Properties of variational estimates of a mixture model for random graphs

Jean-Jacques Daudin
UMR AgroParisTech-INRA518

ECCS Lisbonne 9/16/2010,
Why do people represent data by a network (1)?

- Real networks do exist: electric, transport or www networks... They have been represented for a long time by virtual networks.
- Virtual network is a nice way for representing or even "modelling" many scientific phenomenons: social relations, metabolic pathways, chemical reactions...

FIG. 5: The karate club network of Zachary (figure taken from Girvan and Newman [18]).
Why do people represent data by a network (2)?

- an overall representation of the interactions between many nodes
- the plot reveals the topology of the networks
- nodes may be colored, adding more information

Figure 2 | Yeast protein interaction network. A map of protein-protein interactions in Saccharomyces cerevisiae, which is based on early yeast two-hybrid measurements, illustrates that a few highly connected nodes (which are also known as hubs) hold the network together. The largest cluster, which contains ~78% of all proteins, is shown. The colour of a node indicates the phenotypic effect of removing the corresponding protein (red = lethal, green = non-lethal, orange = slow growth, yellow = unknown). Reproduced with permission from REF.18 © Macmillan Magazines Ltd.
An unusual data set structure

In the relational data set, the core information is the relation between two items.

- lines are not independent
- the data structure is similar to distance, similarity, covariance or correlation matrices
What do we want?

A simple representation of a complex graph, using meta-vertices and meta-edges.

FIG. 5: The karate club network of Zachary (figure taken from Girvan and Newman [18]).
Stochastic Block Model (SBM) a mixture model for random graphs, Snijders and Nowicki (1997), Daudin (2008)

- $i = 1, n$ nodes
- $q = 1, Q$ classes
- $X_{ij} = 1$ if there is an edge from node $i$ to node $j$.
- $Z = Z_{iq}$ discrete latent variable, $Z_{iq} = 1$ if node $i$ pertains to class $q$
- $(Z_{i1}, Z_{i2}, ..., Z_{iQ}) \sim \mathcal{M}(1, \alpha_1, \alpha_2, ..., \alpha_Q)$
- Conditionally to $Z$, $X_{ij}$ are independent Bernoulli RV with

\[ P(X_{ij} = 1/Z_i = q, Z_j = l) = \pi_{ql} \]
# SBM: a flexible model

<table>
<thead>
<tr>
<th>Description</th>
<th>Graph</th>
<th>Q</th>
<th>π</th>
</tr>
</thead>
<tbody>
<tr>
<td>Erdos, no cluster</td>
<td><img src="image1.png" alt="Graph" /></td>
<td>1</td>
<td>(p)</td>
</tr>
</tbody>
</table>
| Hubs                               | ![Graph](image2.png) | 4 | \[
0 1 0 0 \\
1 0 1 0 \\
0 1 0 1 \\
0 0 1 0
\] |
| cluster in "usual sense"           | ![Graph](image3.png) | 2 | \[
1 \varepsilon \\
\varepsilon 1
\] |
| Hierarchical                       | ![Graph](image4.png) | 5 | \[
0 1 1 0 0 \\
0 0 0 1 0 \\
0 0 0 0 1 \\
0 0 0 0 0 \\
0 0 0 0 0
\] |
Log-Likelihood is untractable

- **Complete data likelihood**

\[ \mathcal{L}(X, Z) = \sum_i \sum_q Z_{iq} \ln \alpha_q + \sum_{i < j} \sum_{q, l} Z_{iq} Z_{jl} \ln b(\pi_{ql}, X_{ij}) \]

where \( b(\pi_{ql}, X_{ij}) = \pi_{ql}^{X_{ij}} (1 - \pi_{ql})^{(1-X_{ij})} \)

- **Observed data likelihood**

\[ \mathcal{L}(X) = \ln \sum_Z \exp \mathcal{L}(X, Z) \]

- observed data likelihood requires a sum over \( Q^n \) terms: **untractable**

- EM-like strategies require \( \Pr(Z|X) \): **untractable** (no conditional independence).
**Main Idea:** Replace complicated $\Pr(Z|X)$ by a simple $R_X[Z]$ such that $\text{KL}(R_X[Z], \Pr(Z|X))$ is minimal.

