A reminder on statistical modelling: Poisson log-normal and stochastic block models

S. Robin

NGB, Feb. 2019

Statistical modelling

.. the art of translating a scientific question into mathematical equations.

Typical experiment.

Statistical modelling

.. the art of translating a scientific question into mathematical equations.

Typical experiment.

A question

An experimental design

Data

Analysis (model)

Statistical modelling

.. the art of translating a scientific question into mathematical equations.

Typical experiment.

A question

An experimental design

Data x, Y

Analysis (model)

Statistical modelling

.. the art of translating a scientific question into mathematical equations.

Typical experiment.

A question

An experimental design

Data x, Y

Analysis (model) $Y = f(x, \theta)$

Statistical modelling

.. the art of translating a scientific question into mathematical equations.

Typical experiment.

A question	$\theta =$?
An experimental design	
Data	x, Y
Analysis (model)	$Y = f(x, \theta)$

Statistical modelling

.. the art of translating a scientific question into mathematical equations.

Typical experiment.

A question	$\theta =?$
An experimental design	$x =?$
Data	x, Y
Analysis (model)	$Y = f(x, \theta)$

- $Y =$ response, variable of interest
- $x =$ covariates, environmental conditions, treatments, ...
- $\theta =$ unknown parameters: $\theta = (\alpha, \beta, \gamma, \mu, \pi, \sigma^2, ...)$

Outline

Generalized linear model (GLM)

Poisson log-normal model (PLN)

Stochastic blockmodel (SBM)

Outline

A first example

Question. Does the depth affect the abundance of the deepwater redfish (*Sebastes mentella*)?

Experiment. Sampling in $n = 89$ stations

Data.

- $x =$ depth,
- $Y =$ abundance

```
##   Depth Abundance
## 1   349         7
## 2   382        93
## 3   294        37
## 4   304         0
## 5   384       958
## 6   344       101
```

A first model: Poisson regression

Model. The stations $i = 1 \dots n$ are independent and ($\mathcal{P}$ = Poisson)

$$Y_i \sim \mathcal{P}(e^{\beta_0 + \beta_1 x_i})$$

$\theta = (\beta_0, \beta_1)$: β_0 = intercept, β_1 = effect of the depth.

```
GLM1 = glm(Abundance ~ Depth, data=data.glm1, family='poisson')
```

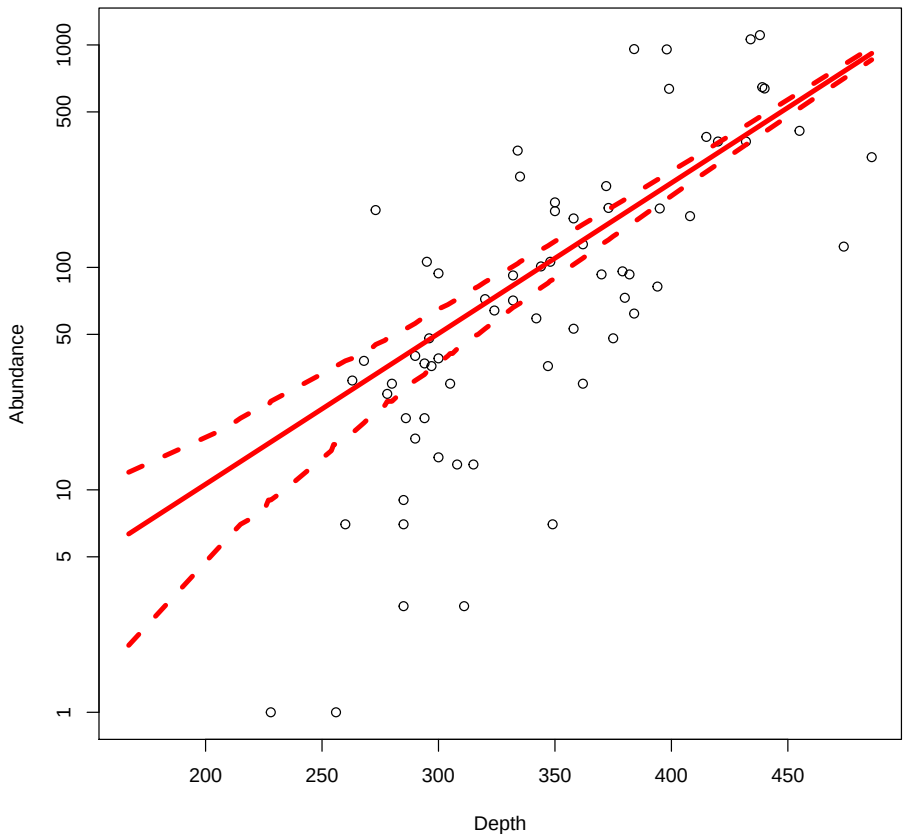
Results.

```
summary(GLM1)$coef

##                Estimate   Std. Error   z value     Pr(>|z|)
## (Intercept) -0.75795114 0.0586298812 -12.92773 3.139613e-38
## Depth        0.01559353 0.0001458807 106.89234 0.000000e+00
```

