X_{ji} = c) \tau_{jl} \left(\psi(\Xi_{lkc}) - \psi(\sum_{u=1}^{C} \Xi_{lku}) \right) \right\}.
\end{aligned}$$

Initialization and choice of K

Initialization of the VBEM algorithm:

- the VBEM is known to be sensitive to its initialization,
- we propose a strategy based on several k-means algorithms with a specific distance:

$$d(i,j) = \sum_{h=1}^{N} \delta(X_{ih} \neq X_{jh}) A_{ih} A_{jh} + \sum_{h=1}^{N} \delta(X_{hi} \neq X_{hj}) A_{hi} A_{hj}.$$

Initialization and choice of K

Initialization of the VBEM algorithm:

- the VBEM is known to be sensitive to its initialization,
- we propose a strategy based on several k-means algorithms with a specific distance:

$$d(i,j) = \sum_{h=1}^{N} \delta(X_{ih} \neq X_{jh}) A_{ih} A_{jh} + \sum_{h=1}^{N} \delta(X_{hi} \neq X_{hj}) A_{hi} A_{hj}.$$

Choice of the number K of groups:

- once the VBEM algorithm has converged, the lower bound $\mathcal{L}(q)$ is a good approximation of the integrated log-likelihood $\log p(X, A)$,
- we thus can use $\mathcal{L}(q)$ as a model selection criterion for choosing K,
- if computed right after the M step,

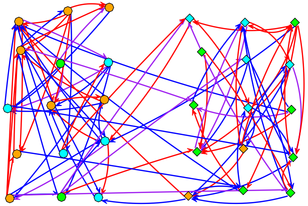
$$\mathcal{L}(q) = \sum_{r,s}^{S} \log\left(\frac{B(a_{rs}, b_{rs})}{B(a^0_{rs}, b^0_{rs})}\right) + \sum_{s=1}^{S} \log\left(\frac{C(\boldsymbol{\chi}_s)}{C(\boldsymbol{\chi}^0_s)}\right) + \sum_{k,l}^{K} \log\left(\frac{C(\boldsymbol{\Xi}_{kl})}{C(\boldsymbol{\Xi}^0_{kl})}\right) - \sum_{i=1}^{N}\sum_{k=1}^{K} \tau_{ik} \log(\tau_{ik}).$$

Outline

Experimental setup

We considered 3 different situations:

- S1 : network without subgraphs and with a preponderant proportion of edges of type 1,
- S2 : network without subgraphs and with balanced proportions of the three edge types,
- S3 : network with 3 subgraphs and with balanced proportions of the three edge types.



Global setup:

- in all cases, the number of (unobserved) groups is $K = 3$ and the network size is $N = 100$,
- we use the adjusted Rand index (ARI) for evaluating the clustering quality (and thus the model fitting).

Choice of the number K of groups

First, a model selection study:

- we aim at validating the use of $\mathcal{L}(q)$ as model selection criteria,
- we simulated 50 RSM networks according to scenario 1 and with $N = 100$,
- and applied our VB-EM algorithm for different values of K ($K = 2, ..., 5$),
- the actual value of K is $K = 3$.

Choice of the number K of groups

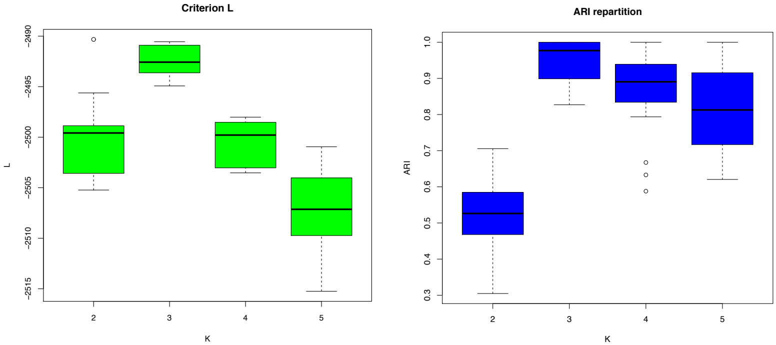


Table : Lower bound $\mathcal{L}$ and ARI averaged over 50 networks simulated according to the RSM model.

