A partial history of latent variable models in genomics

S. Robin + many others

LPSM, Sorbonne université

Statistical Methods for Post Genomic Data, 2024

Introduction

Latent variable model. Model involving unobserved (= 'latent' = 'hidden' = . . .) variables.

Introduction

Latent variable model. Model involving unobserved (= 'latent' = 'hidden' = ...) variables.

Use for modelling. Account for

- missing information (e.g. classification)
- dependency structure
- over-dispersion, ...

Introduction

Latent variable model. Model involving unobserved (= 'latent' = 'hidden' = ...) variables.

Use for modelling. Account for

- missing information (e.g. classification)
- dependency structure
- over-dispersion, ...

Notations.

Y = observed variables (data), θ = model parameters, Z = latent variables

Introduction

Latent variable model. Model involving unobserved (= 'latent' = 'hidden' = ...) variables.

Use for modelling. Account for

- missing information (e.g. classification)
- dependency structure
- over-dispersion, ...

Notations.

Y = observed variables (data), θ = model parameters, Z = latent variables

Parameters $\neq$ latent variables.

- θ can be fixed or random (e.g. frequentist vs Bayesian)
- Z is random, $\simeq$ same dimension as Y

Inference

Intractable likelihood.

$$\log p_\theta(Y) = \log \int p_\theta(Y, z)\, \mathrm{d}z$$

Inference

Intractable likelihood.

$$\log p_\theta(Y) = \log \int p_\theta(Y, z)\, \mathrm{d}z$$

Popular approach. Use the decomposition [DLR77]

$$\log p_\theta(Y) = \mathbb{E}_\theta[\log p_\theta(Y, Z) \mid Y] + \underbrace{\mathcal{H}_\theta(Z \mid Y)}_{\text{entropy}}$$

Inference

Intractable likelihood.

$$\log p_\theta(Y) = \log \int p_\theta(Y, z)\, \mathrm{d}z$$

Popular approach. Use the decomposition [DLR77]

$$\log p_\theta(Y) = \mathbb{E}_\theta[\log p_\theta(Y, Z) \mid Y] + \underbrace{\mathcal{H}_\theta(Z \mid Y)}_{\text{entropy}}$$

and apply the iterative expectation-maximization (EM) algorithm:

$$\theta^{(h+1)} = \underbrace{\arg\max_\theta}_{\text{M step}} \underbrace{\mathbb{E}_{\theta^{(h)}}}_{\text{E step}} [\log p_\theta(Y, Z) \mid Y]$$

to (hopefully) get the MLE $\widehat{\theta} = \arg\max_\theta \log p_\theta(Y)$.

Inference

Intractable likelihood.

$$\log p_\theta(Y) = \log \int p_\theta(Y, z)\, \mathrm{d}z$$

Popular approach. Use the decomposition [DLR77]

$$\log p_\theta(Y) = \mathbb{E}_\theta[\log p_\theta(Y, Z) \mid Y] + \underbrace{\mathcal{H}_\theta(Z \mid Y)}_{\text{entropy}}$$

and apply the iterative expectation-maximization (EM) algorithm:

$$\theta^{(h+1)} = \underbrace{\arg\max_\theta}_{\text{M step}} \underbrace{\mathbb{E}_{\theta^{(h)}}}_{\text{E step}} [\log p_\theta(Y, Z) \mid Y]$$

to (hopefully) get the MLE $\widehat{\theta} = \arg\max_\theta \log p_\theta(Y)$.

Critical step = E step. Given the current $\theta^{(h)}$, compute some moments of the conditional distribution $p_{\theta^{(h)}}(Z \mid Y)$:

$$\mathbb{E}_{\theta^{(h)}}[f(Z) \mid Y].$$

Outline

Differentially expressed genes

Multiple testing. $n \simeq 10^3, 10^4$ genes,

$$H_i = \{\text{gene } i \text{ has the same expression level under conditions } A \text{ and } B\},$$

$P_i = p$-value for gene i ($P_i \sim \mathcal{U}(0,1)$ if H_i holds).

Differentially expressed genes

Multiple testing. $n \simeq 10^3, 10^4$ genes,

$$H_i = \{\text{gene } i \text{ has the same expression level under conditions } A \text{ and } B\},$$

$P_i = p$-value for gene i ($P_i \sim \mathcal{U}(0,1)$ if H_i holds).

Aim.

- Detect which H_i should be rejected
- while avoiding to many false rejections

Differentially expressed genes

Multiple testing. $n \simeq 10^3, 10^4$ genes,

$$H_i = \{\text{gene } i \text{ has the same expression level under conditions } A \text{ and } B\},$$

$P_i = p$-value for gene i ($P_i \sim \mathcal{U}(0,1)$ if H_i holds).

Aim.

- Detect which H_i should be rejected
- while avoiding to many false rejections

Mixture model [MBBTJ06]. 2 component mixture

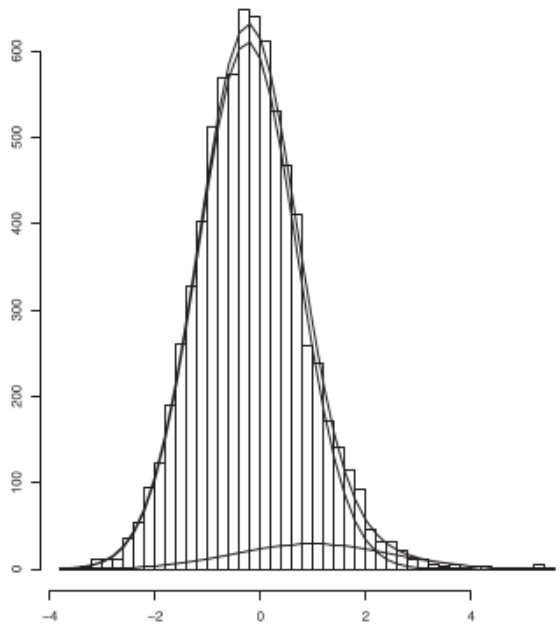
- $Z_i = 0$ if H_i holds, 0 otherwise

$$\pi = \Pr\{Z_i = 1\}$$

- $Y_i = -\Phi^{-1}(P_i)$ (Φ^{-1} = probit):

$$Y_i \mid Z_i = 0 \sim \mathcal{N}(0,1), \quad Y_i \mid Z_i = 0 \sim \mathcal{N}(\mu, \sigma^2)$$