- Optimize in $R_X$ the function $J(R_X)$ given by:

  $$J(R_X[Z]) = \mathcal{L}(X) - \text{KL}(R_X[Z], \Pr(Z|X))$$

  $$= \mathcal{H}(R_X[Z]) - \sum_Z R_X[Z] \mathcal{L}(X, Z)$$

- For simple $R_X$, $J(R_X[Z])$ is tractable,

- At best, $R_X = \Pr(Z|X)$ and $J(R_X[Z]) = \mathcal{L}(X)$. 

Variational estimates for SBM

J.J. Daudin

Background

New results about SBM

Simulations

Analysis of a large PPIN

References
2 Steps Iterative Algorithm, (Daudin et al., 2008)

- **Step 1** Optimize $J(\mathcal{R}_X[Z])$ w.r.t. $\mathcal{R}_X[Z]$:
  - Restriction to a "comfortable" class of functions,
  - $\mathcal{R}_X[Z] = \prod_i h(Z_i; \tau_i)$, with $h(.; \tau_i)$ the multinomial distribution,
  - $\tau_{iq}$ is a variational parameter to be optimized using a fixed point algorithm:
    \[
    \tilde{\tau}_{iq} \propto \alpha_q \prod^{Q}_{j=1} b(\pi_{ql}, X_{ij}) \tilde{\tau}_{jl}
    \]

- **Step 2** Optimize $J(\mathcal{R}_X[Z])$ w.r.t. $(\alpha, \pi)$:
  - Constraint: $\sum_q \alpha_q = 1$
  - $\tilde{\alpha}_q = \sum_i \tilde{\tau}_{iq} / n$
  - $\tilde{\pi}_{ql} = \sum_{ij} \tilde{\tau}_{iq} \tilde{\tau}_{jl} X_{ij} / \sum_{ij} \tilde{\tau}_{iq} \tilde{\tau}_{jl}$
Toy-Example: Karate Club

- nodes are members of the club
- edges between 2 members if they have social relation outside the club
- known properties: the club has split away in two parts (cercles and squares).

SBM results for Karate Club

The split is recovered and the role of the leaders is underlined.
Known results about the consistency of variational estimates

- The variational method is practically effective with many applications till $n = 3000$ (see Picard et al. and http://stat.genopole.cnrs.fr/software/mixnet/).
- But till now no theoretical property is known.

No general result about variational estimates: they maximize a pseudo-likelihood and no general properties have been established.

- Gunawardana and Byrne show that the variational estimates are consistent only for degenerate cases.
- VE have been proved to be consistent in some cases, Markovian models (Hall et al.), latent variable models (Consonni et al. and normal mixture model (Woolrich et al.))
- and not consistent for state space models (Wang et al.)
Consistency of variational estimates

C1: \( \forall (q \neq q') \ \exists l \in (1, Q) : \pi_{ql} \neq \pi_{ql'} \) or \( \pi_{lq} \neq \pi_{lq'} \),
C2: \( \exists a > 0 : \min(\min(\pi_{ql} > 0), \min(1 - \pi_{ql}) > 0) \geq a \),
C3: \( \exists b > 0 : \min(\alpha_q) \geq b \).

Under C1, C2 and C3, the variational estimates \((\hat{\pi}, \hat{\alpha})\) are consistent and asymptotically equivalent to the maximum likelihood estimates. Moreover, when \( n \to \infty \),

\[
\frac{1}{n^2} \left[ \mathcal{L}(x[n]; \alpha, \pi) - J(x[n]; \tau[n], \pi, \alpha) \right] \xrightarrow{P} 0
\]

and

\[
\overline{\tau[n]} \xrightarrow{P} z^*,
\]

with \( z^* \) being the true value for \( Z \).
The proof uses
- the properties of $\mathcal{I}$,
- concentration inequalities
- an extension of classical methods using Empirical Processes for proving consistency.