Comparison with other SBM-based approaches

Second, a comparison with other SBM-based methods:

- binary SBM: the original SBM algorithm was applied on a collapsed version of the data (only the presence of edges); the mixer package was used,
- binary SBM (type 1, 2 or 3): the original SBM algorithm was applied on a collapsed version of the data (only edges of type 1, 2 or 3); the mixer package was used,
- typed SBM: we had to implement the categorical version of SBM since it is not available in existing software; this version of SBM will be available in mixer soon,
- the studied methods were applied to the the three scenarii and results are averaged over 50 networks.

Comparison with other SBM-based approaches

Method	Scenario 1	Scenario 2	Scenario 3
binary SBM (presence)	0.001 ± 0.012	0.001 ± 0.013	0.239 ± 0.061
binary SBM (type 1)	0.976 ± 0.071	0.494 ± 0.233	-0.372 ± 0.262
binary SBM (type 2)	0.001 ± 0.006	-0.003 ± 0.006	0.179 ± 0.097
binary SBM (type 3)	0.959 ± 0.121	0.519 ± 0.219	0.367 ± 0.244
Typed SBM	0.694 ± 0.232	0.472 ± 0.339	0.360 ± 0.162
RSM	$\mathbf{1.000 \pm 0.000}$	$\mathbf{0.981 \pm 0.056}$	$\mathbf{0.939 \pm 0.097}$

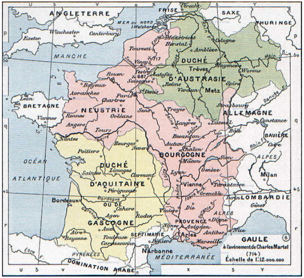
Table : ARI averaged over 50 networks simulated according to the three considered situations.

Outline

The ecclesiastical network

The data:

- 1331 individuals (mostly clergymen) who participated to ecclesiastical councils in Gaul between 480 and 614,
- 4 types of relationships between individuals have been identified (positive, negative, variable or neutral),
- each individual belongs to one of the 5 regions (3 kingdoms et 2 provinces).



Our modeling allows a multi-level analysis:

- Z allows to characterize the found clusters through social positions of the individuals,
- parameter Π describes the relations between the found clusters,
- parameter γ describes the connections between the subgraphs,
- parameter α describes the cluster repartition in the subgraphs.

RSM results: the latent clusters

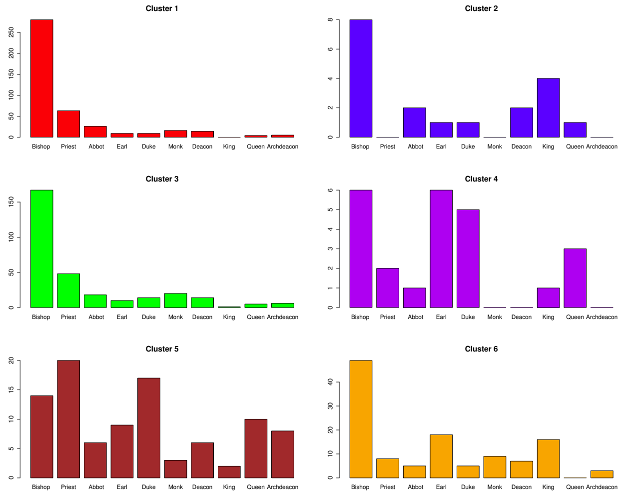


Figure : Characterization of the $K = 6$ clusters found by RSM.

RSM results: the latent clusters

The latent clusters from the historical point of view:

- clusters 1 and 3 correspond to local, provincial of diocesan councils, mostly interested in local issues (ex: council of Arles, 554),
- clusters 2 and 6 correspond to councils dedicated to political questions, usually convened by a king (ex: Orleans, 511),
- clusters 4 and 5 correspond to aristocratic assemblies, where queens and duke and earls are present (ex: Orleans, 529).