- $\theta = (\pi, \mu, \sigma^2)$



Mixture model

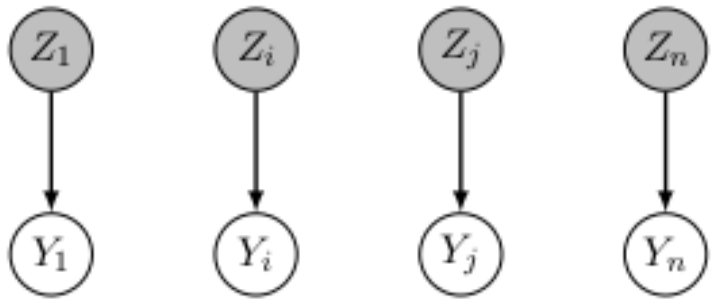
Easy EM. Independent couples (Z_i, Y_i)

$$p_\theta(Z_i \mid Y) = p_\theta(Z_i \mid Y_i)$$

→ Bayes formula:

$$\widehat{\tau}_i = \mathbb{P}_{\widehat{\theta}}\{Z_i = 1 \mid Y_i\}.$$

Graphical model:



Mixture model

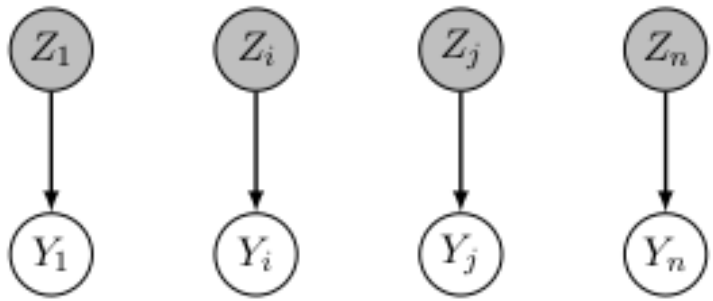
Easy EM. Independent couples (Z_i, Y_i)

$$p_\theta(Z_i \mid Y) = p_\theta(Z_i \mid Y_i)$$

→ Bayes formula:

$$\widehat{\tau}_i = \mathbb{P}_{\widehat{\theta}}\{Z_i = 1 \mid Y_i\}.$$

Graphical model:



False discovery rate = fraction of false positives among positives:

$$\widehat{FDR}(t) = \sum_{i:\widehat{\tau}_i > t} (1 - \widehat{\tau}_i) \Big/ \#\{i : \widehat{\tau}_i > t\}.$$

- Choose the detection threshold t so that $\widehat{FDR}(t) \leqslant 5\%$ (say).

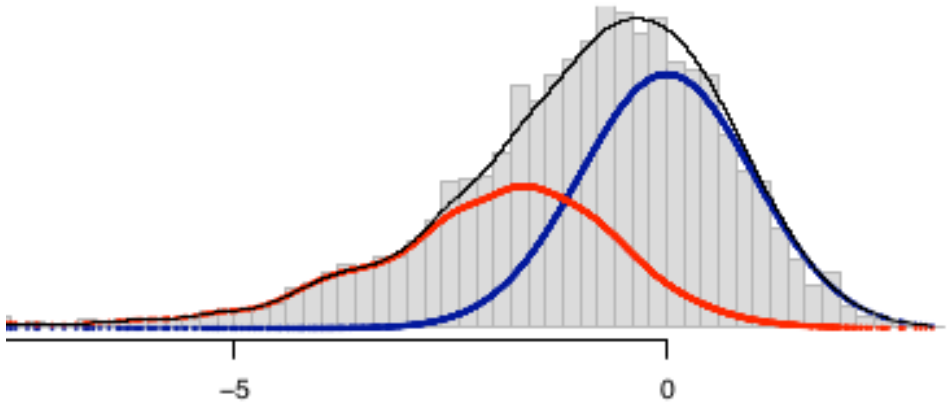
Semi-parametric mixture model

Non-parametric emission distribution.

- Available prior estimate $\widehat{\pi}$ [Sto02]
- Known null distribution

$$p_\theta(Y_i \mid Z_i = 0) = \mathcal{N}(Y_i; 0, 1)$$

- Free alternative distribution



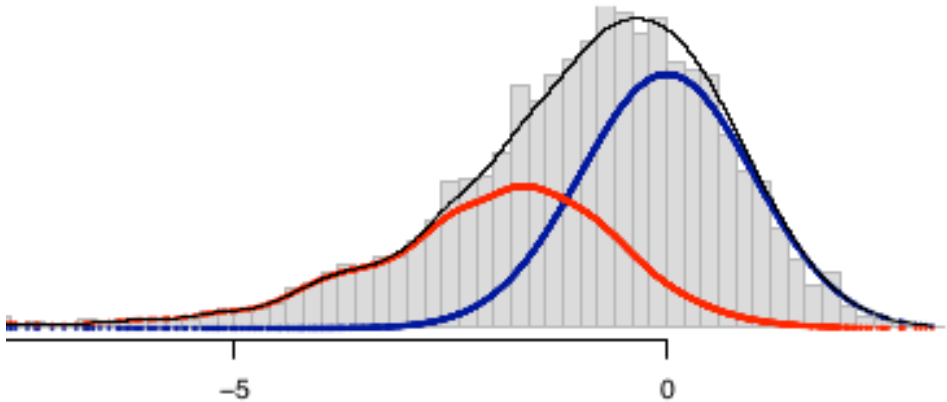
Semi-parametric mixture model

Non-parametric emission distribution.

- Available prior estimate $\widehat{\pi}$ [Sto02]
- Known null distribution

$$p_\theta(Y_i \mid Z_i = 0) = \mathcal{N}(Y_i; 0, 1)$$

- Free alternative distribution



Kernel density estimate of $f_1 = p_\theta(Y_i \mid Z_i = 1)$:

$$\widehat{f}_1(y) = \sum_i \widehat{\tau}_i k(y - y_i) \Big/ \sum_i \widehat{\tau}_i.$$

Composed hypotheses

More complex latent distribution. n genes, Q comparison tests:

$$H_i^q = \{\text{gene } i \text{ is not differentially expressed in comparison } q\}$$

$\rightarrow$ Latent variable: $Z_i = (Z_i^1, \dots Z_i^Q) \in \{0,1\}^Q$, $\Rightarrow 2^Q$ possible configurations.

Composed hypotheses

More complex latent distribution. n genes, Q comparison tests:

$$H_i^q = \{\text{gene } i \text{ is not differentially expressed in comparison } q\}$$

$\rightarrow$ Latent variable: $Z_i = (Z_i^1, \dots Z_i^Q) \in \{0,1\}^Q$, $\Rightarrow 2^Q$ possible configurations.

Mixture model.