Two properties of networks data and the model are important in this proof:
- there are $n^2$ data → strong concentration inequalities
- the asymptotic pdf of $Z|X$ pertains to the factorized class of pdfs in which the variational approximation is searched.

These properties are rarely shared by other data sets and models so the proof is specific to random networks.
Variational Bayes(1)

- **Bayesian setting**: parameters are viewed as unobserved variables
- Two sets of unobserved variables: $Z$ and $\theta = (\alpha, \pi)$.
- Conjugate priors, Latouche et al. give closed-form approximate conditional distributions of both $Z$ and $\theta$.
- Same results can be obtained as an application of the general variational Bayes method with exponential families given by Beal and Ghahramani.
Variational Bayes (2)

Prior $\alpha \sim \mathcal{D}(n^0)$, $\pi_{q\ell} \sim \mathcal{B}(\eta_{q\ell}^0, \zeta_{q\ell}^0)$, $n^0 = (n_1^0, \ldots, n_Q^0)$

$\mathcal{D}$: Dirichlet, $\mathcal{B}$: Beta.

Approximate conditional mean of $Z_{iq}$

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

$\psi$ digamma function.

Approximate posterior $(\alpha|X) \approx \mathcal{D}(\tilde{n})$, $(\pi_{q\ell}|X) \approx \mathcal{B}(\tilde{\eta}_{q\ell}, \tilde{\zeta}_{q\ell})$

$$
\tilde{n}_q = n_q^0 + \sum_i \tau_{iq}^{\text{VB}},
$$

$$
\tilde{\eta}_{q\ell} = \eta_{q\ell}^0 + \left(1 - \frac{1}{2} \mathbf{1}_{q=l}\right) \sum_{i \neq j} X_{ij} \tau_{iq}^{\text{VB}} \tau_{j\ell}^{\text{VB}},
$$

$$
\tilde{\zeta}_{q\ell} = \zeta_{q\ell}^0 + \left(1 - \frac{1}{2} \mathbf{1}_{q=l}\right) \sum_{i \neq j} (1 - X_{ij}) \tau_{iq}^{\text{VB}} \tau_{j\ell}^{\text{VB}}.
$$
Simulation design

- $n = 2, 4, \ldots, 50$
- $\alpha = \begin{pmatrix} 0.6 & 0.4 \end{pmatrix}$, $\pi = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.3 \end{pmatrix}$.
- 500 graphs for each graph size.
Consistency and precision of VEM and VB

Figure: Mean (top) and standard deviations (bottom) of the estimates. From left to right: $\alpha_1$, $\pi_{11}$, $\pi_{12}$, $\pi_{22}$. VEM: red circles, VB: blue crosses.
Figure: Proportion of the simulations where interval with credibility 90% contain the true value of the parameter. $\alpha_1$: black crosses, $\pi_{11}$: red triangles, $\pi_{12}$: blue circles, $\pi_{22}$: green solid circles. Binomial confidence interval: dotted lines.
**Figure:** Width of the 90% credibility as a function of the graph size (in log scale). From left to right: $\alpha_1$, $\pi_{11}$, $\pi_{12}$ and $\pi_{22}$. Straight lines have slope $-0.5$ for $\alpha_1$ and $-1$ for the three others.
MS-Interactome data

- MS-Interactome (Ewing et al.): first large-scale study of protein-protein interactions in human cells using a mass spectrometry approach.
- 3,494 interactions between 1,561 proteins
- Bait proteins chosen based on known functional annotation and implied disease association.
- One third of the 338 bait proteins are disease-related ones, mainly involved in cancer.
- Data previously analyzed by Marras et al. using a two-steps procedure: first a deterministic method allows to find large core and community structures and second a stochastic method (such as mixture model) permits to uncover fine-grained interactome components.
- The following analysis is made using VEM method using package Mixnet.
ICL = \mathcal{J}(x[n]; \tau[n], \pi, \alpha) - (Q - 1) \log n - \frac{Q(Q + 1)}{2} \log \left[ \frac{n(n - 1)}{2} \right]

AIC = \mathcal{J}(x[n]; \tau[n], \pi, \alpha) - (Q - 1) - \frac{Q(Q + 1)}{2}

Best choices: Q = 23 (AIC) and Q = 8 (ICL).
Description of the first groups. The proteins have been affected to one group if their probability of pertaining to the group is greater than 0.5.

