Accuracy of Variational Estimates for Random Graph Mixture Models

S. Gazal, J.-J. Daudin, S. Robin

AgroParisTech / INRA

MSTGA, Toulouse, Dec. 2009

Mixture model (MixNet)

We consider an undirected graph with n nodes $(i, j = 1 \dots n)$ and binary edges:

$$X_{ij} = \mathbb{I}\{i \sim j\}.$$

The nodes are spread into Q groups $(q, \ell = 1 \dots Q)$ with proportions

$$(\alpha_1, \dots, \alpha_Q) =: \alpha.$$

The groups to which each node belongs are independent:

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

The edges are independent conditionally to the node's groups (*Daudin et al. (2008)*):

$$\{X_{ij}\} \text{ independent}, \qquad X_{ij}|Z_{iq}Z_{j\ell} \sim \mathcal{B}(\pi_{q\ell}).$$

Likelihoods

The log-likelihood of the complete data is

$$\log P(X, Z) = \frac{1}{2} \sum_{i,j \neq i} \sum_{q,l} Z_{iq} Z_{j\ell} \log b_{ijq\ell} + \sum_i \sum_q Z_{iq} \log \alpha_q$$

where $b_{ijq\ell} = \pi_{q\ell}^{X_{ij}} (1 - \pi_{q\ell})^{1 - X_{ij}}$.

Its conditional expectation denoted by $\mathcal{Q}$ in the EM literature is

$$\begin{aligned} \mathcal{Q}(X) &= \mathbb{E}[\log P(Z, X) | X] \\ &= \frac{1}{2} \sum_{i,j \neq i} \sum_{q,l} \Delta_{ijq\ell} \log b_{ijq\ell} + \sum_i \sum_q \tau_{iq} \log \alpha_q \end{aligned}$$

where $\tau_{iq} = \mathbb{E}(Z_{iq} | X)$ and $\Delta_{ijq\ell} = \mathbb{E}(Z_{iq} Z_{j\ell} | X)$.

Regular EM

M-step. Parameter estimates are straightforward and similar for all inference methods

$$\pi_{q\ell}^{\text{EM}} = \frac{\sum_{i \neq j} X_{ij} \Delta_{ijq\ell}^{\text{EM}}}{\sum_{i \neq j} \Delta_{ijq\ell}^{\text{EM}}} \qquad \text{and} \qquad \alpha_q^{\text{EM}} = \frac{1}{n} \sum_i \tau_{iq}^{\text{EM}}.$$

E-step. It aims at computing the conditional distribution of the unobserved data Z:

$$\tau_{iq}^{\text{EM}} = \sum_z z_{iq} P(Z = z|X) \qquad \text{and} \qquad \Delta_{ijq\ell}^{\text{EM}} = \sum_z z_{iq} z_{j\ell} P(Z = z|X) \ .$$

Except for small datasets, $\sum_z$ can not be computed.

The methods presented hereafter provide approximations of τ_{iq} and $\Delta_{ijq\ell}$.

Variational EM (VEM)

The variational strategy (*Jaakola (2000)*) aims at maximizing a lower bound of $\log P(X)$

$$\begin{aligned} \mathcal{J}(X) &= \log P(X) - KL(\mathcal{R}_X(Z), P(Z|X)) \\ &= \left[\frac{1}{2} \sum_{i \neq j} \sum_{q,l} \Delta^{\text{VEM}}_{ijq\ell} \log b_{ijq\ell} + \sum_i \sum_q \tau^{\text{VEM}}_{iq} \log \alpha_q \right] \end{aligned}$$

$\mathcal{R}_X$ is chosen in a set of manageable distributions:

$$\mathcal{R}_X(Z) = \prod_i \prod_q (\tau_{iq})^{Z_{iq}}$$

which implies $\Delta^{\text{VEM}}_{ijq\ell} = \tau^{\text{VEM}}_{iq} \tau^{\text{VEM}}_{j\ell}$.

