Averaging Stochastic Block Models to approximate W-graphs: An illustration of variational (Bayes) inference for latent variable models

S. Robin

AgroParisTech / INRA

joint work with J.-J. Daudin, F. Picard, M. Mariadassou, C. Vacher
S. Volant, M.-L. Martin-Magniette,
C. Matias, P. Latouche

Modèles graphiques probabilistes et modèles de données structurées sur graphes, Juillet 2014, Grenoble

Looking for conditional distributions

Many hierarchical / graphical models involve unobserved ('hidden') variables Z (or parameters θ).

Many inference strategies aim at or require to determine their conditional distribution given the observations X.

Looking for conditional distributions

Many hierarchical / graphical models involve unobserved ('hidden') variables Z (or parameters θ).

Many inference strategies aim at or require to determine their conditional distribution given the observations X.

Examples.

- Frequentist maximum likelihood for latent variable models via EM:
  $\rightarrow$ requires to compute $P_\theta(Z|X)$.

- Bayesian inference:
  $\rightarrow$ aims at computing $P(\theta|X)$ or even $P(Z, \theta|X)$.

Outline

1. Incomplete data models

2. Variational approximation / Variational (Bayes) inference

3. Model averaging

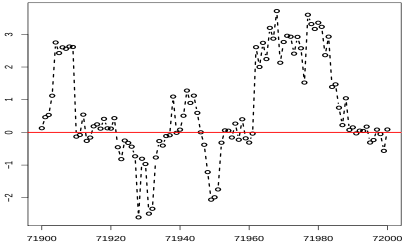
4. From SBM to W-graphs

Incomplete data models

Hidden Markov Model (HMM)

Data. $X = (X_t)$ observed along 'time'.

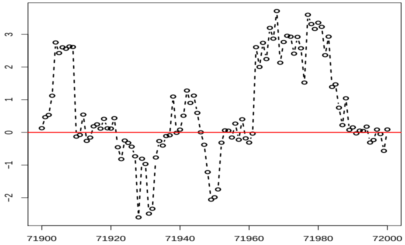
Idea. The distribution of X_t depends on some hidden Z_t.



Hidden Markov Model (HMM)

Data. $X = (X_t)$ observed along 'time'.

Idea. The distribution of X_t depends on some hidden Z_t.



Model.

- $Z = (Z_t)$ is an homogeneous Markov chain with transition π;
- Observations $X = (X_t)$ are independent given Z;
- The distribution of X_t depends on Z_t: $(X_t | Z_t = k) \sim f_k = f(\cdot; \gamma_k)$;

Frequentist maximum likelihood inference

Likelihood. The (log-)likelihood

$$\log P_\theta(X) = \log \sum_z P_\theta(X, z)$$

can not be straightforwardly maximized.

Frequentist maximum likelihood inference

Likelihood. The (log-)likelihood

$$\log P_\theta(X) = \log \sum_z P_\theta(X, z)$$

can not be straightforwardly maximized.

EM trick. But it can be decomposed as

$$\log P_\theta(X) = \mathbb{E}[\log P_\theta(X, Z)|X] + \mathcal{H}[P_\theta(Z|X)],$$

where $\mathcal{H}$ stands for the entropy:

$$\mathcal{H}[P(U)] = -\mathbb{E}[\log P(U)].$$

EM algorithm

Aims at maximizing the log-likelihood

$$\log P_\theta(X)$$

through the alternation of two steps [8].

EM algorithm

Aims at maximizing the log-likelihood

$$\log P_\theta(X)$$

through the alternation of two steps [8].

E-step: calculate $P_{\theta^h}(Z|X)$
$\longrightarrow$ not explicit
$\longrightarrow$ but calculable via the forward-backward recursion.

EM algorithm

Aims at maximizing the log-likelihood

$$\log P_\theta(X)$$

through the alternation of two steps [8].

E-step: calculate $P_{\theta^h}(Z|X)$
$\longrightarrow$ not explicit
$\longrightarrow$ but calculable via the forward-backward recursion.

M-step: maximize $\mathbb{E}_{\theta^h}[\log P_\theta(X, Z)|X]$ with respect to θ
$\longrightarrow$ often similar to standard MLE.

Conditional distribution

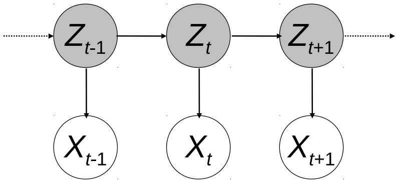
Graphical model representation.

Conditional distribution

Frequentist inference.

- The dependency graph is a tree;

Graphical model representation.

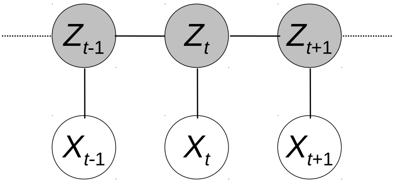


Conditional distribution

Frequentist inference.

- The dependency graph is a tree;
- The moral graph is still a tree;

Graphical model representation.

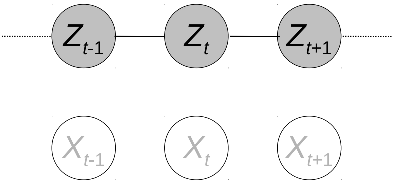


Conditional distribution

Frequentist inference.

- The dependency graph is a tree;
- The moral graph is still a tree;
- $P_\theta(Z|X)$ can be computed efficiently.

Graphical model representation.



Conditional distribution

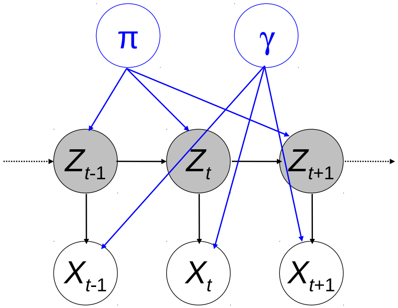
Frequentist inference.

- The dependency graph is a tree;
- The moral graph is still a tree;
- $P_\theta(Z|X)$ can be computed efficiently.

Bayesian inference.

- The dependency graph is not a tree;

Graphical model representation.



Conditional distribution

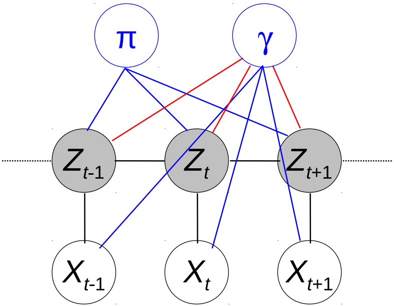
Frequentist inference.

- The dependency graph is a tree;
- The moral graph is still a tree;
- $P_\theta(Z|X)$ can be computed efficiently.

Bayesian inference.

- The dependency graph is not a tree;
- The moral graph is not a tree (Z and $\theta = (\pi, \gamma)$ are not conditionally independent);

Graphical model representation.



Conditional distribution

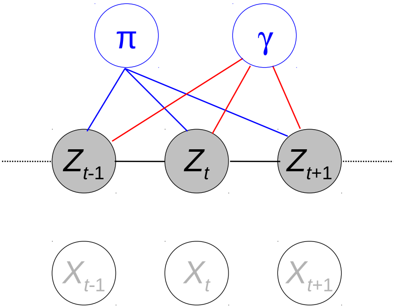
Frequentist inference.

- The dependency graph is a tree;
- The moral graph is still a tree;
- $P_\theta(Z|X)$ can be computed efficiently.

Bayesian inference.

- The dependency graph is not a tree;
- The moral graph is not a tree (Z and $\theta = (\pi, \gamma)$ are not conditionally independent);
- $P(Z, \theta|X)$ can not be computed efficiently.

Graphical model representation.



Conditional distribution

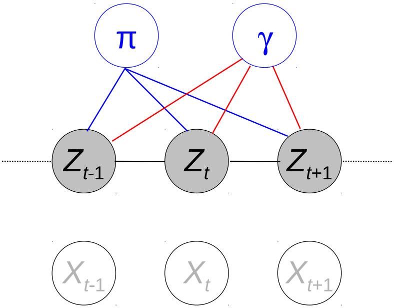
Frequentist inference.

- The dependency graph is a tree;
- The moral graph is still a tree;
- $P_\theta(Z|X)$ can be computed efficiently.

Bayesian inference.

- The dependency graph is not a tree;
- The moral graph is not a tree (Z and $\theta = (\pi, \gamma)$ are not conditionally independent);
- $P(Z, \theta|X)$ can not be computed efficiently.

Graphical model representation.



$\rightarrow$ Variational methods can be used to approximate it.

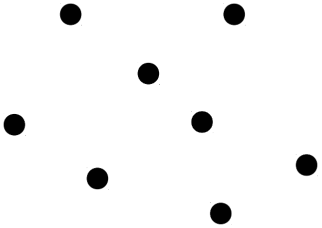
Stochastic Block Model (SBM)

A mixture model for random graphs.