- Significant effect of the depth: '$\beta_1 = 0$' rejected
- Mean abundance multiplied by $e^{\widehat{\beta}_1 100}$ every 100m:

```
exp(GLM1$coef[2]*100)

##    Depth
## 4.755745
```

Fit



- Overdispersion?

More than one covariate (1/2)

Question. Effect of depth and temperature on deepwater redfish

Data. Same stations

- $x_1 =$ depth
- $x_2 =$ temperature
- $Y =$ abundance

```
##   Depth Temperature Abundance
## 1   349        3.95         7
## 2   382        3.75        93
## 3   294        3.45        37
## 4   304        3.65         0
## 5   384        3.35       958
## 6   344        3.65       101
```

Model. Independent stations and

$$Y_i \sim \mathcal{P}(e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}})$$

$\rightarrow$ Multiplicative effect of the temperature ($\times e^{\beta_2}$ every ^oC)

More than one covariate (2/2)

Vector form (compact).

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} = x_i^\intercal \beta, \qquad \text{where} \quad x_i = \begin{bmatrix} 1 \\ x_{i1} \\ x_{i2} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

- $x_i =$ vector of covariates:

$$Y_i \sim \mathcal{P}(e^{x_i^\intercal \beta})$$

- parameter: $\theta = \beta = (\beta_0, \beta_1, \beta_2)$

Results.

```
##                Estimate   Std. Error   z value      Pr(>|z|)
## (Intercept) -2.20356570 0.0767733013 -28.70224 3.577417e-181
## Depth        0.01695337 0.0001595502 106.25727  0.000000e+00
## Temperature  0.45769759 0.0118927306  38.48549  0.000000e+00
```

- Interpretation?

Offset, overdispersion

Sampling effort. 2h sampling in station 1, 5h sampling in station 2:

- under fixed conditions: $\mathbb{E}Y_2 = 2.5 \times \mathbb{E}Y_1$
- define offsets: $o_1 = \log(2)$, $o_2 = \log(5)$, $o_3 = ...$

$$Y_i \sim \mathcal{P}(e^{o_i + x_i^\intercal \beta})$$

Overdispersion. Typical in count data

- good fit for the mean $\mathbb{E}(Y_i)$, bad fit for the variance ($\mathbb{V}(Y_i)$ too small)
- add a random term:

$$Y_i \sim \mathcal{P}(e^{x_i^\intercal \beta + Z_i}) \qquad \text{where} \quad \left\{ \begin{array}{rcl} Z_i & \sim & \mathcal{N}(0, \sigma^2) \\ \log Z_i & \sim & \mathcal{G}\text{am}(a, a) \end{array} \right.$$

$\rightarrow$ generalized linear mixed model: $\theta = (\beta, \sigma^2)$ or $\theta = (\beta, a)$

Generalized linear models

General form. Modelling the mean response:

$$g\left(\mathbb{E}(Y_i)\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots = x_i^\intercal \beta$$

$g =$ *link* function (identity, log, logit, ...)

Estimation.