- Z_i = gene configuration:

$$p_\theta(Z) = p_\theta(Z_i^1, \dots Z_i^Q)$$

- Y_i = probit p-values

$$p_\theta(Y_i \mid Z_i) = \prod_q p_\theta(Y_i^q \mid Z_i^q)$$

Composed hypotheses

More complex latent distribution. n genes, Q comparison tests:

$$H_i^q = \{\text{gene } i \text{ is not differentially expressed in comparison } q\}$$

$\rightarrow$ Latent variable: $Z_i = (Z_i^1, \dots Z_i^Q) \in \{0,1\}^Q$, $\Rightarrow 2^Q$ possible configurations.

Mixture model.

- Z_i = gene configuration:

$$p_\theta(Z) = p_\theta(Z_i^1, \dots Z_i^Q)$$

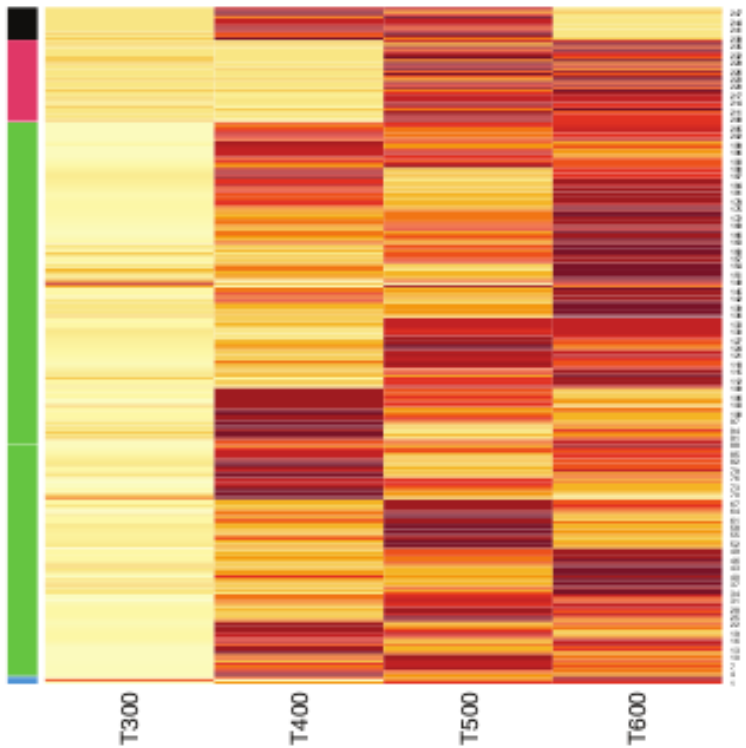
- Y_i = probit p-values

$$p_\theta(Y_i \mid Z_i) = \prod_q p_\theta(Y_i^q \mid Z_i^q)$$

- Right: 5 time points.

$$H^q = \{\mathbb{E}Y(t_{q-1}) = \mathbb{E}Y(t_q)\}.$$

Genes with at least 2 successive differences.



$Z = (0110), (0011), (0111), (1111)$

And also

Avatars of the Poisson distribution.

- Zero inflated Poisson = Mixture Dirac(0) + Poisson
  $\rightarrow$ Regular mixture

- Over-dispersed Poisson = Negative binomial = Poisson-Gamma
  $\rightarrow$ Close form conditional distribution, thanks to conjugacy

And also

Avatars of the Poisson distribution.

- Zero inflated Poisson = Mixture Dirac(0) + Poisson
  $\rightarrow$ Regular mixture

- Over-dispersed Poisson = Negative binomial = Poisson-Gamma
  $\rightarrow$ Close form conditional distribution, thanks to conjugacy

Gaussian mixed models.

- Z = Gaussian random effect
  $\rightarrow$ Close form conditional distribution

Outline

Copy number variation

Data. n locus along the genome,

$$Y_t = \text{noisy measurement of the number of copies at locus } t$$

(should be two for diploids).

Copy number variation

Data. n locus along the genome,

$$Y_t = \text{noisy measurement of the number of copies at locus } t$$

(should be two for diploids).

Asumption. Neighbor loci often share the same number of copies.

Copy number variation

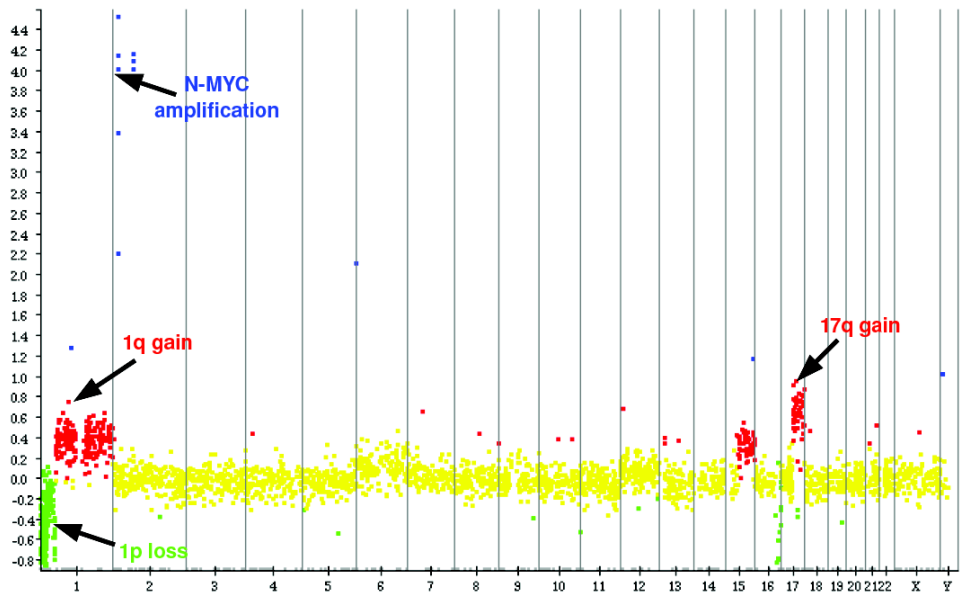
Data. n locus along the genome,

$$Y_t = \text{noisy measurement of the number of copies at locus } t$$

(should be two for diploids).

Asumption. Neighbor loci often share the same number of copies.

Example. [Hup08]



Hidden Markov model

Latent variable model.

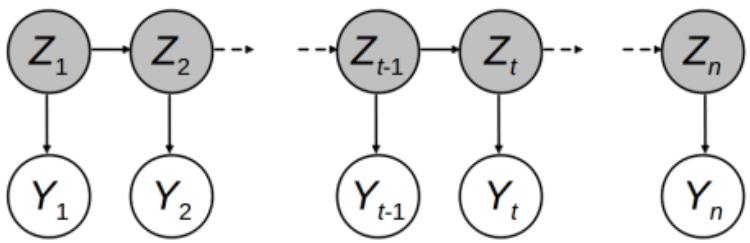
- Z_t = status at locus t:

$$Z = (Z_t)_{1 \leqslant t \leqslant n} \sim MC(\nu, \pi), \qquad (\nu = \text{initial dist}, \pi = \text{transition matrix})$$

- $(Y_t)_{1 \leqslant t \leqslant n}$ independent $\mid Z$:

$$(Y_t \mid Z_t = k) \sim \mathcal{N}(\mu_k, \sigma^2).$$

- $\theta = (\nu, \pi, \mu, \sigma^2)$
- Graphical model:



Hidden Markov model

Latent variable model.