Stochastic Block Model (SBM)

A mixture model for random graphs.

- Consider n nodes ($i = 1..n$);



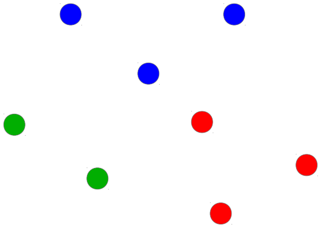
Stochastic Block Model (SBM)

A mixture model for random graphs.

- Consider n nodes ($i = 1..n$);
- Z_i = unobserved label of node i:

$$\{Z_i\} \text{ i.i.d. } \sim \mathcal{M}(1; \pi)$$

$$\pi = (\pi_1, ...\pi_K);$$



Stochastic Block Model (SBM)

A mixture model for random graphs.

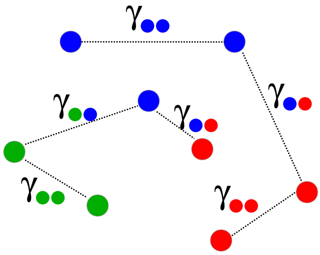
- Consider n nodes ($i = 1..n$);
- Z_i = unobserved label of node i:

$$\{Z_i\} \text{ i.i.d. } \sim \mathcal{M}(1; \pi)$$

$$\pi = (\pi_1, \dots \pi_K);$$

- Edge X_{ij} depends on the labels: $\{X_{ij}\}$ independent given $\{Z_i\}$,

$$(X_{ij}) \sim \mathcal{B}(\gamma_{Z_i, Z_j})$$



Stochastic Block Model (SBM)

A mixture model for random graphs.

- Consider n nodes ($i = 1..n$);

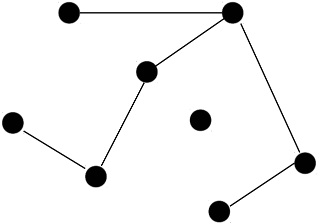
- Z_i = unobserved label of node i:

$$\{Z_i\} \text{ i.i.d. } \sim \mathcal{M}(1; \pi)$$

$$\pi = (\pi_1, \dots \pi_K);$$

- Edge X_{ij} depends on the labels: $\{X_{ij}\}$ independent given $\{Z_i\}$,

$$(X_{ij}) \sim \mathcal{B}(\gamma_{Z_i, Z_j})$$



Frequentist maximum likelihood inference

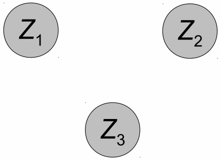
Maximum likelihood. MLE of θ obtained via the EM algorithm ...
provided that we can calculate

$$P_\theta(Z|X).$$

Frequentist maximum likelihood inference

Maximum likelihood. MLE of θ obtained via the EM algorithm ... provided that we can calculate

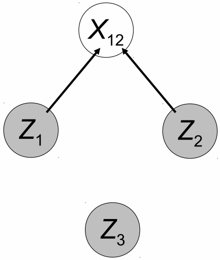
$$P_\theta(Z|X).$$



Frequentist maximum likelihood inference

Maximum likelihood. MLE of θ obtained via the EM algorithm ... provided that we can calculate

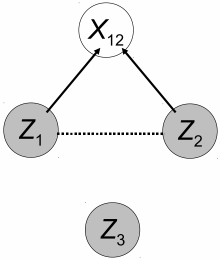
$$P_\theta(Z|X).$$



Frequentist maximum likelihood inference

Maximum likelihood. MLE of θ obtained via the EM algorithm ... provided that we can calculate

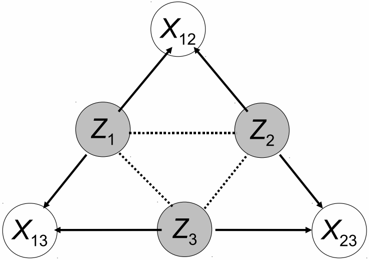
$$P_\theta(Z|X).$$



Frequentist maximum likelihood inference

Maximum likelihood. MLE of θ obtained via the EM algorithm ... provided that we can calculate

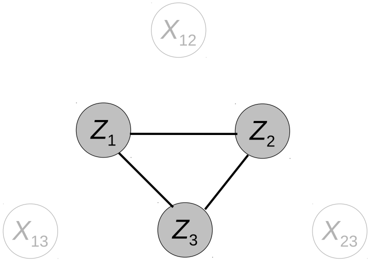
$$P_\theta(Z|X).$$



Frequentist maximum likelihood inference

Maximum likelihood. MLE of θ obtained via the EM algorithm ... provided that we can calculate

$$P_\theta(Z|X).$$



Conditional distribution. The dependency graph of Z given X is a clique.
$\rightarrow$ No factorization can be hoped (unlike for HMM).
$\rightarrow$ $P_\theta(Z|X)$ can not be computed (efficiently).
$\rightarrow$ Variational techniques provide

$$Q(Z) \simeq P_\theta(Z|X).$$

Bayesian inference

We are now interested in

$$P(Z, \theta | X)$$

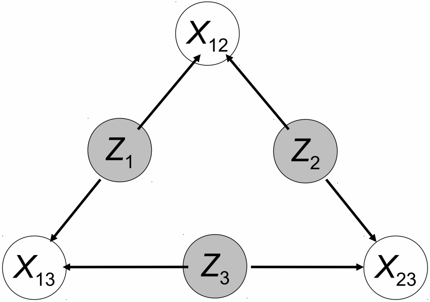
$\rightarrow$ more intricate than $P_\theta(Z|X)$.

Bayesian inference

We are now interested in

$$P(Z, \theta | X)$$

$\rightarrow$ more intricate than $P_\theta(Z|X)$.

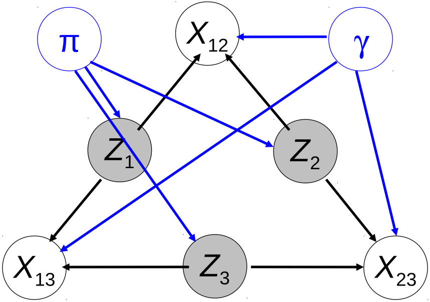


Bayesian inference

We are now interested in

$$P(Z, \theta | X)$$

$\rightarrow$ more intricate than $P_\theta(Z|X)$.

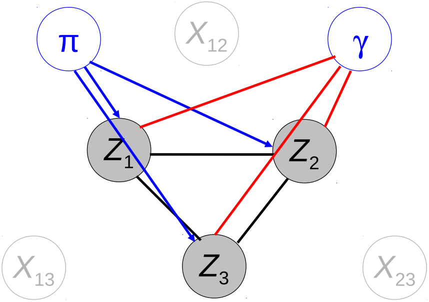


Bayesian inference

We are now interested in

$$P(Z, \theta | X)$$

$\rightarrow$ more intricate than $P_\theta(Z|X)$.



Variational techniques can provide an approximation

$$Q(Z, \theta) \simeq P(Z, \theta | X).$$

Variational approximation / Variational (Bayes) inference

Küllback-Leibler divergence

Definition.

$$KL[Q(\cdot), P(\cdot)] = \mathbb{E}_Q\left(\log \frac{Q}{P}\right) = \int Q \log Q - \int Q \log P$$

Küllback-Leibler divergence

Definition.

$$KL[Q(\cdot), P(\cdot)] = \mathbb{E}_Q \left(\log \frac{Q}{P} \right) = \int Q \log Q - \int Q \log P$$

Some properties.

- Not a distance, only a 'divergence'.
- Always positive.
- Null iff $P = Q$.
- Contrast consistent with maximum-likelihood inference.

Variational inference

Lower bound of the log-likelihood. For any distribution $Q(Z)$ [13, 25],

$$\log P(X) \geq \log P(X) - KL[Q(Z), P(Z|X)]$$

Variational inference

Lower bound of the log-likelihood. For any distribution $Q(Z)$ [13, 25],

$$\begin{aligned} \log P(X) &\geq \log P(X) - KL[Q(Z), P(Z|X)] \\ &= \int Q(z) \log P(X, z) \, \mathrm{d}z - \int Q(z) \log Q(z) \, \mathrm{d}z \end{aligned}$$

Variational inference

Lower bound of the log-likelihood. For any distribution $Q(Z)$ [13, 25],

$$\begin{aligned}\log P(X) &\geq \log P(X) - KL[Q(Z), P(Z|X)] \\ &= \int Q(z) \log P(X, z) \; \mathrm{d}z - \int Q(z) \log Q(z) \; \mathrm{d}z \\ &= \mathbb{E}_Q[\log P(X, Z)] + \mathcal{H}[Q(Z)]\end{aligned}$$

Variational inference

Lower bound of the log-likelihood. For any distribution $Q(Z)$ [13, 25],

$$\begin{aligned} \log P(X) &\geq \log P(X) - KL[Q(Z), P(Z|X)] \\ &= \int Q(z) \log P(X, z) \, \mathrm{d}z - \int Q(z) \log Q(z) \, \mathrm{d}z \\ &= \mathbb{E}_Q[\log P(X, Z)] + \mathcal{H}[Q(Z)] \end{aligned}$$

Link with EM. This is similar to

$$\log P(X) = \mathbb{E}[\log P(X, Z)|X] + \mathcal{H}[P(Z|X)]$$

replacing $P(Z|X)$ with $Q(Z)$.