- Most popular method: maximum likelihood (ML)

$$\max_\theta p_\theta(\{Y_i\}; \{x_i\}) = \max_\theta \prod_{i=1}^n p_\theta(Y_i; x_i)$$

- Easy if the likelihood $p_\theta(\{Y_i\}; \{x_i\})$ is nice (depends on g)
- Estimates and tests for combination of parameters: e.g. $\beta_1 - \beta_2$
- More difficult for mixed models, because the Z_i are not observed

Outline

Joint modelling of species abundance

Questions.

- Which environmental conditions do affect the abundance of p species?
- Do the respective species abundances vary independantly from each other?

Data.

- n sites or samples, p species
- Y_{ij} = abundance of species j in site i
- x_i = vector of d covariates for site i
- o_{ij} = offset (e.g. sampling effort) for species j in site i

Barents data

- $n = 89$ stations from the Barents sea
- $p = 30$ fish species
- $d = 4$ covariates (latitude, longitude, depth, temperature)
- same sampling effort for all species in all stations

$X : n \times d$ matrix

```
##       Latitude Longitude Depth
## [1,]    71.10     22.43   349
## [2,]    71.32     23.68   382
## [3,]    71.60     24.90   294
## [4,]    71.27     25.88   304
## [5,]    71.52     28.12   384
## [6,]    71.48     29.10   344
```

$Y : n \times p$ matrix

```
##       Re_hi An_de An_mi Hi_pl
## [1,]     0     0     0    31
## [2,]     0     0     0     4
## [3,]     0     0     0    27
## [4,]     0     0     1    13
## [5,]     0     0     0    23
## [6,]     1     0     0    20
```

- No offset

Joint modelling of species abundances

Modelling. Need to account for

- the nature of abundance data (counts + overdispersion)
- the effect of each covariate on each species
- the 'correlation' between each pair of species
- the specificities of the experimental design (e.g. unbalanced sampling efforts)

Poisson log-normal model (PLN)

Model.

- for each site i, draw a random (independent, latent) vector

$$Z_i \sim \mathcal{N}_p(0, \Sigma)$$

- for each site i and each species, draw the (independent) abundance

$$Y_{ij} \sim \mathcal{P}(e^{o_{ij} + x_{ij}^\mathsf{T} \beta + Z_{ij}})$$

Model parameters. $\theta = (\beta, \Sigma)$

- $\beta_{hj} =$ effect of covariate h on species j
- $\beta = [\beta_{hj}] : d \times p$ matrix of regression coefficients
- $\sigma_j^2 =$ additional variability of species j
- $\sigma_{jk} =$ 'covariance' between species j and k
- $\Sigma = [\sigma_{jk}] : p \times p$ covariance matrix

Example: 3 species, 2 covariates

$\widehat{\beta}$

```
##                Re_hi   Se_me   Se_ma
## (Intercept) -3.150 -5.590  1.292
## Depth        0.015  0.023 -0.012
## Temperature -0.864  0.536  0.748
```

$\widehat{\Sigma}$

```
##         Re_hi   Se_me   Se_ma
## Re_hi   0.750   0.334 -0.393
## Se_me   0.334   3.260 -0.640
## Se_ma -0.393 -0.640  3.351
```

Overdispersion $\widehat{\sigma}_j$

```
## Re_hi Se_me Se_ma
## 0.866 1.806 1.831
```

Correlations $\widehat{\rho}_{jk}$

```
##         Re_hi   Se_me   Se_ma
## Re_hi   1.000   0.213 -0.248
## Se_me   0.213   1.000 -0.194
## Se_ma -0.248 -0.194  1.000
```