τ^{VEM}_{iq} satisfy a fix-point equation:

$$\tau^{\text{VEM}}_{iq} \propto \alpha^{\text{VEM}}_q \prod_{j \neq i} \prod_l (b_{ijq\ell})^{\tau^{\text{VEM}}_{j\ell}}.$$

Belief Propagation (BP)

Based on a message passing principle (*MacKay (2003)*), we get a similar τ_{iq}, but a new version of $\Delta_{ijq\ell}$:

$$\Delta^{\text{BP}}_{ijq\ell} \quad \propto \quad \alpha_q \alpha_\ell b_{ijq\ell} \prod_{k \neq i,j} \prod_r \left(b_{ikqr}\right)^{\tau^{\text{BP}}_{kl}} \left(b_{kjqr}\right)^{\tau^{\text{BP}}_{kr}} .$$

It can be seen as a slight modification of the VEM approximation:

$$\Delta^{\text{BP}}_{ijq\ell} \propto \tau^{\text{BP}}_{iq} \tau^{\text{BP}}_{j\ell} \frac{b_{ijq\ell}}{\prod_r \left[\left(b_{ijqr}\right)^{\tau^{\text{BP}}_{jr}} \left(b_{ijrl}\right)^{\tau^{\text{BP}}_{ir}}\right]}.$$

Variational Bayes EM (VB)

The variational approximation can be applied to Bayesian inference; the parameter $\theta = (\alpha, \pi)$ is also viewed as an unobserved variable (*Beal and Ghahramani (2003)*).
We aim at finding

$$\mathcal{R}_X^* = \arg\min KL(\mathcal{R}_X(\theta, Z), P(\theta, Z|X))$$

If conjugate priors are used:

$$P(X, \theta, Z) \propto \exp\{\phi(\theta)'[u^0 + u(X, Z)]\},$$

close-form approximate conditional distributions of the form

$$\mathcal{R}_X(\theta, Z) = \mathcal{R}_{X,\theta}(\theta)\mathcal{R}_{X,Z}(Z)$$

can be derived:

$$\begin{aligned} \mathcal{R}_{X,\theta}(\theta) &\propto \exp\{\phi(\theta)'\tilde{u}(X)\} \qquad \tilde{u}(X) = u^0 + \overline{u}(X) \\ \mathcal{R}_{X,Z}(Z) &\propto \exp\{\overline{\phi}' u(X, Z)\} \end{aligned}$$

VB for MixNet

For Dirichlet and Beta priors, we get:

$\mathcal{R}_{X,Z}(Z)$.

$$\tau_{iq}^{\mathsf{VB}} \quad \propto \quad e^{\psi(\widetilde{n}_q)-\psi\left(\sum_{l=1}^{Q}\widetilde{n}_\ell\right)} \prod_{j\neq i}^{n}\prod_{l=1}^{Q} e^{\tau_{j\ell}^{\mathsf{VB}}\left\{\psi(\widetilde{\zeta}_{q\ell})-\psi(\widetilde{\eta}_{q\ell}+\widetilde{\zeta}_{q\ell})+X_{ij}[\psi(\widetilde{\eta}_{q\ell})-\psi(\widetilde{\zeta}_{q\ell})]\right\}}$$

where ψ is the first derivative of the Γ function.

$\mathcal{R}_{X,\theta}(\theta)$.

$$(\alpha|X) \approx \mathcal{D}(\widetilde{n}), \qquad (\pi_{q\ell}|X) \approx \mathsf{B}(\widetilde{\eta}_{q\ell}, \widetilde{\zeta}_{q\ell})$$

where

$$\begin{aligned}
\widetilde{n}_q &= n_q + \textstyle\sum_i \tau_{iq}^{\mathsf{VB}}, \\
\widetilde{\eta}_{q\ell} &= \eta_{q\ell} + \left(1 - \tfrac{1}{2}\mathbb{1}_{q=l}\right) \textstyle\sum_{i\neq j} X_{ij}\tau_{iq}^{\mathsf{VB}}\tau_{j\ell}^{\mathsf{VB}}, \\
\widetilde{\zeta}_{q\ell} &= \zeta_{q\ell} + \left(1 - \tfrac{1}{2}\mathbb{1}_{q=l}\right) \textstyle\sum_{i\neq j}(1 - X_{ij})\tau_{iq}^{\mathsf{VB}}\tau_{j\ell}^{\mathsf{VB}}.
\end{aligned}$$

Simulation Design

2-group MixNet model with parameters:

$$\text{Case 1: } \alpha = \begin{pmatrix} 0.6 & 0.4 \end{pmatrix}, \pi = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.5 \end{pmatrix};$$

$$\text{Case 2: } \alpha = \begin{pmatrix} 0.6 & 0.4 \end{pmatrix}, \pi = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.3 \end{pmatrix}.$$

500 graphs are simulated for each case and each graph size.