Variational EM algorithm

Variational EM.

- M-step: compute

$$\widehat{\theta} = \arg\max_{\theta} \mathbb{E}_{Q^*}[\log P_\theta(X, Z)].$$

Variational EM algorithm

Variational EM.

- M-step: compute

$$\widehat{\theta} = \arg\max_{\theta} \mathbb{E}_{Q^*}[\log P_\theta(X, Z)].$$

- E-step: replace the calculation of $P(Z|X)$ with the search of

$$Q^*(Z) = \arg\min_{Q \in \mathcal{Q}} KL[Q(Z), P(Z|X)].$$

which amounts at a functional optimization problem:

$$\min_{Q \in \mathcal{Q}} \int Q(z) \log \frac{Q(z)}{P(z|X)} \, \mathrm{d}z.$$

A functional optimization problem

We want to minimize, with respect to the function Q, the functional

$$\mathcal{F}(Q) = \int \mathcal{L}[Q(z), z] \, \mathrm{d}z.$$

A functional optimization problem

We want to minimize, with respect to the function Q, the functional

$$\mathcal{F}(Q) = \int \mathcal{L}[Q(z), z] \,\mathrm{d}z.$$

Example. For variational EM, take either $z = z$ or $z = (z, \theta)$ and set

$$\mathcal{L}[Q(z), z] = Q(z) \log \left[Q(z)/P(X, z) \right].$$

A functional optimization problem

We want to minimize, with respect to the function Q, the functional

$$\mathcal{F}(Q) = \int \mathcal{L}[Q(z), z] \, \mathrm{d}z.$$

Example. For variational EM, take either $z = z$ or $z = (z, \theta)$ and set

$$\mathcal{L}[Q(z), z] = Q(z) \log \left[Q(z) / P(X, z) \right].$$

Theorem

The function

$$Q^* = \arg\min_Q \mathcal{F}(Q)$$

satisfies

$$\forall z, \qquad \left. \frac{\partial \mathcal{L}[Q(z), z]}{\partial Q(z)} \right|_{Q^*(z)} = 0.$$

Stochastic Block Model (SBM)

Likelihoods. Z_i = label, X_{ij} = edge, $\theta = (\pi, \gamma)$, $Z_{ik} = \mathbb{I}\{i \in k\}$,

$$P_\theta(X, Z) = \prod_i \prod_k \pi_k^{Z_{ik}} \times \prod_{i \neq j} \prod_{k,\ell} \underbrace{\left[\gamma_{k\ell}^{X_{ij}} (1 - \gamma_{k\ell})^{1-X_{ij}}\right]}_{f_{k\ell}(X_{ij})}^{Z_{ik} Z_{j\ell}}.$$

Stochastic Block Model (SBM)

Likelihoods. Z_i = label, X_{ij} = edge, $\theta = (\pi, \gamma)$, $Z_{ik} = \mathbb{I}\{i \in k\}$,

$$P_\theta(X, Z) = \prod_i \prod_k \pi_k^{Z_{ik}} \times \prod_{i \neq j} \prod_{k,\ell} \underbrace{\left[\gamma_{k\ell}^{X_{ij}} (1 - \gamma_{k\ell})^{1-X_{ij}}\right]}_{f_{k\ell}(X_{ij})}^{Z_{ik} Z_{j\ell}}.$$

Distribution class. $\mathcal{Q}$ = set of factorisable distributions:

$$\mathcal{Q} = \{Q : Q = \prod_i Q_i\}, \qquad Q_i(Z_i) = \prod_k \tau_{ik}^{Z_{ik}}.$$

$$\Rightarrow \quad Q_{ij} = Q_i Q_j.$$

Stochastic Block Model (SBM)

Likelihoods. $Z_i =$ label, $X_{ij} =$ edge, $\theta = (\pi, \gamma)$, $Z_{ik} = \mathbb{I}\{i \in k\}$,

$$P_\theta(X, Z) = \prod_i \prod_k \pi_k^{Z_{ik}} \times \prod_{i \neq j} \prod_{k,\ell} \underbrace{\left[\gamma_{k\ell}^{X_{ij}} (1 - \gamma_{k\ell})^{1-X_{ij}}\right]}_{f_{k\ell}(X_{ij})}^{Z_{ik} Z_{j\ell}}.$$

Distribution class. $\mathcal{Q} =$ set of factorisable distributions:

$$\mathcal{Q} = \{Q : Q = \prod_i Q_i\}, \qquad Q_i(Z_i) = \prod_k \tau_{ik}^{Z_{ik}}.$$

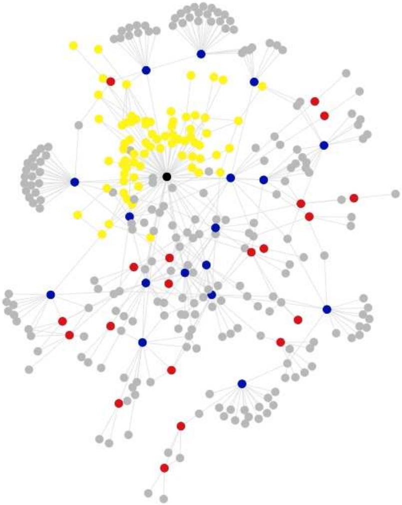
$$\Rightarrow \quad Q_{ij} = Q_i Q_j.$$

Applying the theorem leads to a fix-point relation:

$$\tau_{ik} \propto \pi_k \prod_{j \neq i} \prod_\ell f_{k\ell}(X_{ij})^{\tau_{j\ell}}$$

also known as mean-field approximation in physics. [7]

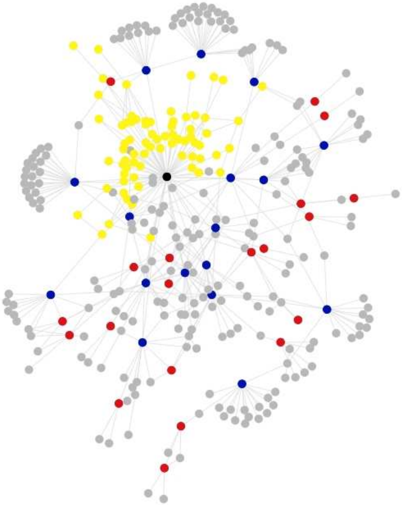
Application to *E. coli* operon network



Parameter estimates. $K = 5$

$\widehat{\gamma}_{k\ell}$ (%)	1	2	3	4	5
1	.	.	.	.	.
2	6.40	1.50	1.34	.	.
3	1.21	.	.	.	.
4	.	.	.	.	.
5	8.64	17.65	.	72.87	11.01
$\widehat{\pi}$ (%)	65.49	5.18	7.92	21.10	0.30

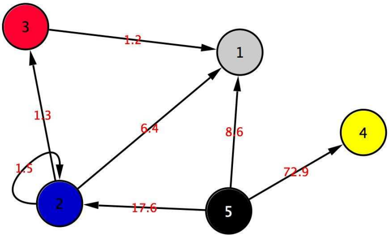
Application to *E. coli* operon network



Parameter estimates. $K = 5$

$\widehat{\gamma}_{k\ell}$ (%)	1	2	3	4	5
1	.	.	.	.	.
2	6.40	1.50	1.34	.	.
3	1.21	.	.	.	.
4	.	.	.	.	.
5	8.64	17.65	.	72.87	11.01
$\widehat{\pi}$ (%)	65.49	5.18	7.92	21.10	0.30

Meta-graph representation. [21]



Properties of the variational approximation

Not much is known about the quality of variational inference.

- VEM algorithm converges to a different optimum than ML, except for 'degenerated' models [12].
- Mean field approximation is asymptotically exact for models with infinite range dependency [20].

Properties of the variational approximation

Not much is known about the quality of variational inference.

- VEM algorithm converges to a different optimum than ML, except for 'degenerated' models [12].
- Mean field approximation is asymptotically exact for models with infinite range dependency [20].

Specific case of graphs.

- Specific asymptotic framework: n^2 data, $'p' = n$ 'variables' per individual.
- $P(Z|X)$ concentrates towards a Dirac mass when $n \to \infty$ [5], [17].
- The Dirac mass belongs to $\mathcal{Q}$...