Barents: no covariate

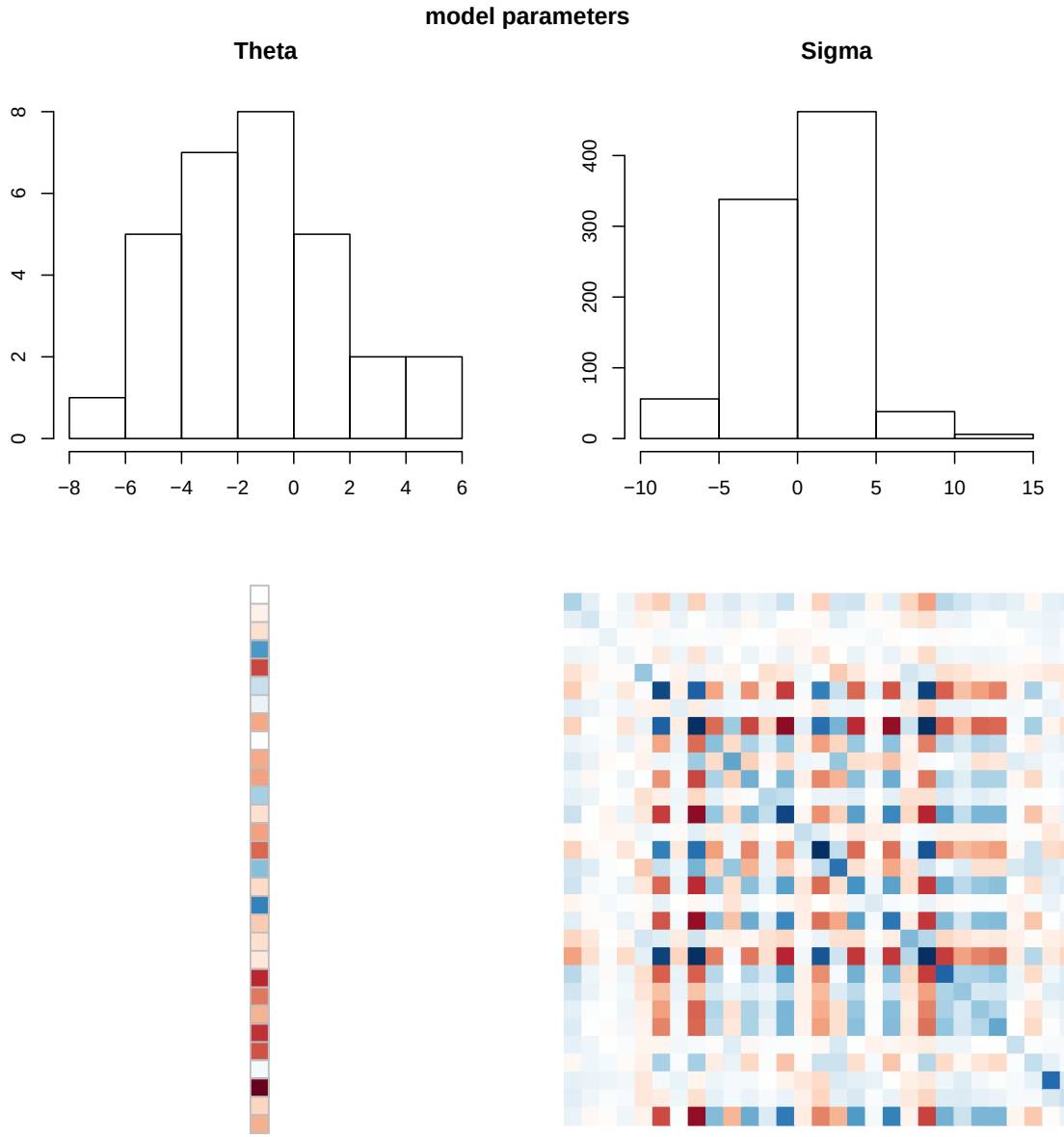
Model:

$$Z_i \sim \mathcal{N}(0, \Sigma)$$

$$Y_{ij} \sim \mathcal{P}(e^{\beta_{0j} + Z_{ij}})$$

Interpretation:

- β_{0j}:
- Σ_{jk}:



Barents: all covariates

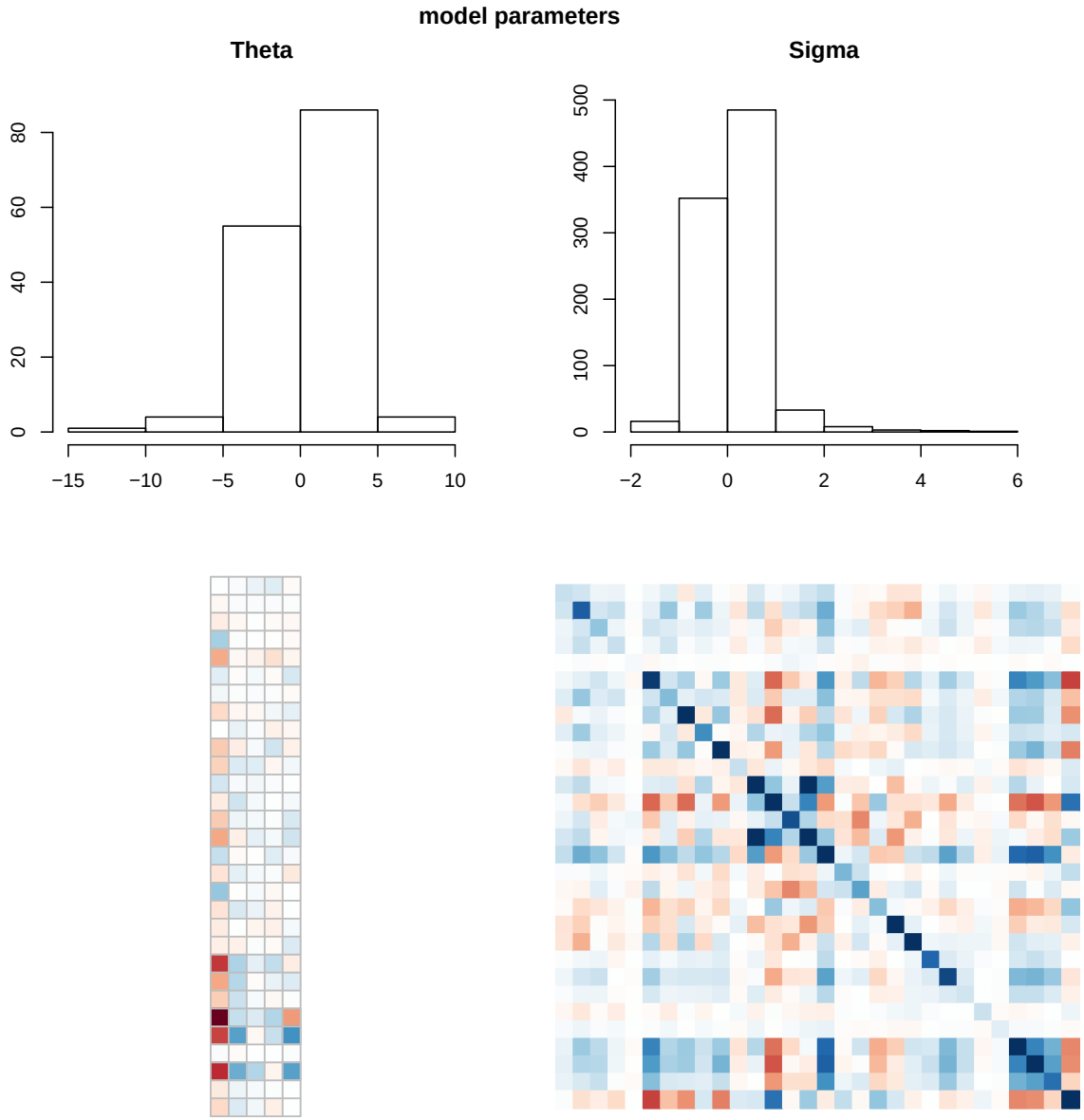
Model:

$$Z_i \sim \mathcal{N}(0, \Sigma)$$

$$Y_{ij} \sim \mathcal{P}(e^{x_i^\intercal \beta_j + Z_{ij}})$$

Interpretation:

- β_{hj}:
- Σ_{jk}:



Comparing samples: Linear discriminant analysis (LDA)

Question. Does the treatment affect the species abundance distribution?

Data. Collect samples under different treatment levels (covariates + abundances): (X^A, Y^A), (X^B, Y^B), (X^C, Y^C), ...