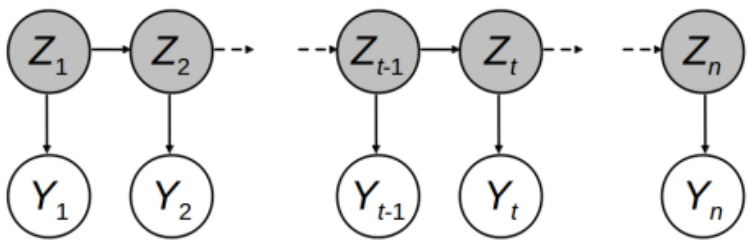
- Z_t = status at locus t:

$$Z = (Z_t)_{1 \leqslant t \leqslant n} \sim MC(\nu, \pi), \qquad (\nu = \text{initial dist}, \pi = \text{transition matrix})$$

- $(Y_t)_{1 \leqslant t \leqslant n}$ independent $\mid Z$:

$$(Y_t \mid Z_t = k) \sim \mathcal{N}(\mu_k, \sigma^2).$$

- $\theta = (\nu, \pi, \mu, \sigma^2)$
- Graphical model:



Inference. E-step = forward-backward (= Baum-Welsh = Kalman = . . .) recursion.

Hidden Markov model

Latent variable model.

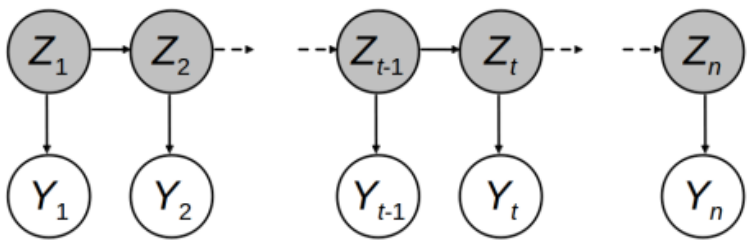
- Z_t = status at locus t:

$$Z = (Z_t)_{1 \leqslant t \leqslant n} \sim MC(\nu, \pi), \qquad (\nu = \text{initial dist}, \pi = \text{transition matrix})$$

- $(Y_t)_{1 \leqslant t \leqslant n}$ independent $\mid Z$:

$$(Y_t \mid Z_t = k) \sim \mathcal{N}(\mu_k, \sigma^2).$$

- $\theta = (\nu, \pi, \mu, \sigma^2)$
- Graphical model:



Inference. E-step = forward-backward (= Baum-Welsh = Kalman = . . .) recursion.

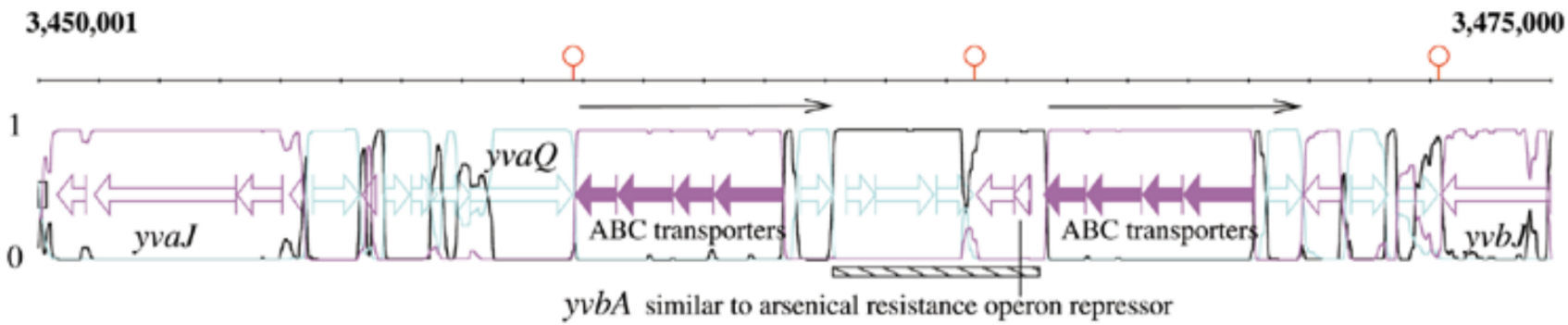
Classification. Most probable hidden path $\hat{Z}$ = Viterbi (dynamic programming) algorithm.

Gene detection

Aim starting from $Y =$

```
ATCTTTTTCGGCTTTTTTTAGTATCCACAGAGGTTATCGACAACATTTTCACATTACCAACCCCTGTGGACAAGGTTTTT
TCAACAGGTTGTCCGCTTTGTGGATAAGATTGTGACAACCATTGCAAGCTCTCGTTTATTTTGGTATTATATTTGTGTTT
TAACTCTTGATTACTAATCCTACCTTTCCTCTTTATCCACAAAGTGTGGATAAGTTGTGGATTGATTTCACACAGCTTGT
GTAGAAGGTTGTCCACAAGTTGTGAAATTTGTCGAAAAGCTATTTATCTACTATATTATATGTTTTCAACATTTAATGTG
TACGAATGGTAAGCGCCATTTGCTCTTTTTTTGTGTTCTATAACAGAGAAAGACGCCATTTTCTAAGAAAAGGAGGGACG
...
```

get



Modelling the Markov structure

Gene detection. [NBM+02]