Model selection: Choice of K

- Using Laplace approximation ($p = \#$parms) [22]:

$$\text{BIC}(K) = \log P_{\widehat{\theta}_K}(X) - \frac{p}{2}\log n \approx \log P(X, K).$$

- For classification purposes (penalize for the posterior entropy: [3])

$$\text{ICL}(K) = \text{BIC}(K) - \mathcal{H}[P_{\widehat{\theta}_K}(Z|X)] \approx \int \mathbb{E}\left[\log P(X, Z, \theta, K)|X\right] \, \text{d}\theta$$

Model selection: Choice of K

- Using Laplace approximation ($p = \#\text{parms}$) [22]:

$$\text{BIC}(K) = \log P_{\widehat{\theta}_K}(X) - \frac{p}{2} \log n \approx \log P(X, K).$$

- For classification purposes (penalize for the posterior entropy: [3])

$$\text{ICL}(K) = \text{BIC}(K) - \mathcal{H}[P_{\widehat{\theta}_K}(Z|X)] \approx \int \mathbb{E}\left[\log P(X, Z, \theta, K)|X\right] \, \mathrm{d}\theta$$

VEM inference for SBM:

$$\widetilde{\text{ICL}}(K) = \mathbb{E}_{Q^*} \log P_{\widehat{\theta}_K}(X, Z) - \frac{K-1}{2} \log n - \frac{K^2}{2} \log n(n-1)$$

$\widetilde{\text{BIC}}(K) \simeq \widetilde{\text{ICL}}(K)$ because $\mathcal{H}[Q^*(Z)] \simeq 0$.

Variational Bayes inference

Variational Bayes EM. Directly search for

$$Q^*(Z,\theta) = \arg\min_{Q \in \mathcal{Q}} KL[Q(Z,\theta), P(Z,\theta|X)].$$

Variational Bayes inference

Variational Bayes EM. Directly search for

$$Q^*(Z,\theta) = \arg\min_{Q \in \mathcal{Q}} KL[Q(Z,\theta), P(Z,\theta|X)].$$

Exponential family / conjugate prior setting.

- If the joint distribution $P(X,Z|\theta)$ is in the exponential family

$$P(x,z|\theta) \propto \exp\{\phi(\theta)^\top u(x,z)\};$$

- If the prior distribution of θ is conjugate

$$P(\theta) \quad \propto \quad \exp\{\phi(\theta)^\top \nu\}$$

Variational Bayes inference

Variational Bayes EM. Directly search for

$$Q^*(Z,\theta) = \arg\min_{Q \in \mathcal{Q}} KL[Q(Z,\theta), P(Z,\theta|X)].$$

Exponential family / conjugate prior setting.

- If the joint distribution $P(X,Z|\theta)$ is in the exponential family

$$P(x,z|\theta) \propto \exp\{\phi(\theta)^\top u(x,z)\};$$

- If the prior distribution of θ is conjugate

$$\begin{aligned} P(\theta) &\propto \exp\{\phi(\theta)^\top \nu\} \\ \Rightarrow P(x,z,\theta) &\propto \exp\{\phi(\theta)^\top [u(x,z)+\nu]\}; \end{aligned}$$

Variational Bayes inference

Variational Bayes EM. Directly search for

$$Q^*(Z,\theta) = \arg\min_{Q \in \mathcal{Q}} KL[Q(Z,\theta), P(Z,\theta|X)].$$

Exponential family / conjugate prior setting.

- If the joint distribution $P(X,Z|\theta)$ is in the exponential family

$$P(x,z|\theta) \propto \exp\{\phi(\theta)^\top u(x,z)\};$$

- If the prior distribution of θ is conjugate

$$\begin{aligned} P(\theta) &\propto \exp\{\phi(\theta)^\top \nu\} \\ \Rightarrow P(x,z,\theta) &\propto \exp\{\phi(\theta)^\top [u(x,z)+\nu]\}; \end{aligned}$$

Class of approximate distributions.

- If we restrict the approximate distribution $Q(Z,\theta)$ to

$$Q \in \mathcal{Q} := \{Q : Q = Q_Z Q_\theta\}.$$

Variational Bayes EM algorithm

The optimal distribution $Q^*(Z, \theta) = Q_Z^*(Z) Q_\theta^*(\theta)$:

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

satisfies the two following relations [2].

Remind that

$$\log P(X, Z, \theta) = \phi(\theta)[u(X, Z) + \nu]$$

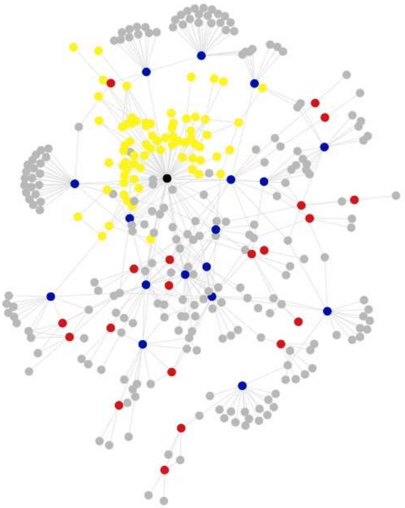
VB E-step.

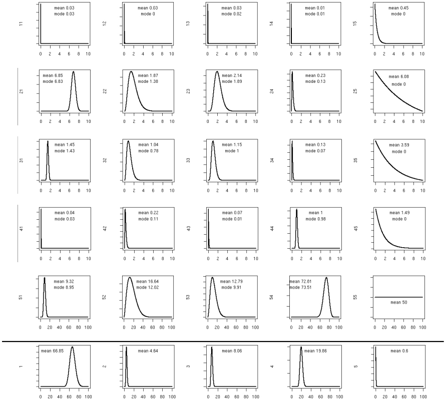
$$\begin{aligned} \log Q_Z^*(Z) &= \mathbb{E}_{Q_\theta}[\log P(X, Z, \theta)] + \text{cst} \\ &= \mathbb{E}_{Q_\theta}[\phi(\theta)]^\top u(X, Z) + \text{cst} \end{aligned}$$

VB M-step.

$$\log Q_\theta^*(\theta) \quad \propto \quad \phi(\theta)^\top \left\{ \mathbb{E}_{Q_Z}[u(X, Z)] + \nu \right\} + \text{cst}$$

Application to *E. coli* operon network





Variational Bayes approximation: Simulation Study

Few is known about the properties of variational-Bayes inference:

- Asymptotic normality of the approximate posterior [26].
- Consistency is proved for some incomplete data models [27, 19].
- In practice, VB-EM often under-estimates the posterior variances.

Variational Bayes approximation: Simulation Study

Few is known about the properties of variational-Bayes inference:

- Asymptotic normality of the approximate posterior [26].
- Consistency is proved for some incomplete data models [27, 19].
- In practice, VB-EM often under-estimates the posterior variances.

Simulation design: [10]

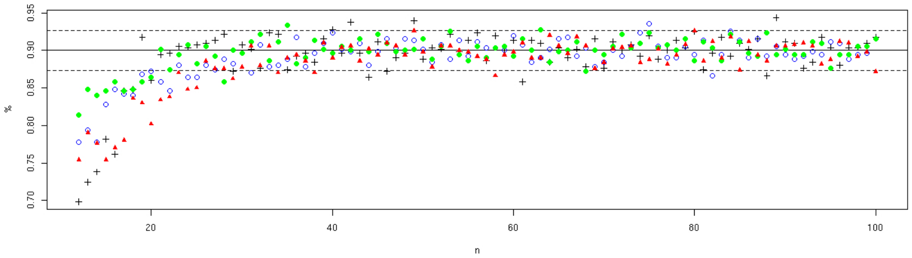
- 2-group binary SBM with parameters

$$\pi = \begin{pmatrix} 0.6 & 0.4 \end{pmatrix}, \qquad \gamma = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.3 \end{pmatrix}$$

for each scenario and each graph size.

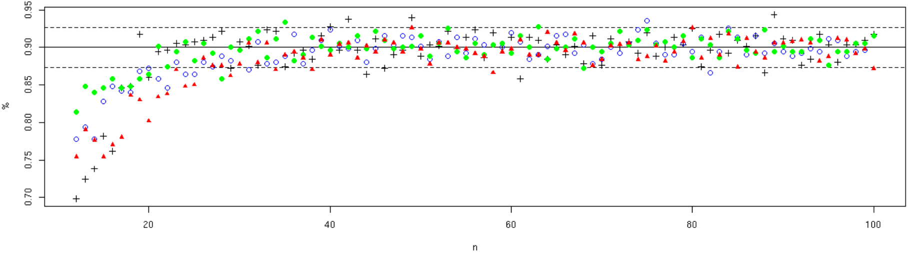
VBEM Credibility intervals

Actual level as a function of n: π_1: $+$, γ_{11}: $\triangle$, γ_{12}: $\circ$, γ_{22}: $\bullet$



VBEM Credibility intervals

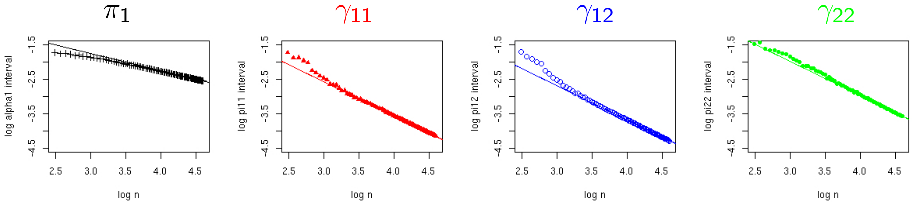
Actual level as a function of n: π_1: $+$, γ_{11}: $\triangle$, γ_{12}: $\circ$, γ_{22}: $\bullet$



For all parameters, VBEM posterior credibility intervals achieve the nominal level (90%), as soon as $n \geq 30$.
$\rightarrow$ The VBEM approximation seems to work well.