<table>
<thead>
<tr>
<th>group</th>
<th># proteins</th>
<th># unrecognized proteins</th>
<th>GO Term</th>
<th>Corrected P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>44</td>
<td>2</td>
<td>Cellular metabolic Process &amp; Apoptose</td>
<td>$4.10^{-7}$</td>
</tr>
<tr>
<td>2</td>
<td>79</td>
<td>11</td>
<td>RNA Processing</td>
<td>$5.10^{-3}$</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td></td>
<td>cell proliferation</td>
<td>$8.10^{-3}$</td>
</tr>
<tr>
<td>4</td>
<td>211</td>
<td>24</td>
<td>intracellular transport</td>
<td>$9.10^{-8}$</td>
</tr>
<tr>
<td>5</td>
<td>55</td>
<td>11</td>
<td>macromolecule localization</td>
<td>$1.10^{-4}$</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td></td>
<td>protein targeting and transport</td>
<td>$1.10^{-6}$</td>
</tr>
<tr>
<td>7</td>
<td>353</td>
<td>57</td>
<td>Cellular metabolic Process</td>
<td>$5.10^{-12}$</td>
</tr>
<tr>
<td>8</td>
<td>111</td>
<td>12</td>
<td>macromolecule modification</td>
<td>$3.10^{-16}$</td>
</tr>
<tr>
<td>9</td>
<td>372</td>
<td>73</td>
<td>protein complex assembly</td>
<td>$3.10^{-8}$</td>
</tr>
<tr>
<td>10</td>
<td>96</td>
<td>14</td>
<td>phosphorylation</td>
<td>$7.10^{-7}$</td>
</tr>
<tr>
<td>11</td>
<td>5</td>
<td>2</td>
<td>negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle</td>
<td>$1.10^{-5}$</td>
</tr>
<tr>
<td>12</td>
<td>15</td>
<td></td>
<td>negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle</td>
<td>$2.10^{-38}$</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td></td>
<td>RNA metabolic process</td>
<td>$1.10^{-2}$</td>
</tr>
</tbody>
</table>
More about the groups

- most of the groups can be identified by at least one GO term with low corrected P-values
- 234 proteins were not recognized by GO term Finder → SBM proposes a classification for unknown proteins.
- 17th group composed of two proteins highly related with tumor progression: the Von Hippel Lindau (VHL) tumor suppression protein and MCC, which blocks cell cycle progression.
- group 13, composed of two proteins Tgfb1i4 (transforming growth factor beta 1 induced transcript), which is a growth factor, and RNSP1, which is a part of a post-splicing multiprotein complex regulating exons.
Meta-Network obtained with SBM

Figure: Representation of the 18 groups obtained with SBM. Edges between two nodes are present only if the probability of connection between them is greater than 0.015. The size of each node and the size of the police are proportional to the number of proteins contained in it. The width of the edges are proportional to the probability of connection between the corresponding nodes.
Conclusions

- SBM is a flexible model which allows to replace a complicated network by a simple meta-network,
- VEM and VB are theoretically validated,
- It is possible to use SBM to cluster large networks,
- Freely available package, URL: http://pbil.univ-lyon1.fr/software/MixNet.
Common work with

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
<th>Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alain Celisse</td>
<td>UMR CNRS 8524, Univ. Lille 1</td>
<td>Consistency of variationnal Estimates</td>
</tr>
<tr>
<td>Steven Gazal</td>
<td>INSERM, Paris</td>
<td>Simulations</td>
</tr>
<tr>
<td>Stéphane Robin</td>
<td>UMR 518 AgroParisTech/INRA</td>
<td>Variationnal Bayes</td>
</tr>
</tbody>
</table>


Variational method

PPIN Example