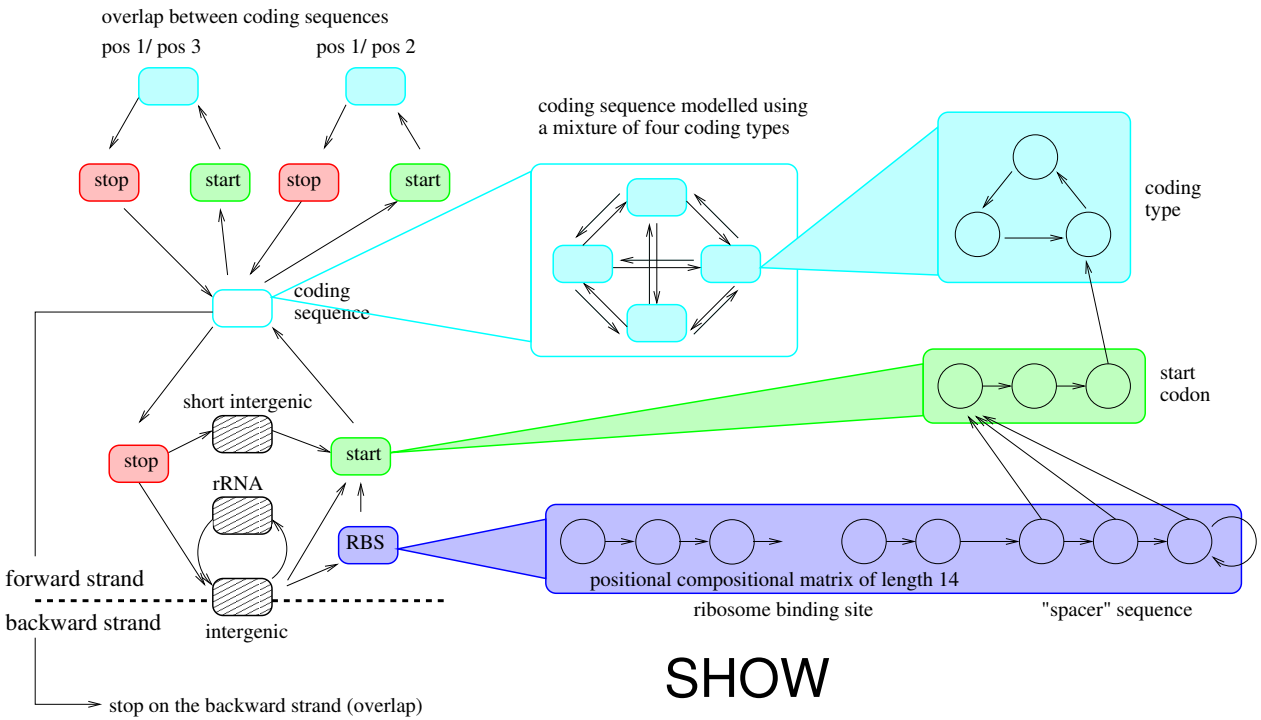
- $Y = (Y_t)_{1 \leqslant t \leqslant n}$ = (bacterial) genome ($n \sim 10^6$)
- Z_t = coding status of nucleotide t

Modelling the Markov structure

Gene detection. [NBM+02]

- $Y = (Y_t)_{1 \leqslant t \leqslant n}$ = (bacterial) genome ($n \sim 10^6$)
- Z_t = coding status of nucleotide t

Transition graph of (Z_t).



Not modelling the emission distribution

Non-parametric HMM. Application of [AMR09] to HMM

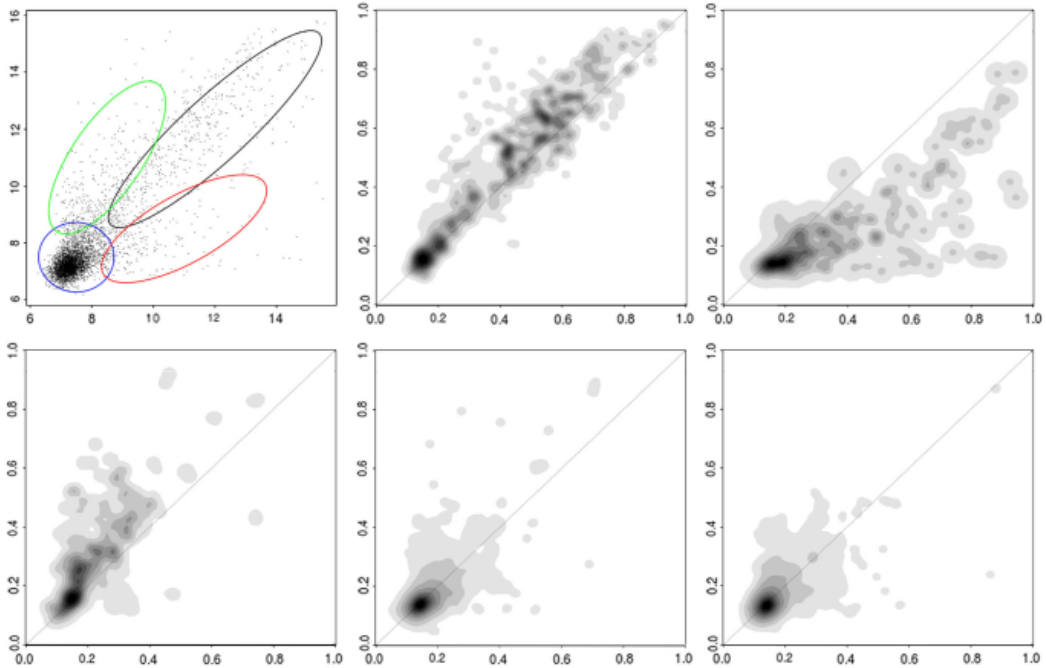
- HMM with non-parametric emission distributions are generically identifiable.
- No distribution assumptions (Gaussian, Poisson, Negative binomial, etc).

Not modelling the emission distribution

Non-parametric HMM. Application of [AMR09] to HMM

- HMM with non-parametric emission distributions are generically identifiable.
- No distribution assumptions (Gaussian, Poisson, Negative binomial, etc).

Differentially expressed regions along the genome. Kernel-density estimates in each of the $K = 5$ states



Tree-shaped Markov models

Remark. The graphical model of an HMM is tree-shaped, likewise this of many evolutionary models.

[1] Main problem in phylogeny = find the tree

[2] Breaks down for a network (e.g. horizontal transfert)

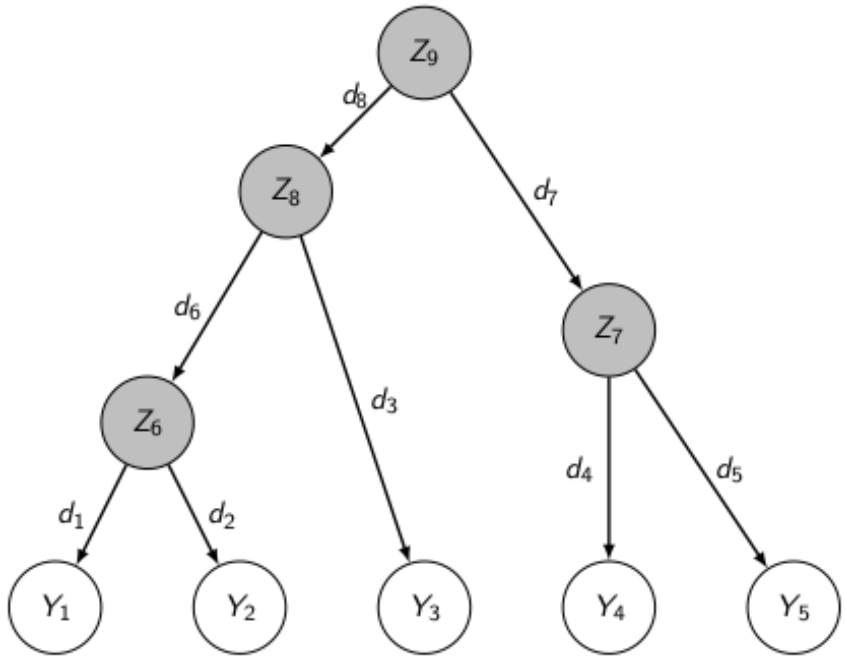
Tree-shaped Markov models

Remark. The graphical model of an HMM is tree-shaped, likewise this of many evolutionary models.