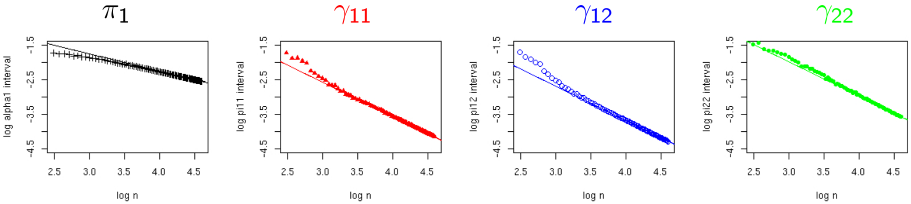
Convergence rate of the VBEM estimates

Width of the posterior credibility intervals.



Convergence rate of the VBEM estimates

Width of the posterior credibility intervals.



- The width decreases as $1/\sqrt{n}$ for π_1.
- It decreases as $1/n = 1/\sqrt{n^2}$ for γ_{11}, γ_{12} and γ_{22}.
- Consistent with the penalty of the ICL criterion proposed by [7].

Explicit inference formulas and model selection criteria for SBM: no need for Laplace approximation [14]

Model averaging

Bayesian model averaging

Problem. Consider a parameter of interest Δ that can be estimated with different models $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_K, \dots$

Bayesian model averaging

Problem. Consider a parameter of interest Δ that can be estimated with different models $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_K, \dots$

Two solutions.

1. Select a 'best model' $\mathcal{M}_{\widehat{K}}$ and take

$$\widehat{\Delta} = \widehat{\Delta}_{\widehat{K}} := \mathbb{E}(\Delta | X, \widehat{K}).$$

Bayesian model averaging

Problem. Consider a parameter of interest Δ that can be estimated with different models $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_K, \dots$

Two solutions.

1. Select a 'best model' $\mathcal{M}_{\widehat{K}}$ and take
$$\widehat{\Delta} = \widehat{\Delta}_{\widehat{K}} := \mathbb{E}(\Delta | X, \widehat{K}).$$

2. Average the estimates:
$$\widehat{\Delta} = \mathbb{E}(\Delta | X) = \sum_K P(K|X) \mathbb{E}(\Delta | X, K) = \sum_K P(K|X) \widehat{\Delta}_K.$$

Bayesian model averaging

Problem. Consider a parameter of interest Δ that can be estimated with different models $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_K, \dots$

Two solutions.

1. Select a 'best model' $\mathcal{M}_{\widehat{K}}$ and take

$$\widehat{\Delta} = \widehat{\Delta}_{\widehat{K}} := \mathbb{E}(\Delta | X, \widehat{K}).$$

2. Average the estimates:

$$\widehat{\Delta} = \mathbb{E}(\Delta | X) = \sum_K P(K|X) \mathbb{E}(\Delta | X, K) = \sum_K P(K|X) \widehat{\Delta}_K.$$

or more generally

$$P(\Delta | X) = \sum_K P(K|X) P(\Delta | X, K)$$

Variational weights

The weights $P(K|X)$ can generally not be computed

$$P(K|X) = \iint P(z, \theta, K|X) \, \mathrm{d}z \, \mathrm{d}\theta$$

Variational weights

The weights $P(K|X)$ can generally not be computed

$$P(K|X) = \iint P(z, \theta, K|X) \text{ d}z \text{ d}\theta$$

We can approximate the conditional $P(Z, \theta, K|X)$ with

$$Q^*(Z, \theta, K) = \arg\min_{Q \in \mathcal{Q}} KL[Q(Z, \theta, K), P(Z, \theta, K|X)]$$

Variational weights

The weights $P(K|X)$ can generally not be computed

$$P(K|X) = \iint P(z, \theta, K|X) \text{ d}z \text{ d}\theta$$

We can approximate the conditional $P(Z, \theta, K|X)$ with

$$Q^*(Z, \theta, K) = \arg\min_{Q \in \mathcal{Q}} KL[Q(Z, \theta, K), P(Z, \theta, K|X)]$$

Variational weights. Taking

$$\mathcal{Q} = \{Q : Q = Q_K Q_{Z|K} Q_{\theta|K}\}$$

Variational weights

The weights $P(K|X)$ can generally not be computed

$$P(K|X) = \iint P(z, \theta, K|X) \text{ d}z \text{ d}\theta$$

We can approximate the conditional $P(Z, \theta, K|X)$ with

$$Q^*(Z, \theta, K) = \arg\min_{Q \in \mathcal{Q}} KL[Q(Z, \theta, K), P(Z, \theta, K|X)]$$

Variational weights. Taking

$$\mathcal{Q} = \{Q : Q = Q_K Q_{Z|K} Q_{\theta|K}\}$$

the variational approximation gives [24]

$$\begin{aligned} Q_K^*(K) &\propto P(K) e^{\log P(X|K) - KL[Q^*(Z,\theta|K); P(Z,\theta|X,K)]} \\ &= P(K|X) e^{-KL[Q^*(Z,\theta|K); P(Z,\theta|X,K)]} \end{aligned}$$

VBMA: the recipe

For $K = 1 \dots K_{\max}$

- Use regular VB-EM to compute the approximate conditional posterior

$$Q^*_{Z,\theta|K} = Q^*_{Z|K} Q^*_{\theta|K}$$

- Compute the conditional lower bound of the log-likelihood

$$\log P(X|K) - KL[Q^*(Z,\theta|K); P(Z,\theta|X,K)] =: \log P(X|K) - KL_K.$$

VBMA: the recipe

For $K = 1 \dots K_{\max}$

- Use regular VB-EM to compute the approximate conditional posterior

$$Q^*_{Z,\theta|K} = Q^*_{Z|K} Q^*_{\theta|K}$$

- Compute the conditional lower bound of the log-likelihood

$$\log P(X|K) - KL[Q^*(Z,\theta|K); P(Z,\theta|X,K)] =: \log P(X|K) - KL_K.$$

Compute the approximate posterior of model $\mathcal{M}_K$:

$$w_K := Q_K(K) \propto P(K|X) \exp(-KL_K).$$

VBMA: the recipe

For $K = 1 \dots K_{\max}$

- Use regular VB-EM to compute the approximate conditional posterior

$$Q^*_{Z,\theta|K} = Q^*_{Z|K} Q^*_{\theta|K}$$

- Compute the conditional lower bound of the log-likelihood

$$\log P(X|K) - KL[Q^*(Z,\theta|K); P(Z,\theta|X,K)] =: \log P(X|K) - KL_K.$$

Compute the approximate posterior of model $\mathcal{M}_K$:

$$w_K := Q_K(K) \propto P(K|X) \exp(-KL_K).$$

Deduce the approximate posterior of δ:

$$P(\delta|X) \approx \sum_K w_K Q^*_{\theta|K}(\delta|K)$$

From SBM to W-graphs

Modeling network heterogeneity

Latent variable models allow to capture the underlying structure of a network (e.g. SBM).

Modeling network heterogeneity

Latent variable models allow to capture the underlying structure of a network (e.g. SBM).

General setting for binary graphs. [4]:

- A latent (unobserved) variable Z_i is associated with each node:

$$\{Z_i\} \text{ iid } \sim \pi$$

- Edges $Y_{ij} = \mathbb{I}\{i \sim j\}$ are independent conditionally to the Z_i's:

$$\{Y_{ij}\} \text{ independent } |\{Z_i\} : \Pr\{Y_{ij} = 1\} = \gamma(Z_i, Z_j)$$

See [18] for a review.