A possible model. For species j in sample i collected under treatment t:

$$Z_i^t \sim \mathcal{N}_p(0, \Sigma)$$
$$Y_{ij}^t \sim \mathcal{P}(\exp(o_{ij}^t + \beta_{0j}^t + \beta_{1j} x_{1i}^t + Z_{ij}^t))$$

PLN syntax:

```
PLNLDA(abundance ~ covariates + offset(o), grouping=treatment)
```

Model assumptions.

- Different mean species abundance under each treatment: β_{0j}^t
- Same effect of the covariates under all treatments: β_{1j}
- Same species dependency under all treatments: Σ

Dimension reduction: Principle component analysis (PCA)

Questions.

- Can we visualize the data in few (2, 3) dimensions?
- Can we determine main trends in the species covariations?

'Probabilistic' PCA. The Z_i actually lay in $q \ll p$ dimensions

$$\Sigma = BB^\intercal, \qquad B : p \times q$$

i.e. draw $W_i \sim \mathcal{N}_q(0, I)$, then $Z_i = BW_i$

PLN-PCA. PLN model, where Σ is forced to have rank q

Modelling the dependencies: Network inference

Question. Which species are in direct interaction?

Gaussian property: Suppose $Z \sim \mathcal{N}_p(0, \Sigma)$ and denote $\Omega = \Sigma^{-1}$:

$$\begin{aligned} & \{\Omega_{jk} = 0\} \\ \Leftrightarrow \quad & \{(Z_j, Z_k) \text{ independent conditionally on all other coordinates}\} \end{aligned}$$

PLN-network. PLN model, where Ω is forced to be 'sparse'

About estimation

Fitting PLN is not easy mostly because the latent Z_i's are not observed.

EM-like strategy: iteratively

- E-step: 'retrieve' Z
- M-step: update the parameter estimates $\widehat{\theta}$

Specificity of PLN. Intractable E-step (big nasty integral)

- Resort to a so-called variational approximation
- Price to pay:
  $\rightarrow$ only approximate standard deviations of the estimates
  $\rightarrow$ no formal test
- Ongoing works

Outline

Interaction data

Questions.

- Do individuals interact 'uniformly' (or 'randomly') with each other?
- Do some individuals play similar role in the interaction network?
- Do covariates (partially) explain the organization of the network?

Data.

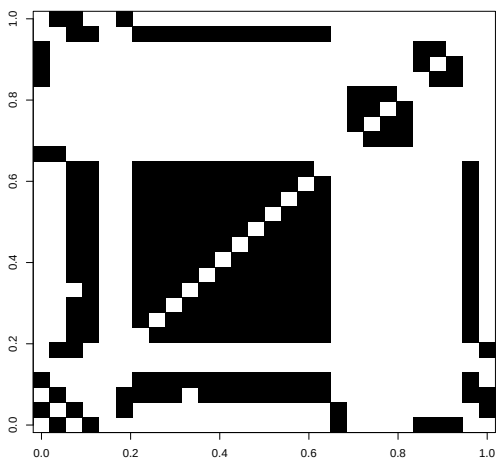
- p individuals
- $Y_{ij} =$ interaction between individuals i and j
- $x_{ij} =$ vector of d covariates for the individuals pair (i, j)

Zebra network

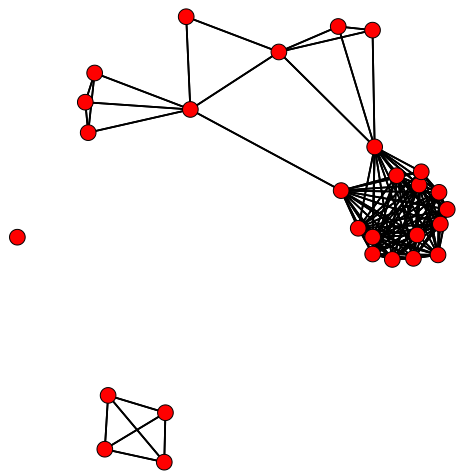
- $p = 26$ zebras
- $Y_{ij} = 1$ if i and j socially interact

$Y = n \times n$ adjacency matrix

```
##       [,1] [,2] [,3] [,4] [,5]
## [1,]    0    1    0    1    0
## [2,]    1    0    1    0    0
## [3,]    0    1    0    0    0
## [4,]    1    0    0    0    0
## [5,]    0    0    0    0    0
## [6,]    0    1    1    0    0
```



Network



Stochastic blockmodel (SBM)

Rational.