Phylogeny.

- Y = genomes of extant species
- Z = ancestral genomes

Likelihood for a given[1] tree[2].
Felsentein's algorithm [Fel81]
= 'upward-downward' recursion



[1] Main problem in phylogeny = find the tree

[2] Breaks down for a network (e.g. horizontal transfert)

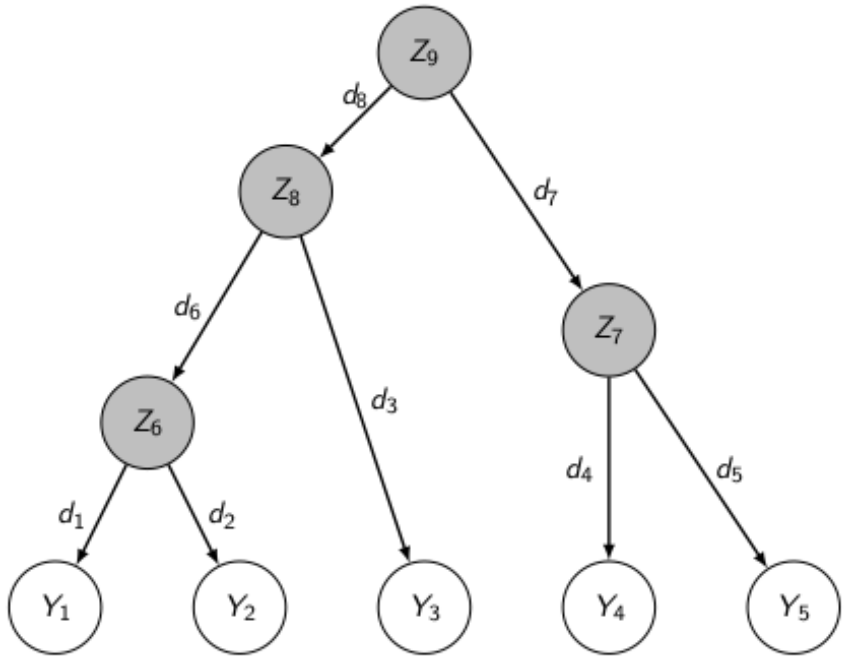
Tree-shaped Markov models

Remark. The graphical model of an HMM is tree-shaped, likewise this of many evolutionary models.

Phylogeny.

- Y = genomes of extant species
- Z = ancestral genomes

Likelihood for a given[1] tree[2].
Felsentein's algorithm [Fel81]
= 'upward-downward' recursion



Ancestral trait reconstruction. Same problem and same solution for continuous (e.g. Brownian) latent Z [Lar14].

[1] Main problem in phylogeny = find the tree

[2] Breaks down for a network (e.g. horizontal transfert)

Other complex latent structures

More complex latent structures may remain manageable using relevant algebraic properties.

- Sequence partitioning (segmentation): dynamic programming [AL89] and/or simple matrix products
  → quadratic complexity

- Spanning tree[3]-shaped graphical models: maximum spanning tree [Kru56] and/or simple determinant calculation [Cha82]
  → cubic complexity

. . . no generic rule to identify manageable E-steps.

[3] Phylogenetic trees are not spanning trees

Outline

Interaction networks

Data. $Y = n \times n$ interaction matrix:

$$Y_{ij} = 1 \text{ if entities } i \text{ and } j \text{ interact}, \quad = 0 \text{ otherwise}$$

Entities = protein, genes, operons, ...

Aim. Cluster entities according to their interaction profile.

Interaction networks

Data. $Y = n \times n$ interaction matrix:

$$Y_{ij} = 1 \text{ if entities } i \text{ and } j \text{ interact}, \quad = 0 \text{ otherwise}$$

Entities = protein, genes, operons, ...

Aim. Cluster entities according to their interaction profile.

Stochastic block-model.

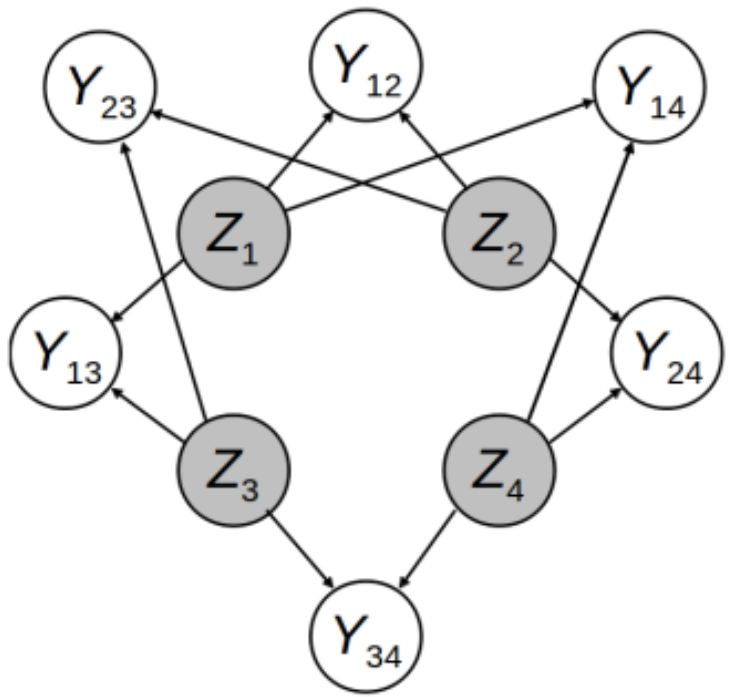
- Z_i cluster membership of entity i

$$\pi_i = \mathbb{P}\{Z_i = k\}$$

- Y_{ij} interaction (i, j)

$$\gamma_{k\ell} = \mathbb{P}\{Y_{ij} = 1 \mid Z_i = k, Z_j = \ell\}$$

(can include covariates, strength, ...).