W-graph model

Latent variables:

$$(Z_i) \text{ iid } \sim \mathcal{U}_{[0,1]},$$

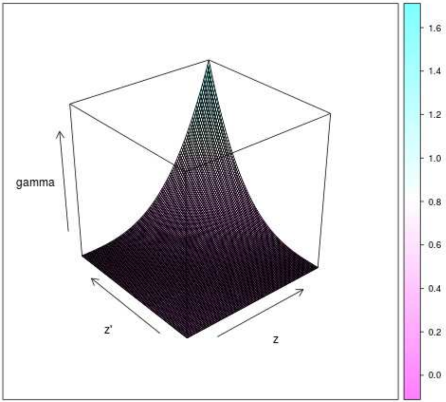
Graphon function γ:

$$\gamma(z, z') : [0,1]^2 \to [0,1]$$

Edges:

$$\Pr\{Y_{ij} = 1\} = \gamma(Z_i, Z_j)$$

Graphon function $\gamma(z, z')$



Inference of the graphon function

Probabilistic point of view.

- W-graph have been mostly studied in the probability literature as a limit for dense graphs: [16], [9]
- Intrinsic un-identifiability of the graphon function γ is often overcome by imposing that $u \mapsto \int \gamma(u, v) \, dv$ is monotonous increasing.
- Motif (sub-graph) frequencies are invariant characteristics of a W-graph.

Inference of the graphon function

Probabilistic point of view.

- W-graph have been mostly studied in the probability literature as a limit for dense graphs: [16], [9]
- Intrinsic un-identifiability of the graphon function γ is often overcome by imposing that $u \mapsto \int \gamma(u, v) \, \mathrm{d}v$ is monotonous increasing.
- Motif (sub-graph) frequencies are invariant characteristics of a W-graph.

Statistical point of view.

- Not much attention has been paid to its inference until very recently: [6], [1], [28], ...
- The two latter also uses SBM as a proxy for W-graph.

SBM as a W-graph model

Latent variables:

$$(Z_i) \text{ iid } \sim \mathcal{M}(1, \pi)$$

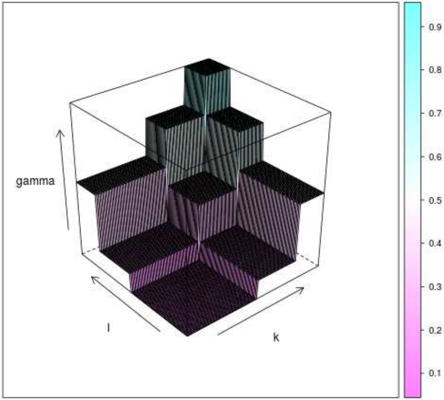
Blockwise constant graphon:

$$\gamma(z, z') = \gamma_{k\ell}$$

Edges:

$$\Pr\{Y_{ij} = 1\} = \gamma(Z_i, Z_j)$$

Graphon function $\gamma_K^{SBM}(z, z')$



$\rightarrow$ block widths $= \pi_k$, block heights $\gamma_{k\ell}$

Variational Bayes estimation of $\gamma(z, z')$

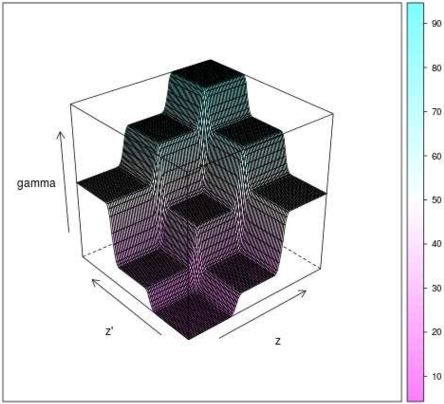
VBEM inference provides the approximate posteriors:

$$(\pi|X) \approx \text{Dir}(\pi^*)$$
$$(\gamma_{k\ell}|X) \approx \text{Beta}(\gamma_{k\ell}^{0*}, \gamma_{k\ell}^{1*})$$

Estimate of $\gamma(u, v)$. BEM: due to the uncertainty of the block widths, the posterior expectation of γ_K^{SBM} is smooth

(Explicit integration using [11])

Posterior mean $\mathbb{E}_{Q_K^*}(\gamma_K^{SBM}(z, z')|X)$



Averaging SBMs

Model averaging: There is no 'true K' in the W-graph model.

Apply VBMA recipe. For $K = 1..K_{\max}$, fit an SBM model via VBEM and compute

$$\widehat{\gamma}_K^{SBM}(z, z') = \mathbb{E}_{Q_K^*}[\gamma_{C(z),C(z')}].$$

Averaging SBMs

Model averaging: There is no 'true K' in the W-graph model.

Apply VBMA recipe. For $K = 1..K_{\max}$, fit an SBM model via VBEM and compute

$$\widehat{\gamma}_K^{SBM}(z, z') = \mathbb{E}_{Q_K^*}[\gamma_{C(z),C(z')}].$$

Perform model averaging as

$$\widehat{\gamma}(z, z') = \sum_K w_K \widehat{\gamma}_K^{SBM}(z, z')$$

where w_K is the variational weights arising from variational Bayes inference.

Some simulations

Design. Symetric graphon:

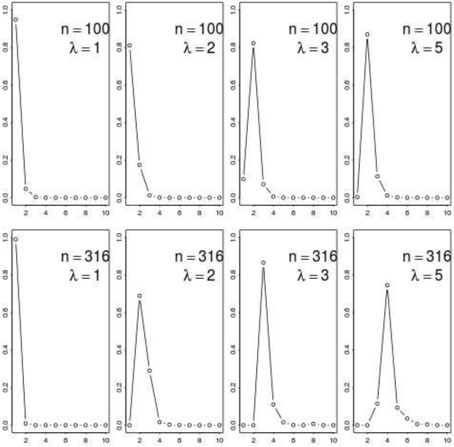
$$\gamma(u, v) = \rho\lambda^2(uv)^{\lambda-1}$$

- $\lambda\uparrow$: imbalanced graph
- $\rho\uparrow$: dense graph

Results.

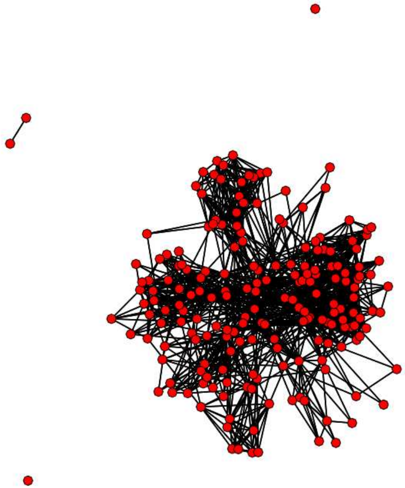
- More complex models as n and $\lambda\uparrow$
- Posterior fairly concentrated

Variational posterior for K: $Q^*(K)$.



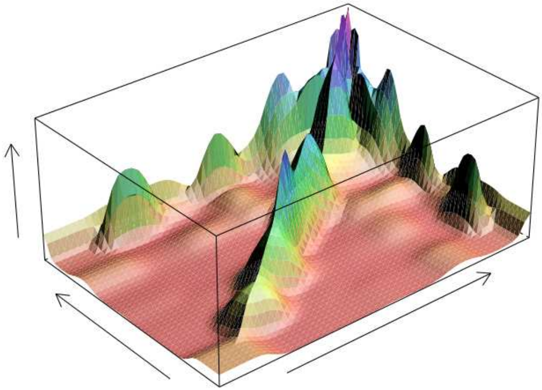
French political blogosphere

Website network. French political blogs: 196 nodes, 1432 edges.



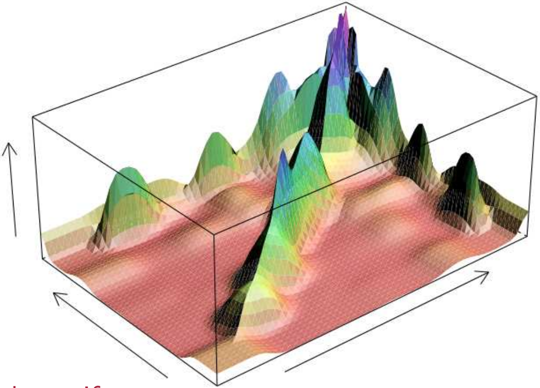
French political blogosphere.

Infered graphon.



French political blogosphere.

Infered graphon.



Extension to network motifs.

- Network motifs have sociological interpretation (e.g. triangles)
- Motif probability can be estimated as $\widehat{\mu}(m) = \mathbb{E}_{Q^*}(\mu(m)|X)$
  $\rightarrow$ Goodness of fit criterion?

Some conclusions

Variational approximation.

- Reasonably simple tool to get approximate conditional distributions;
- More scalable than standard MCMC (?);
- Few general theoretical guaranties;
- Happens to be efficient in some specific frameworks (e.g. graphs[1]);
- Can be combined with Monte-Carlo approaches, e.g. for importance sampling [15].

[1] not true for sparse graphs

E. M Airoldi, T. B Costa, and S. H Chan.
Stochastic blockmodel approximation of a graphon: Theory and consistent estimation.
Technical report, arXiv:1311.1731, 2013.

J. Beal, M. and Z. Ghahramani.
The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures.
Bayes. Statist., 7:543–52, 2003.