The complete comparison of the 4 methods is only made on small graphs ($n = 18$) because of the computation time required by for EM.

Estimates, standard deviation and likelihood

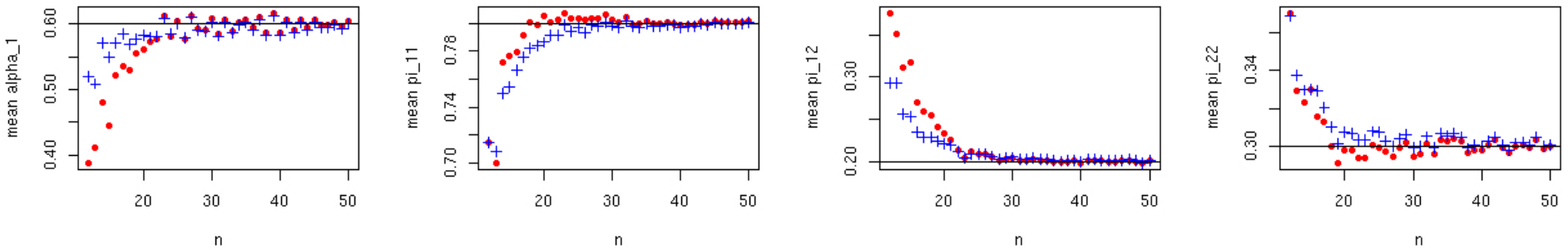
$n = 18$	α_1	π_{11}	π_{12}	π_{22}	$\log P(X)$
True value	60%	80%	20%	50%	
EM	59.1 (13.1)	78.5 (13.5)	20.9 (8.4)	50.9 (15.4)	-90.68
VEM	57.7 (16.6)	78.8 (12.4)	22.4 (10.7)	50.3 (14.6)	-90.87
BP	57.9 (16.2)	78.9 (12.3)	22.2 (10.5)	50.3 (14.5)	-90.85
VB	58.1 (13.3)	78.2 (9.7)	21.6 (7.7)	50.8 (13.3)	-90.71
True value	60%	80%	20%	30%	-
EM	59.5 (14.1)	78.7 (15.6)	21.2 (8.7)	30.3 (14.3)	-88.18
VEM	55.6 (19.0)	80.1 (14.0)	24.0 (11.8)	30.8 (13.8)	-88.54
BP	56.6 (17.8)	80.0 (13.6)	23.2 (11.0)	30.8 (13.8)	-88.40
VB	58.4 (14.6)	77.9 (12.0)	22.3 (9.3)	32.1 (12.3)	-88.26

- All methods provide similar results.
- EM achieves the best ones.
- BP does not significantly improve VEM.

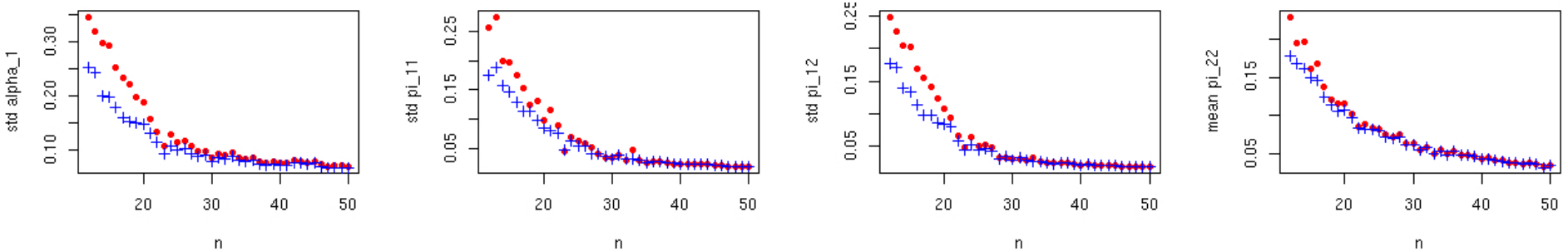
Influence of the graph size

Comparison of VEM: • and VB: + in case 2 (difficult).
Left to right: α_1, π_{11}, π_{12}, π_{22}.