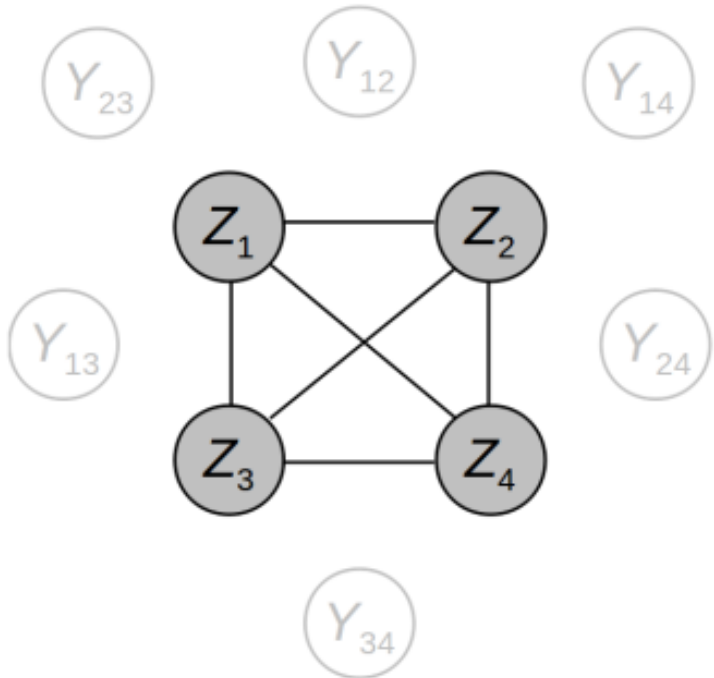


Stochastic block-models

Stochastic block-models

Nasty conditional distribution.

- 'Moralization' makes parents married
- The graphical model of $p_\theta(Z \mid Y)$ is a clique
- No nice factorization can be hoped to integrate it
- The E-step is intractable for, say, $n \geqslant 20$



→ Regular EM does not apply

Metabarcoding & PLN model

Data. n samples, p species,

$$Y_{ij} = \#\text{ of individuals (reads) from species } j \text{ in sample } i$$

Metabarcoding & PLN model

Data. n samples, p species,

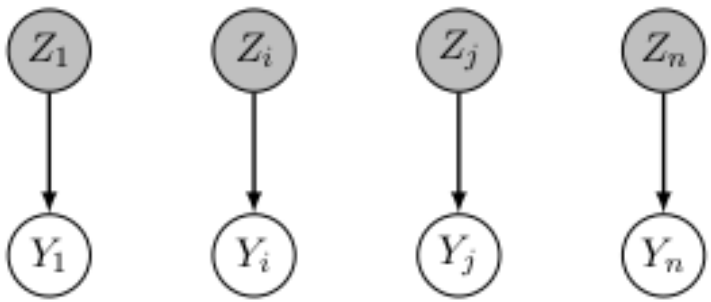
$$Y_{ij} = \#\text{ of individuals (reads) from species } j \text{ in sample } i$$

Poisson log-normal model.

- Independent $(Z_i)_{1 \leqslant i \leqslant n}$: $Z_i \sim \mathcal{N}_p(0, \Sigma)$,
- Conditionally independent counts (Y_{ij}):

$$(Y_{ij} \mid Z_{ij}) \sim \mathcal{P}(\exp(Z_{ij}))$$

(can include covariates, offset, ..)



$Y_i, Z_i \in \mathbb{R}^p$

Metabarcoding & PLN model

Data. n samples, p species,

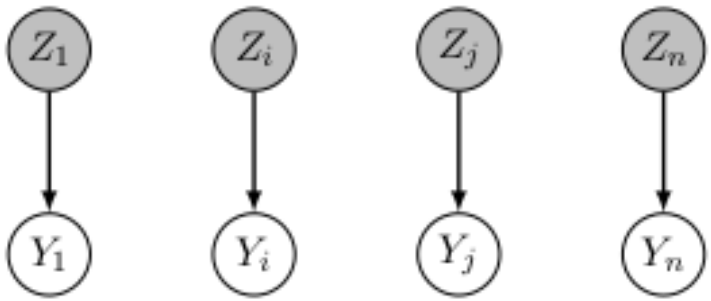
$$Y_{ij} = \#\text{ of individuals (reads) from species } j \text{ in sample } i$$

Poisson log-normal model.

- Independent $(Z_i)_{1 \leqslant i \leqslant n}$: $Z_i \sim \mathcal{N}_p(0, \Sigma)$,
- Conditionally independent counts (Y_{ij}):

$$(Y_{ij} \mid Z_{ij}) \sim \mathcal{P}(\exp(Z_{ij}))$$

(can include covariates, offset, ..)



$Y_i, Z_i \in \mathbb{R}^p$

Nasty conditional distribution. $p_\theta(Z_i \mid Y_i) = p_\theta(Y_i, Z_i)/p_\theta(Y_i)$ but

$$\text{no close form for} \quad p_\theta(Y_i) = \int p_\theta(Y_i, z_i)\, \mathrm{d}z_i$$

- Intractable E step
- Regular EM does not apply

Variational EM

Principle. Replace $p_\theta(Z \mid Y)$ with some approximation $q_\psi(Z)$

- q_ψ chosen within a parametric distribution class $\mathcal{Q}$
- $\psi =$ variational parameter

Variational EM

Principle. Replace $p_\theta(Z \mid Y)$ with some approximation $q_\psi(Z)$

- q_ψ chosen within a parametric distribution class $\mathcal{Q}$
- $\psi =$ variational parameter

Evidence lower bound (ELBO).

$$\begin{aligned} ELBO(\theta, \psi) &= \log p_\theta(Y) - KL(q_\psi(Z) || p_\theta(Z \mid Y)) \\ &= \mathbb{E}_\psi[\log p_\theta(Y, Z)] - \mathcal{H}(q_\psi) \end{aligned}$$

Variational EM

Principle. Replace $p_\theta(Z \mid Y)$ with some approximation $q_\psi(Z)$

- q_ψ chosen within a parametric distribution class $\mathcal{Q}$
- $\psi =$ variational parameter

Evidence lower bound (ELBO).

$$\begin{aligned} ELBO(\theta, \psi) &= \log p_\theta(Y) - KL(q_\psi(Z) || p_\theta(Z \mid Y)) \\ &= \mathbb{E}_\psi[\log p_\theta(Y, Z)] - \mathcal{H}(q_\psi) \end{aligned}$$

Variational EM.

$$\psi^{h+1} = \arg\max_\psi ELBO(\theta^{(h)}, \psi) = \arg\min_\psi KL(q_\psi(Z) || p_{\theta^{(h)}}(Z \mid Y)),$$

$$\theta^{(h+1)} = \arg\max_\theta ELBO(\theta, \psi^{(h+1)}) = \arg\max_\theta \mathbb{E}_{\psi^{(h+1)}}[\log p_\theta(Y, Z)]$$

Variational EM

Application. Critical choice = choice of $\mathcal{Q}$

SBM: $\mathcal{Q}$ = factorable distributions

$$q_\psi(Z) = \prod_i q_{\psi_i}(Z_i)$$

PLN: $\mathcal{Q}$ = Gaussian distributions

$$q_{\psi_i}(Z_i) = \mathcal{N}(Z_i; m_i, S_i)$$

Variational EM

Application. Critical choice = choice of $\mathcal{Q}$

SBM: $\mathcal{Q}$ = factorable distributions

$$q_\psi(Z) = \prod_i q_{\psi_i}(Z_i)$$

PLN: $\mathcal{Q}$ = Gaussian distributions

$$q_{\psi_i}(Z_i) = \mathcal{N}(Z_i; m_i, S_i)$$

What do we know about VEM?