C. Biernacki, G. Celeux, and G. Govaert.
Assessing a mixture model for clustering with the integrated completed likelihood.
IEEE Trans. Pattern Anal. Machine Intel., 22(7):719–725, 2000.

B. Bollobás, S. Janson, and O. Riordan.
The phase transition in inhomogeneous random graphs.
Rand. Struct. Algo., 31(1):3–122, 2007.

A. Celisse, J.-J. Daudin, and L. Pierre.
Consistency of maximum-likelihood and variational estimators in the Stochastic Block Model.
Technical Report 1105.3288, ArXiv e-prints, 2011.

S. Chatterjee.
Matrix estimation by Universal Singular Value Thresholding.
Technical report, arXiv:1212.1247, 2012.

J.-J. Daudin, F. Picard, and S. Robin.
A mixture model for random graphs.
Stat. Comput., 18(2):173–83, Jun 2008.

A. P. Dempster, N. M. Laird, and D. B. Rubin.
Maximum likelihood from incomplete data via the EM algorithm.
J. R. Statist. Soc. B, 39:1–38, 1977.

P. Diaconis and S. Janson.
Graph limits and exchangeable random graphs.

Rend. Mat. Appl., 7(28):33–61, 2008.

Steven Gazal, Jean-Jacques Daudin, and Stéphane Robin.
Accuracy of variational estimates for random graph mixture models.
Journal of Statistical Computation and Simulation, 82(6):849–862, 2012.

A.A. Gouda and T. Szántai.
On numerical calculation of probabilities according to Dirichlet distribution.
Ann. Oper. Res., 177:185–200, 2010.
DOI: 10.1007/s10479-009-0601-9.

A. Gunawardana and W. Byrne.
Convergence theorems for generalized alternating minimization procedures.
J. Mach. Learn. Res., 6:2049–73, 2005.

T. Jaakkola.
Advanced mean field methods: theory and practice, chapter Tutorial on variational approximation methods.
MIT Press, 2000.

P. Latouche, E. Birmelé, and C. Ambroise.
Variational bayesian inference and complexity control for stochastic block models.
Statis. Model., 12(1):93–115, 2012.

Sébastien Li-Thiao-Té, Daudin Jean-Jacques, and Robin Stéphane.
Bayesian model averaging for estimating the number of classes: applications to the total number of species in metagenomics.
Journal of Applied Statistics, 39(7):1489–1504, 2012.

L. Lovász and B. Szegedy.
Limits of dense graph sequences.
Journal of Combinatorial Theory, Series B, 96(6):933 – 957, 2006.

M. Mariadassou and C. Matias.
Convergence of the groups posterior distribution in latent or stochastic block models.

Technical report, arXiv:1206.7101, 2012.

C. Matias and S. Robin.
Modeling heterogeneity in random graphs: a selective review.
Technical report, arXiv:1402.4296, 2014.

C. A. McGrory and D. M. Titterington.
Variational Bayesian analysis for hidden Markov models.
Austr. & New Zeal. J. Statist., 51(2):227–44, 2009.

M. Opper and O. Winther.
Advanced mean field methods: Theory and practice, chapter From Naive Mean Field Theory to the TAP Equations.
The MIT Press, 2001.

F. Picard, V. Miele, J.-J. Daudin, L. Cottret, and S. Robin.
Deciphering the connectivity structure of biological networks using mixnet.
BMC Bioinformatics, Suppl 6:S17, 2009.
doi:10.1186/1471-2105-10-S6-S17.

G. Schwarz.
Estimating the dimension of a model.
Ann. Statist., 6:461–4, 1978.

W. Sun and T. T. Cai.
Large-scale multiple testing under dependence.
J. R. Statist. Soc. B, 71(2):393–424, 2009.

Stevenn Volant, Marie-Laure Martin Magniette, and Stéphane Robin.
Variational bayes approach for model aggregation in unsupervised classification with markovian dependency.
Comput. Statis. & Data Analysis, 56(8):2375 – 2387, 2012.

M. J. Wainwright and M. I. Jordan.
Graphical models, exponential families, and variational inference.
Found. Trends Mach. Learn., 1(1–2):1–305, 2008.
http:/dx.doi.org/10.1561/2200000001.

B. Wang and M. Titterington, D.
Convergence and asymptotic normality of variational bayesian approximations for exponential family models with missing values.
In *Proc. 20th Conf. Uncertainty in Artificial Intelligence*, volume 70 of *ACM Internat. Conf. Proc.*, pages 577–84, 2004.

B. Wang and M. Titterington, D.
Convergence properties of a general algorithm for calculating variational bayesian estimates for a normal mixture model.
Bayes. Anal., 1(3):625–50, 2006.

P. J. Wolfe and S. C. Olhede.
Nonparametric graphon estimation.
Technical report, arXiv:1309.5936, 2013.

Sketch of proof

Sketch of proof

1. Q^* is optimal if, for any function H ('perturbation', 'direction'),

$$\frac{\partial}{\partial t}\mathcal{F}(Q^* + tH)\Big|_{t=0} = 0.$$

Sketch of proof

1. Q^* is optimal if, for any function H ('perturbation', 'direction'),

$$\frac{\partial}{\partial t}\mathcal{F}(Q^* + tH)\bigg|_{t=0} = 0.$$

2. At $t = 0$, under regularity conditions (and since $(f \circ g)' = g' \times f' \circ g$)

$$\frac{\partial}{\partial t}\mathcal{F}(Q+tH) = \int \frac{\partial}{\partial t}\mathcal{L}[Q(z)+tH(z), z] \, \mathrm{d}z = \int H(z)\frac{\partial \mathcal{L}[Q(z), z]}{\partial Q(z)} \, \mathrm{d}z.$$

Sketch of proof

1. Q^* is optimal if, for any function H ('perturbation', 'direction'),
$$\left.\frac{\partial}{\partial t}\mathcal{F}(Q^* + tH)\right|_{t=0} = 0.$$

2. At $t = 0$, under regularity conditions (and since $(f \circ g)' = g' \times f' \circ g$)
$$\frac{\partial}{\partial t}\mathcal{F}(Q+tH) = \int \frac{\partial}{\partial t}\mathcal{L}[Q(z)+tH(z), z] \; \mathrm{d}z = \int H(z)\frac{\partial\mathcal{L}[Q(z), z]}{\partial Q(z)} \; \mathrm{d}z.$$

3. The 'fundamental lemma of variation calculus' says that
$$\forall h, \int f(z)h(z) \; \mathrm{d}z = 0 \qquad \Rightarrow \qquad \forall z, \; f(z) = 0.$$

Variational optimization

We want to minimize $KL[Q(Z,\theta), P(Z,\theta|X)]$ within $\mathcal{Q}$, that is

$$\mathcal{F}(Q_Z, Q_\theta) = \int \int Q_Z(z) Q_\theta(\theta) \log \frac{Q_Z(z) Q_\theta(\theta)}{P(X, z, \theta)} \text{ d}z \text{ d}\theta.$$

Variational optimization

We want to minimize $KL[Q(Z,\theta), P(Z,\theta|X)]$ within $\mathcal{Q}$, that is

$$\mathcal{F}(Q_Z, Q_\theta) = \int\int Q_Z(z)Q_\theta(\theta) \log \frac{Q_Z(z)Q_\theta(\theta)}{P(X,z,\theta)} \, \mathrm{d}z \, \mathrm{d}\theta.$$

To get the optimal Q_Z^*, we set

$$\mathcal{L}_Z[Q_Z(z), z] = Q_Z(z) \int Q_\theta(\theta) \log \frac{Q_Z(z)Q_\theta(\theta)}{P(X,z,\theta)} \, \mathrm{d}\theta.$$

Variational optimization

We want to minimize $KL[Q(Z,\theta), P(Z,\theta|X)]$ within $\mathcal{Q}$, that is

$$\mathcal{F}(Q_Z, Q_\theta) = \int \int Q_Z(z) Q_\theta(\theta) \log \frac{Q_Z(z) Q_\theta(\theta)}{P(X, z, \theta)} \, \mathrm{d}z \, \mathrm{d}\theta.$$

To get the optimal Q_Z^*, we set

$$\mathcal{L}_Z[Q_Z(z), z] = Q_Z(z) \int Q_\theta(\theta) \log \frac{Q_Z(z) Q_\theta(\theta)}{P(X, z, \theta)} \, \mathrm{d}\theta.$$

According to Theorem 1, Q_Z^* satisfies

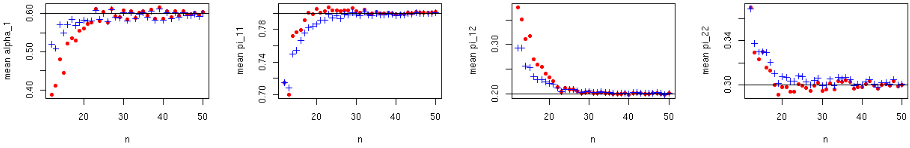
$$\frac{\partial \mathcal{L}_Z[Q_Z(z), z]}{\partial Q_Z(z)} = \int Q_\theta(\theta) \log \frac{Q_\theta(\theta)}{P(X, z, \theta)} \, \mathrm{d}\theta + \log Q_Z(z) - 1 = 0$$

Influence of the graph size

Comparison of VEM: • and VBEM: +

Left to right: π_1, γ_{11}, γ_{12}, γ_{22}.