- Individuals are spread into K groups
- Interactions between individuals are ruled by their respective group membership

Model.

- for each individuals i, draw a random (independent, latent) group

$$P(Z_i = k) = \pi_k$$

- for each pair of individuals (i, j), draw the (independent) interaction

$$\text{if } Z_i = k, Z_j = \ell : \qquad P(Y_{ij} = 1) = \alpha_{k\ell}$$

Model parameters. $\theta = (\pi, \alpha)$

- $\pi = [\pi_k]$: group proportions
- $\alpha = [\alpha_{k\ell}]$: interaction probability between groups

Zebra (binary) network

$\widehat{\pi}$

```
## [1] 0.465 0.535
```

$\widehat{\alpha}$

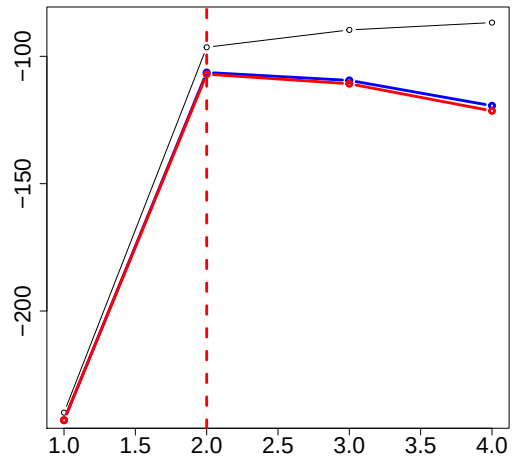
```
##       [,1]  [,2]
## [1,] 0.229 0.025
## [2,] 0.025 0.975
```

Mean degree

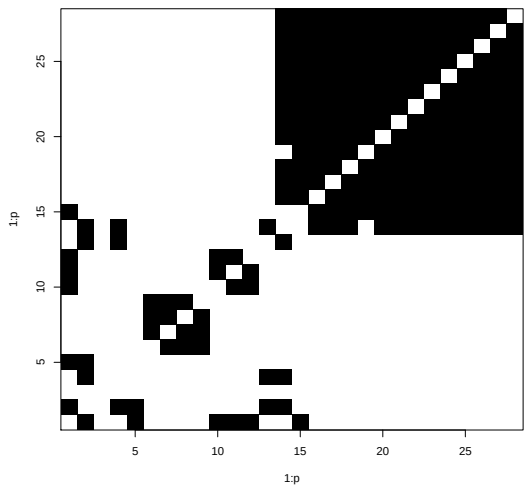
```
## [1]  3.23 14.40
```

$d_k = (p-1)\sum_\ell \pi_\ell \alpha_{k\ell}$

Model selection



Node clustering



Zebra dataset

Data. Actually:

- $p = 26$ zebras
- $Y_{ij} =$ number of interactions btw i and j
- $d = 2$ *node* covariates (sex, age)

A model?

- for each individuals i, draw a random (independent, latent) group

$$P(Z_i = k) = \pi_k$$

- for each pair of individuals (i, j), draw the (independent) interaction

$$\text{if } Z_i = k, Z_j = \ell : \qquad Y_{ij} \sim \mathcal{P}(e^{\alpha_{kl} + x_{ij}^\intercal \beta})$$

$x_{ij} =$ vector of *edge* covariate

Valued SBM

Encoding edge covariates. A possibility:

- $x_{1,ij} = 1$ if i and j have same sex, 0 otherwise
- $x_{2,ij} = 1$ if i and j have same age, 0 otherwise
- $x_{3,ij} = 1$ if i and j have both same sex and same age, 0 otherwise

Model parameters. $\theta = (\pi, \alpha, \beta)$

- π_k : proportion of group k
- $\alpha_{k\ell}$: interaction between groups k and ℓ
- $\beta_1, \beta_2, \beta_3$: effect of the sex, age, both