RSM results: the relationships between clusters

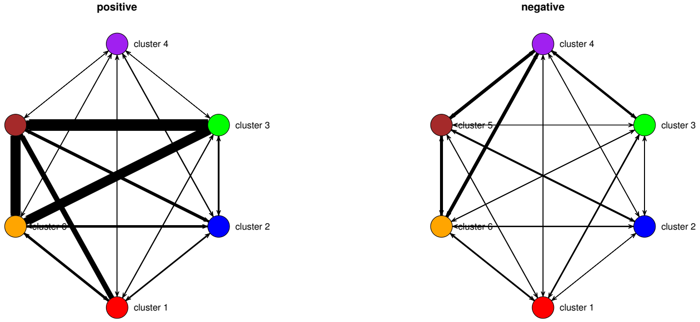


Figure : Characterization of the relationships between clusters (parameter Π).

RSM results: the relationships between clusters

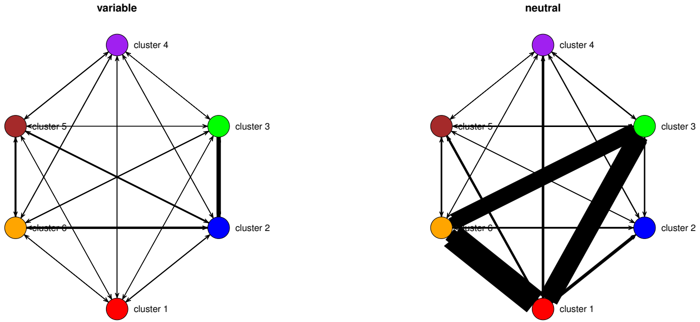


Figure : Characterization of the relationships between clusters (parameter Π).

RSM results: the relationships between clusters

The clusters relationships from the historical point of view:

- positive relations between clusters 3, 5 and 6 mainly corresponds to personal friendships between bishops (source effect),
- negative and variable relations betweens clusters 4, 5 and 6 report the conflicts in the hierarchy of the power,
- neutral relations between clusters 1, 3 and 6 were expected because they deal with different issues (local / political).

RSM results: the relationships between regions

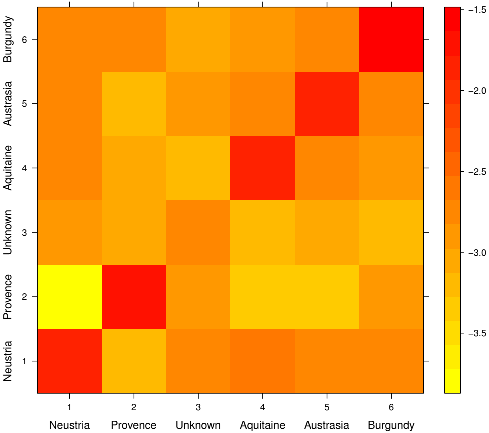


Figure : Characterization of the relationships between the regions (parameter γ in log scale).

RSM results: comparison of the regions

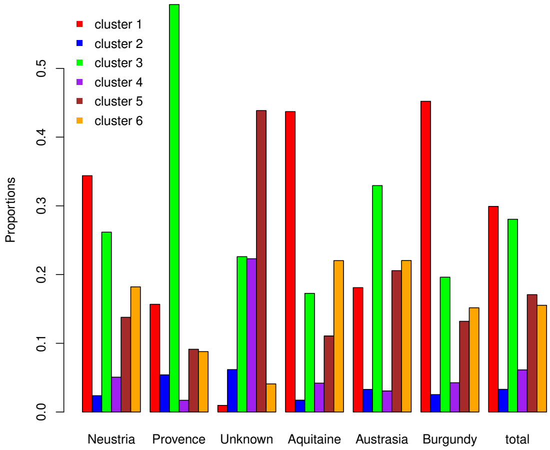


Figure : Characterization of regions through cluster repartition (parameter α).

RSM results: comparison of the regions

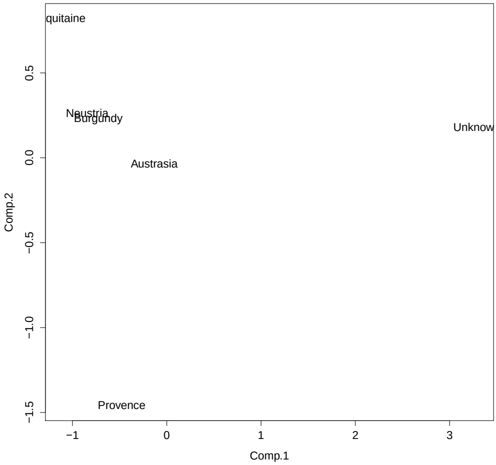


Figure : PCA for compositional data on the parameter α.