Means.

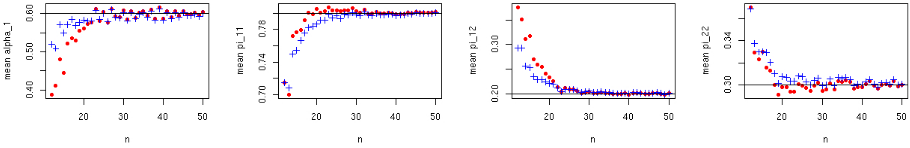


Influence of the graph size

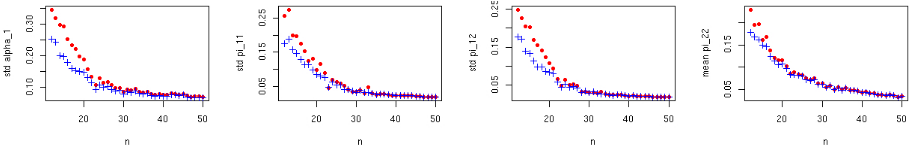
Comparison of VEM: • and VBEM: $+$

Left to right: π_1, γ_{11}, γ_{12}, γ_{22}.

Means.



Standard deviations.

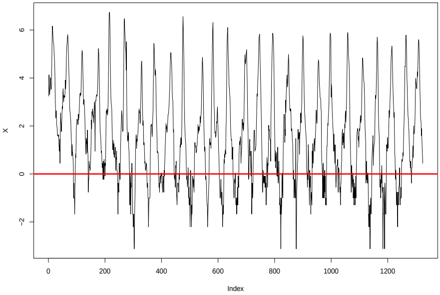


- VBEM estimates converge more rapidly than VEM estimates.
- Their precision is also better.

Classification with HMM

Context. A 2-state (normal/alert) HMM where the 'normal' distribution if known.

Influenza incidence rate: Weekly data.



(Réseau sentinelle).

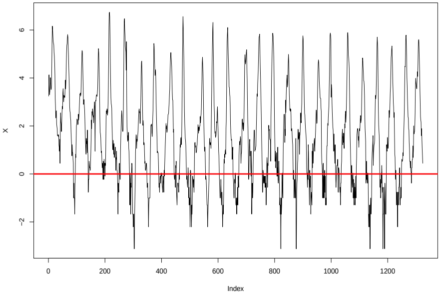
Classification with HMM

Context. A 2-state (normal/alert) HMM where the 'normal' distribution if known.

Model.

- $X_t | Z_t = 0 \sim \phi$
- $X_t | Z_t = 1 \sim g = \sum_k \pi_k f_k$

Influenza incidence rate: Weekly data.



(Réseau sentinelle).

Classification with HMM

Context. A 2-state (normal/alert) HMM where the 'normal' distribution if known.

Model.

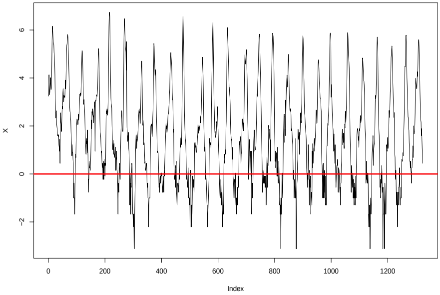
- $X_t | Z_t = 0 \sim \phi$
- $X_t | Z_t = 1 \sim g = \sum_k \pi_k f_k$

Classification

$$\tau_t = \Pr\{Z_t = 1 | X\}$$

$\widehat{\tau}_K$ for a K-component mixture.

Influenza incidence rate: Weekly data.



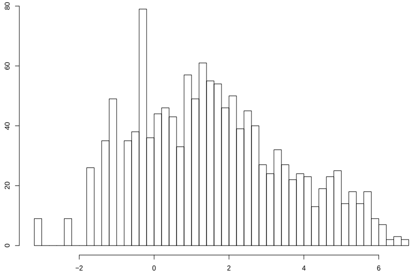
(Réseau sentinelle).

Classification with HMM

Mixture emission. For each K, estimates of g and τ are obtained.

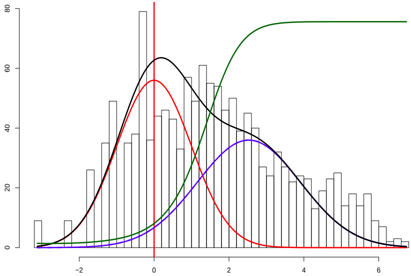
Classification with HMM

Mixture emission. For each K, estimates of g and τ are obtained.



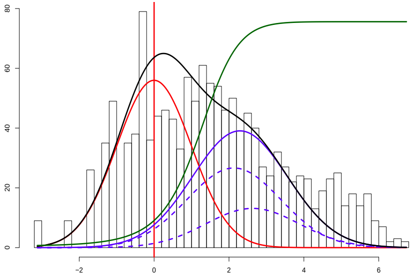
Classification with HMM

Mixture emission. For each K, estimates of g and τ are obtained.



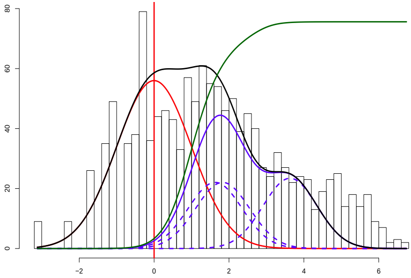
Classification with HMM

Mixture emission. For each K, estimates of g and τ are obtained.



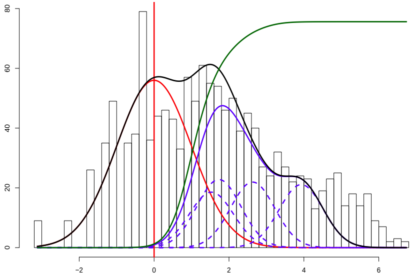
Classification with HMM

Mixture emission. For each K, estimates of g and τ are obtained.



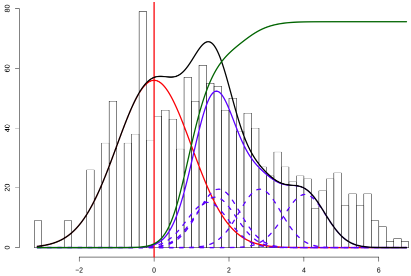
Classification with HMM

Mixture emission. For each K, estimates of g and τ are obtained.



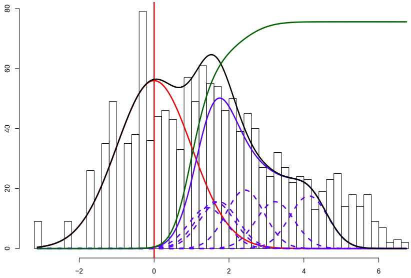
Classification with HMM

Mixture emission. For each K, estimates of g and τ are obtained.



Classification with HMM

Mixture emission. For each K, estimates of g and τ are obtained.

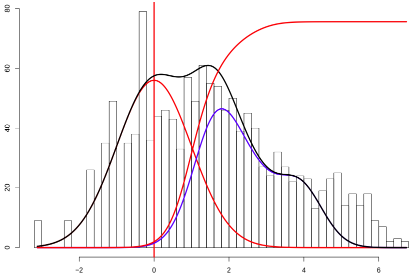


Classification with HMM

Mixture emission. For each K, estimates of g and τ are obtained.

Model averaging.

$$\widehat{\tau} = \sum_K w_K \widehat{\tau}_K$$

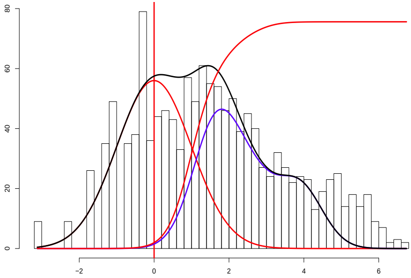


Classification with HMM

Mixture emission. For each K, estimates of g and τ are obtained.

Model averaging.

$$\widehat{\tau} = \sum_K w_K \widehat{\tau}_K$$



The estimation of τ_t provides a control of the FDR under dependency [23]:

$$\widehat{FDR}(s) = \frac{\sum_{t:\widehat{\tau}_t \leq s} \widehat{\tau}_t}{\#\{t : \widehat{\tau}_t \leq s\}}.$$