Means.



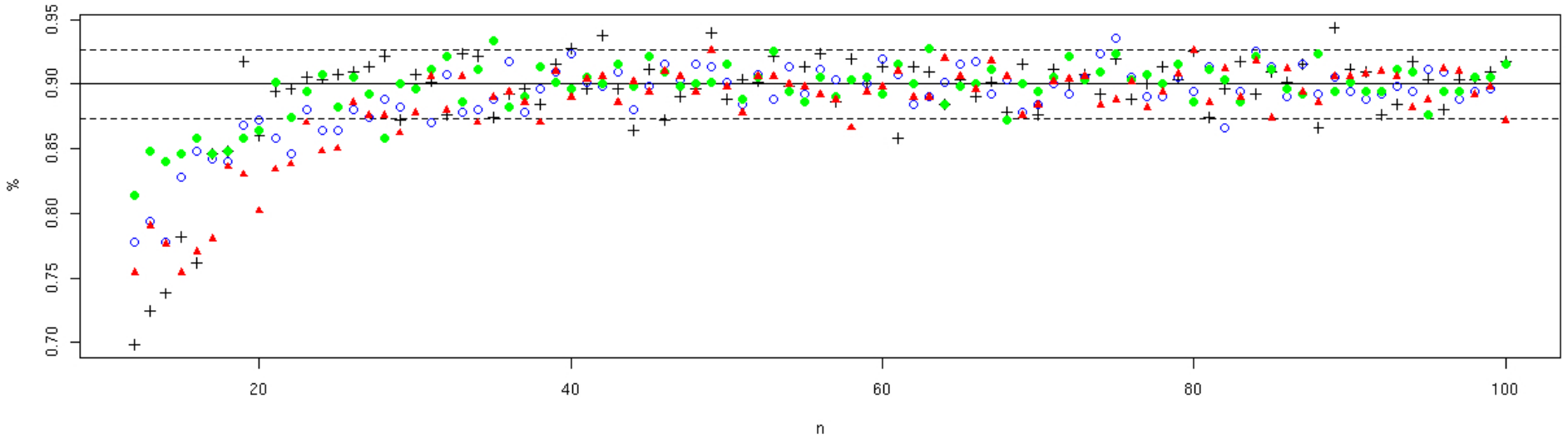
Standard deviations.



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

VB Credibility intervals

Actual level as a function of n.

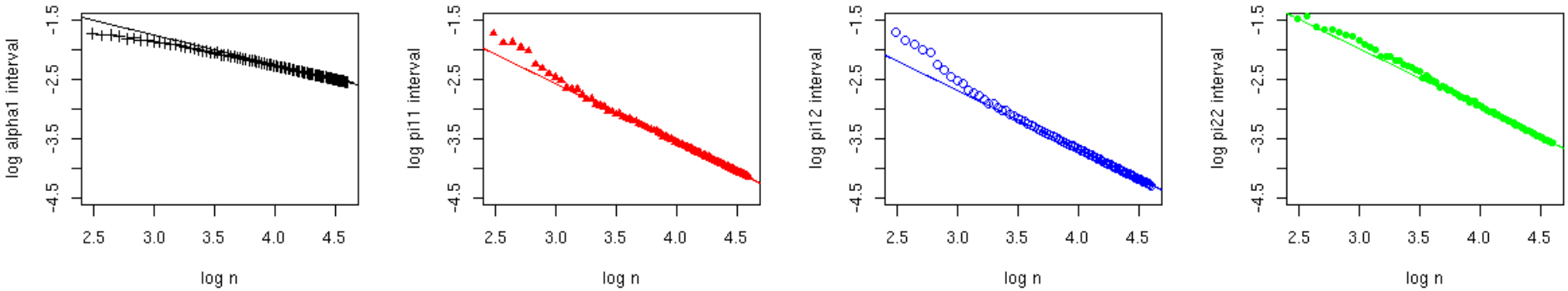


α_1: $+$, π_{11}: $\triangle$, π_{12}: $\circ$, π_{22}: $\bullet$

- For all parameters, VB posterior credibility intervals achieve the nominal level (90%), as soon as $n \geq 25$.
- $\longrightarrow$ the VB approximation seems to work well.