$+Z_i$: group of node i

Valued zebra network

$\widehat{\pi}$

```
## [1] 0.426 0.074 0.250 0.250
```

$\widehat{\beta}$

```
## [1]  0.881 -0.058  0.082
```

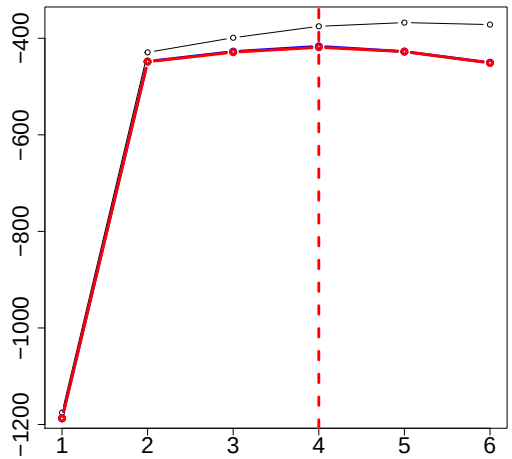
$\widehat{\alpha}$

```
##         [,1]   [,2]   [,3]   [,4]
## [1,] -1.649 -1.925 -3.411 -3.694
## [2,] -1.925  1.807 -0.558 -0.975
## [3,] -3.411 -0.558  1.255  1.183
## [4,] -3.694 -0.975  1.183  1.925
```

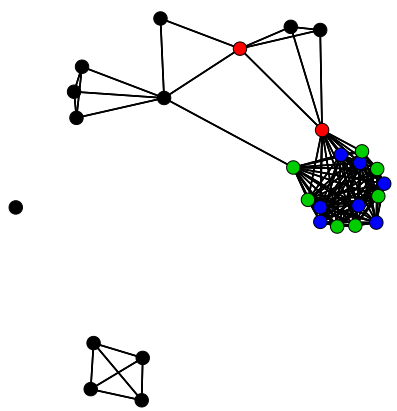
Prediction. i in group 1, j in group 2, (i, j) same sex but not same age. Expected number contacts Y_{ij}:

```
exp(alpha[1, 2] + beta[1])

## [1] 0.3520556
```

Model selection



Node clustering



Valued zebra data: no cluster ($K = 1$)

$\widehat{\alpha}$

```
## [1] -0.47
```

$\widehat{\beta}$

```
## [1]  1.948 -0.274 -0.042
```

Prediction. i and j have same sex but not same age, expected number contacts Y_{ij}:

```
## [1] 4.385
```

A simple equivalent model?

- Poisson regression (GLM)

```
## (Intercept)       X.mat1       X.mat2       X.mat3
##      -0.470        1.948       -0.274       -0.041
```

Latent blockmodel (LBM)

Question.

- Effect of the plant genotype on the microbial diversity of its rhizosphere.
- Preferential relations between groups of genotype and microbes (OTUs)

Data.

- n OTUs ($i = 1..n$), m genotypes ($j = 1..m$)
- Y_{ij} = abundance of OTU i for genotype j
- o_j = sampling effort for genotype j

A model.

$$Z_i^O = \text{ cluster of OTU } i : \qquad P(Z_i^O = k) = \pi_k^O$$

$$Z_j^G = \text{ cluster of genotype } j : \qquad P(Z_j^G = \ell) = \pi_\ell^G$$

$$\text{if } Z_i^O = k \text{ and } Z_j^G = g : \qquad Y_{ij} \sim \mathcal{P}(e^{o_{ij} + \alpha_{k\ell}})$$

$\theta = (\pi^O, \pi^G, \alpha)$

Other extensions

Multiplex networks.

- Different types of interactions

Multilayer networks.

- Different types of nodes

Dynamic networks.

- Interactions occurring along time

About estimation

Fitting SBM is not easy mostly because the latent Z_i's are not observed.

EM-like strategy: iteratively

- E-step: 'retrieve' Z
- M-step: update the parameter estimates $\widehat{\theta}$

Specificity of SBM. Intractable E-step (too many ways to assign nodes to groups: big nasty sum)

- ... same story as PLN
- ... ongoing works