Outline

A maritime flow network

We considered the data from Ducruet (2013):

- data from Lloyd's List (Voyage Record) covering the period October-November 2004,
- huge work to extract from paper versions and complement the lacks (capacity, ...),
- the data contains 28277 vessels between 1815 ports,
- 4 types of relations between ports are considered: liquid bulk, passengers, containers and solid bulk.

The softwares:

- package Mixer for R which implements SBM,
- package Rambo for R which implements RSM.

A maritime flow network

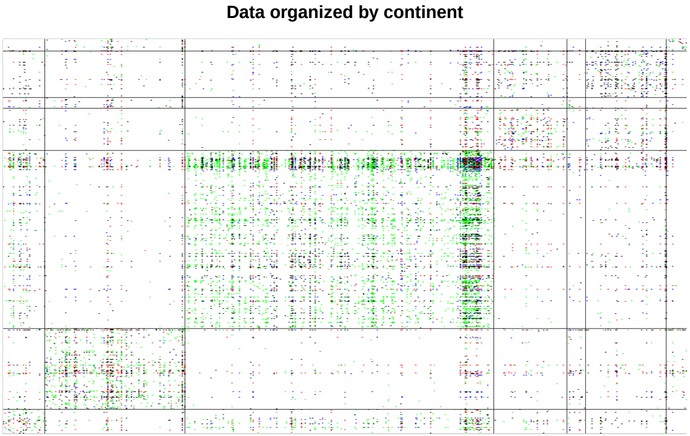


Figure : Adjacency matrix organized by continent with categorical edges (containers, solid bulk, liquid bulk and passengers).

Results of SBM

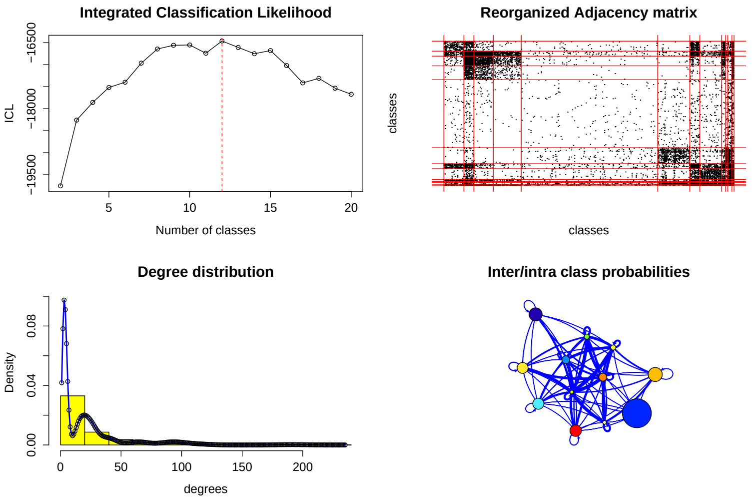


Figure : Output from the mixer package (SBM).

Results of SBM

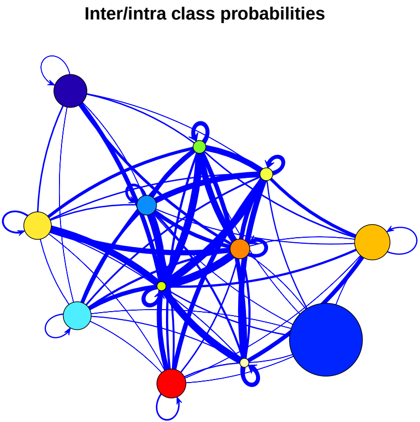


Figure : Connection probabilities between groups (matrix $\mathbf{\Pi}$).