Convergence rate of the VB estimates

Width of the posterior credibility intervals. α_1, π_{11}, π_{12}, π_{22}



- The width decreases as $1/\sqrt{n}$ for α_1.
- It decreases as $1/n$ for π_{11}, π_{12} and π_{22}.
- Consistent with the penalty of the ICL criterion of *Daudin* et al. *(2008)*:

$$(Q-1)\log n + Q^2 \log[n(n-1)/2].$$

Why does VB work so well? (off the record)

Work in progress: Daudin & Celisse are about to prove the concentration of $P(Z|X)$ around the true value z^*, i.e.

$$P(Z|X) \underset{n\to\infty}{\longrightarrow} \delta_{z^*}(Z)$$

Intuition: If this holds,

(*i*) The limit distribution $\delta_{z^*}(Z)$ belongs to the distribution class over which VB approximation achieves maximisation, so it is reached;

(*ii*) The joint conditional distribution $P(\theta, Z|X) = P(\theta|Z, X)P(Z|X)$ tends to $P(\theta|Z, X)\delta_{z^*}(Z)$. Again $P(\theta|Z, X)$ belongs to the distribution class of VB, so it is also reached;

And the variational approximation tends to be ... exact.

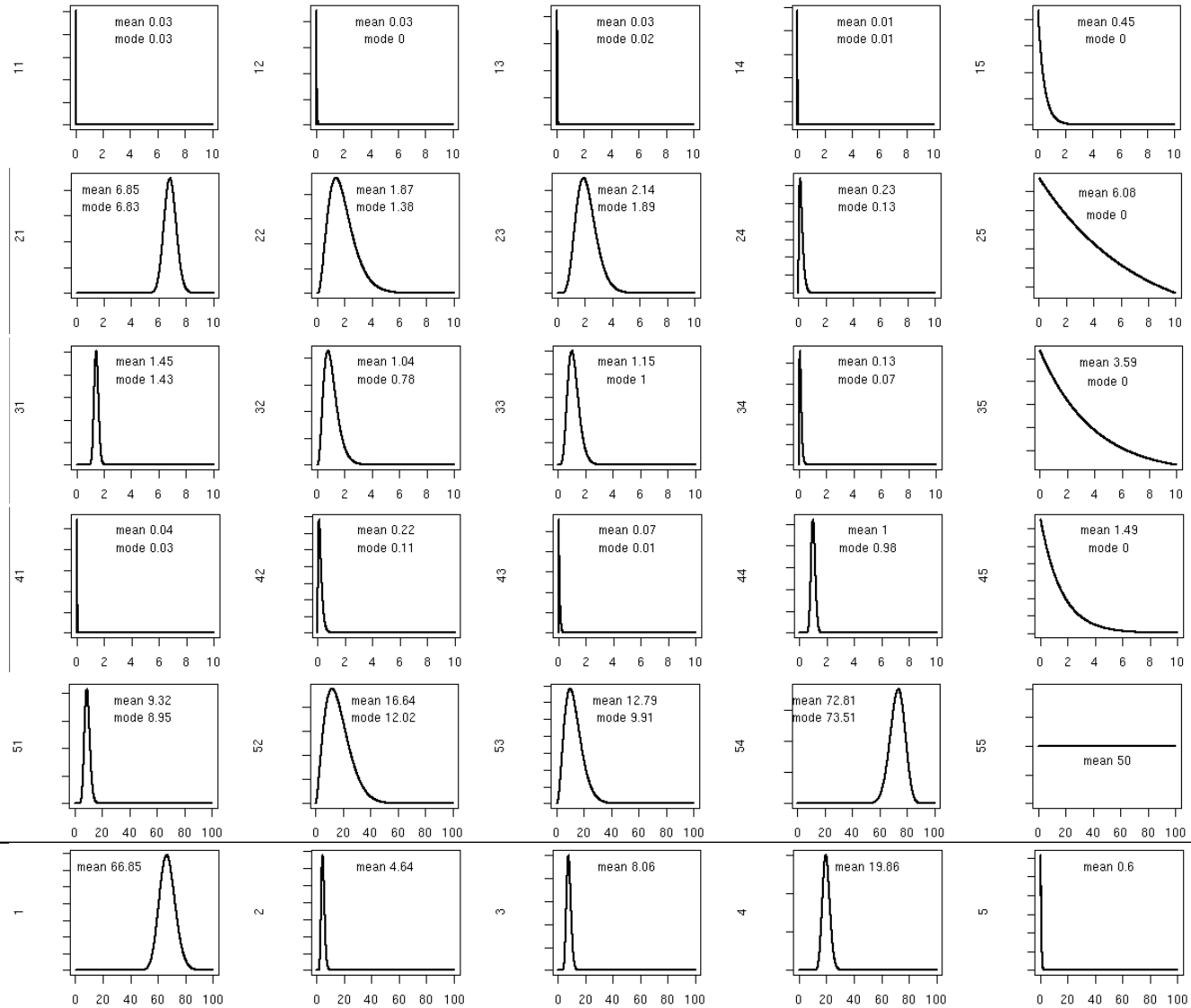
Comparison of VEM and VB

Network are n = 338 operons, linked if one encodes a transcription factor that directly regulates the other one.

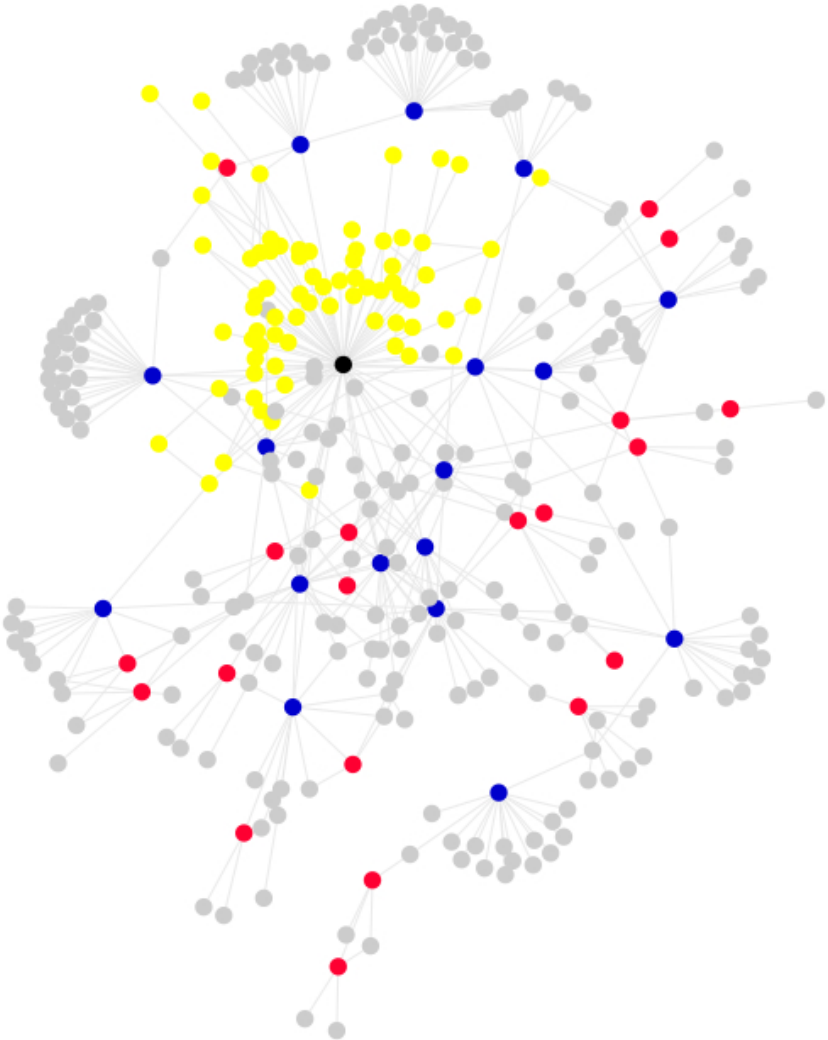
π_{ql}	1	2	3	4	5
1	0.03	0.00	0.03	0.00	0.00
2	6.40	1.50	1.34	0.44	0.00
3	1.21	0.89	0.58	0.00	0.00
4	0.00	0.09	0.00	0.95	0.00
5	8.64	17.65	0.05	72.87	11.01
α	65.49	5.18	7.92	21.10	0.30
1	[0.02;0.04]	[0.00;0.10]	[0.01;0.08]	[0.00;0.03]	[0.02;1.34]
2	[6.14;7.60]	[0.61;3.68]	[1.07;3.50]	[0.05;0.54]	[0.33;17.62]
3	[1.20;1.72]	[0.35;2.02]	[0.56;1.92]	[0.03;0.30]	[0.19;10.57]
4	[0.01;0.07]	[0.04;0.51]	[0.01;0.20]	[0.76;1.27]	[0.08;4.43]
5	[6.35;12.70]	[4.60;33.36]	[4.28;24.37]	[63.56;81.28]	[5.00;95.00]
α	[59.65;74.38]	[2.88;6.74]	[5.68;10.77]	[16.02;24.04]	[0.11;1.42]

VEM and VB estimates for the $Q = 5$ group model (approximate 90% credibility intervals).

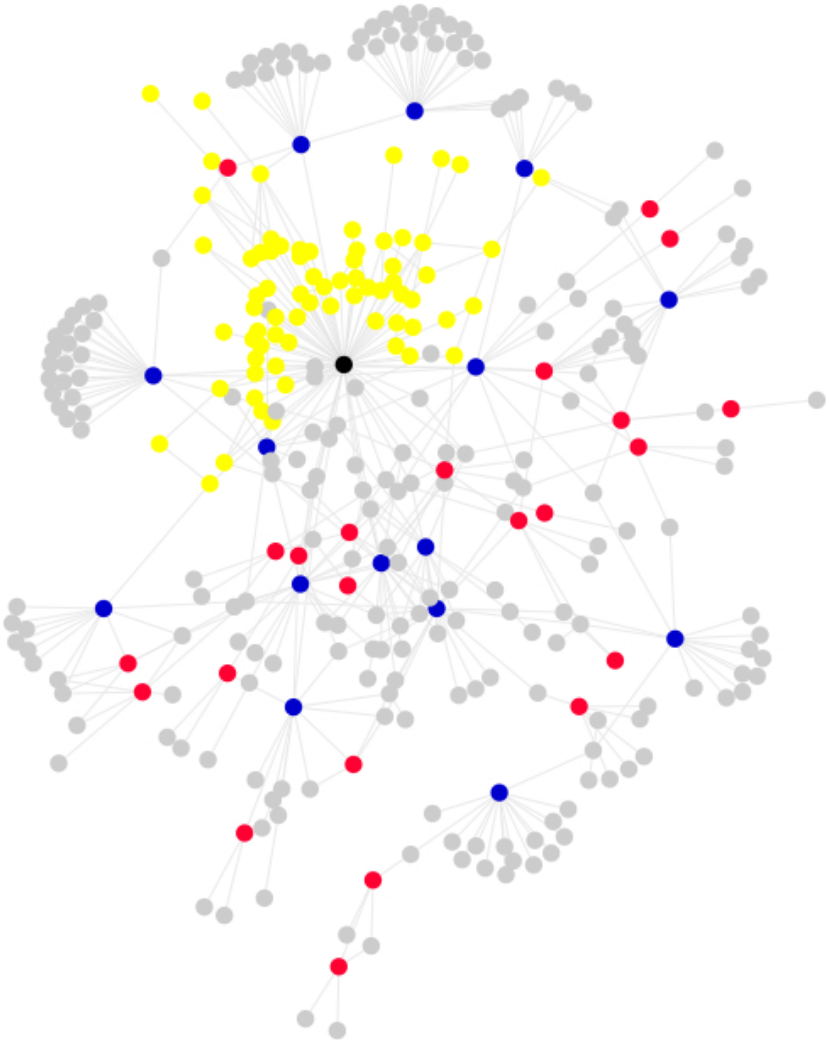
Approximate posterior distribution



Comparison of VEM and VB classifications



VEM



VB

- Only 4 nodes have different classifications.

Beal, J., M. and Ghahramani, Z. (2003). The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. In *Bayesian Statistics 7*, 543–52. Oxford University Press.

Daudin, J.-J., Picard, F. and Robin, S. (2008). A mixture model for random graphs. *Stat Comput*. **18** 173–183.

Jaakola, T. S. (2000). Tutorial on variational approximation methods.

MacKay, D. J. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.