Results of SBM

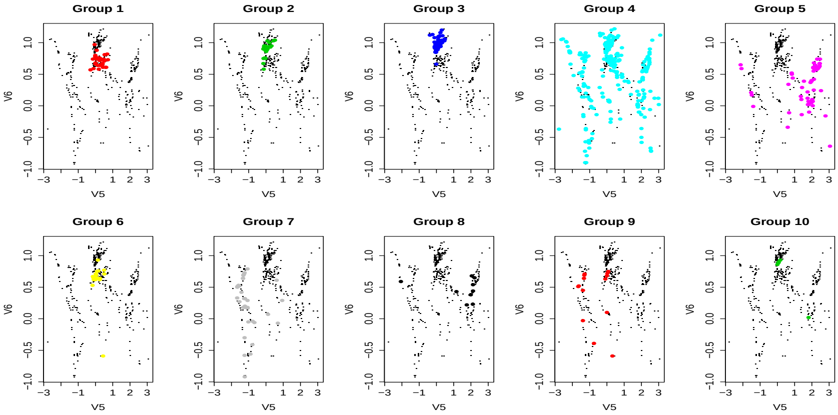


Figure : Geography of the clusters.

Results of SBM

Figure : Adjacency matrix organized according to the SBM groups (containers, solid bulk, liquid bulk and passengers).

Results of RSM

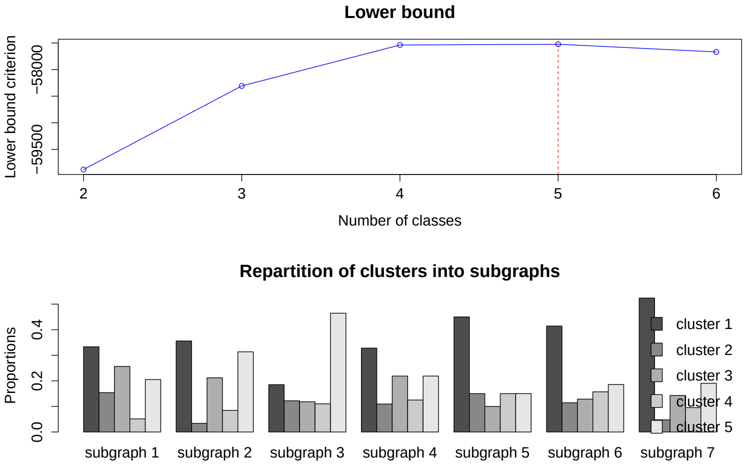


Figure : Output of the Rambo package (RSM).

Results of RSM

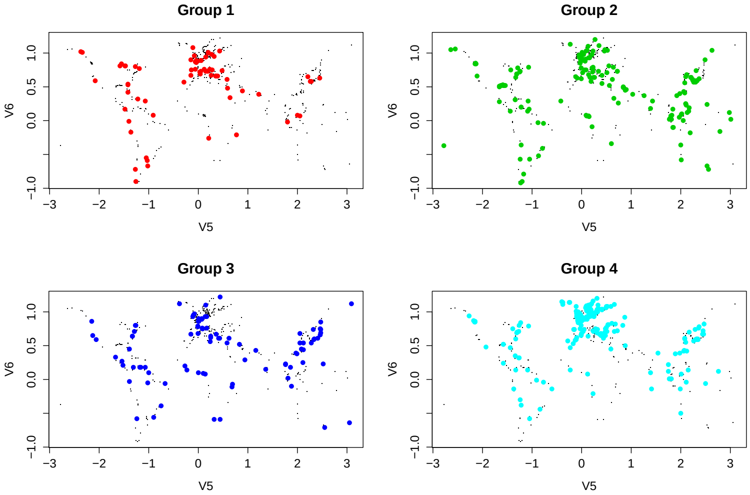


Figure : Geography of the clusters.

Results of RSM

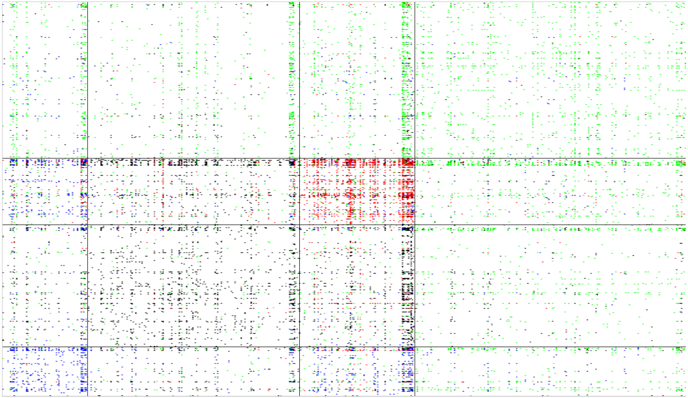


Figure : Adjacency matrix organized according to the RSM groups (containers, solid bulk, liquid bulk and passengers).

Outline

Conclusion

Our contribution:

- the model takes into account an existing partition into subgraphs,
- this modeling allows afterward a comparison of the subgraphs,
- inference is done in a Bayesian framework using a VBEM algorithm.

Interesting problems to address:

- temporality of the network (evolution of relations, offices or social positions),
- visualization of this kind of networks.

Conclusion

Our contribution:

- the model takes into account an existing partition into subgraphs,
- this modeling allows afterward a comparison of the subgraphs,
- inference is done in a Bayesian framework using a VBEM algorithm.

Interesting problems to address:

- temporality of the network (evolution of relations, offices or social positions),
- visualization of this kind of networks.

Software:

package Rambo for the R software is available on the CRAN

Publication:

C. Bouveyron, L. Jegou, Y. Jernite, S. Lamassé, P. Latouche & P. Rivera, *The random subgraph model for the analysis of an ecclesiastical network in merovingian Gaul*, The Annals of Applied Statistics, 8(1), 377-405, 2014.

`http://arxiv.org/abs/1212.5497`

The EM, VEM and VBEM algorithms

First, it necessary to write the log-likelihood as:

$$\log(p(X|\theta)) = \mathcal{L}(q(Z);\theta) + KL(q(Z)||p(Z|X,\theta)),$$

where:

- $\mathcal{L}(q(Z);\theta) = \sum_Z q(Z)\log(p(X,Z|\theta)/q(Z))$ is a lower bound of the log-likelihood,
- $KL(q(Z)||p(Z|X,\theta)) = -\sum_Z q(Z)\log(p(Z|X,\theta)/q(Z))$ is the KL divergence between $q(Z)$ and $p(Z|X,\theta)$.

The EM algorithm:

- E step: θ is fixed and $\mathcal{L}$ is maximized over $q \Rightarrow q^*(Z) = p(Z|X,\theta)$
- M step: $\mathcal{L}(q^*(Z),\theta^{old})$ is now maximized over θ

$$\begin{aligned}\mathcal{L}(q^*(Z),\theta^{old}) &= \sum_Z p(Z|X,\theta^{old})\log(p(X,Z|\theta)/p(Z|X,\theta^{old}))\\ &= E[\log(p(X,Z|\theta)|\theta^{old}] + c.\end{aligned}$$

The EM, VEM and VBEM algorithms

The variational approach:

- let us now suppose that $p(X, Z|\theta)$ is, for some reason, intractable,
- the variational approach restrict the range of functions for q such that the problem is tractable,
- a popular variational approximation is to assume that q factorizes:

$$q(Z) = \prod_i q_i(Z_i).$$

The VEM algorithm:

- V-E step: θ is fixed and $\mathcal{L}$ is maximized over $q \Rightarrow$ $\log q_j^*(Z_j) = E_{i \neq j}[\log p(X, Z|\theta)] + c$
- V-M step: $\mathcal{L}(q^*(Z), \theta^{old})$ is now maximized over θ

The EM, VEM and VBEM algorithms

We consider now the Bayesian framework:

- we aim at estimating the posterior distribution $p(Z, \theta | X)$,
- we have here the relation:

$$\log(p(X)) = \mathcal{L}(q(Z, \theta)) + KL(q(Z, \theta) || p(Z, \theta | X)),$$

- we also assume that q factorizes over Z and θ:

$$q(Z, \theta) = \prod_i q_i(Z_i) q_\theta(\theta).$$

The VBEM algorithm:

- VB-E step: $q_\theta(\theta)$ is fixed and $\mathcal{L}$ is maximized over the q_i $\Rightarrow$ $\log q_j^*(Z_j) = E_{i \neq j, \theta}[\log p(X, Z, \theta)] + c$
- VB-M step: all $q_i(Z_i)$ are now fixed and $\mathcal{L}$ is maximized over q_θ $\Rightarrow$ $\log q_\theta^*(\theta) = E_Z[\log p(X, Z, \theta)] + c$