- Computationally efficient
- Works well in practice (estimation, classification, ...)
- Almost no theoretical guaranty
- Provides no measure of uncertainty

→ Need to consider alternatives or post-processing

Sequential Monte-Carlo for SBM

Weighted SBM with covariates.

$$Y_{ij} \sim \mathcal{P}(e^{\alpha_{k\ell} + x_i^\top \beta})$$

VEM provides : an estimate $\widehat{\theta}_{VEM}$ and an approximate $q_\psi(Z)$.
$\rightarrow$ Build an approximate posterior $p_{VEM}(Z, \theta \mid Y)$.

Sequential Monte-Carlo for SBM

Weighted SBM with covariates.

$$Y_{ij} \sim \mathcal{P}(e^{\alpha_{k\ell} + x_i^\top \beta})$$

VEM provides : an estimate $\widehat{\theta}_{VEM}$ and an approximate $q_\psi(Z)$.
$\rightarrow$ Build an approximate posterior $p_{VEM}(Z, \theta \mid Y)$.

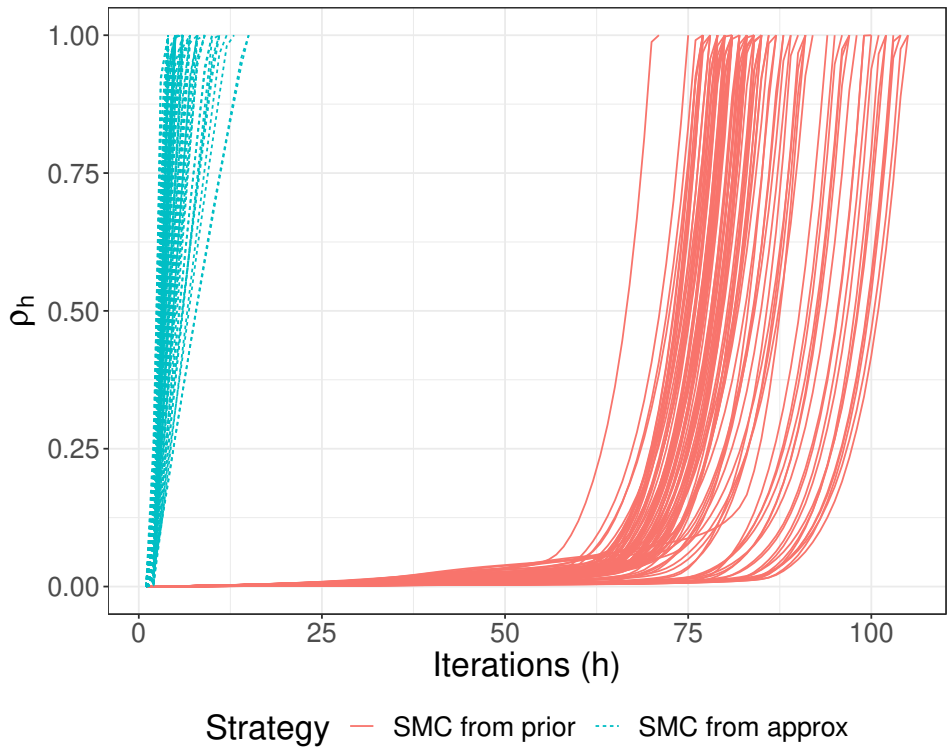
SMC algorithm. Sample from a sequence of distributions

$$p_h(Z, \theta \mid Y) \propto p(Z, \theta \mid Y)^{\rho_h} \; p_0(Z, \theta \mid Y)^{1-\rho_h}$$

- p_0 = starting distribution (p_{VEM}, prior, ...).
- p_H = posterior, when $\rho_H = 1$, .
- $\rho_1, \ldots, \rho_H$ tuned so to keep a sufficient effective sample size (ESS) at each step.

Sequential Monte-Carlo for SBM

Choice of p_0. Number of steps to reach the posterior from p_{VEM} or from the prior

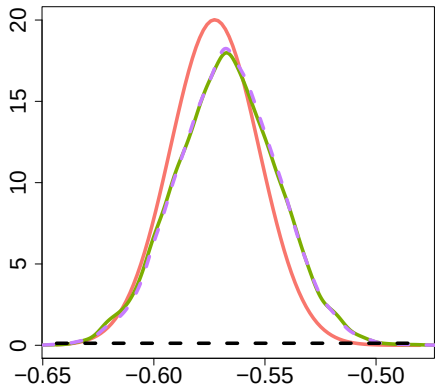


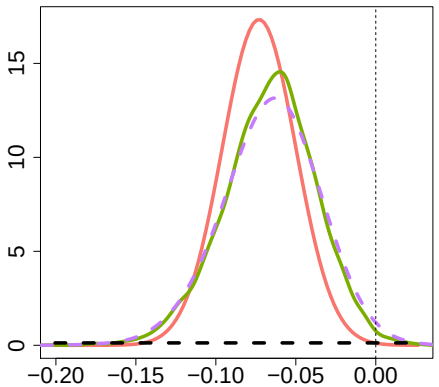
(Synthetic data)

An ecological example ($p = 51$ species)

Posterior distribution for β:

- p_{VEM}
- posterior with $\widehat{K}$
- (posterior with model averaging)

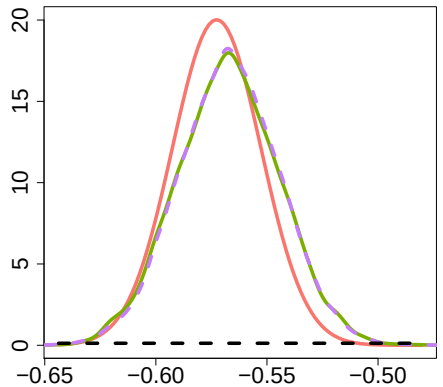


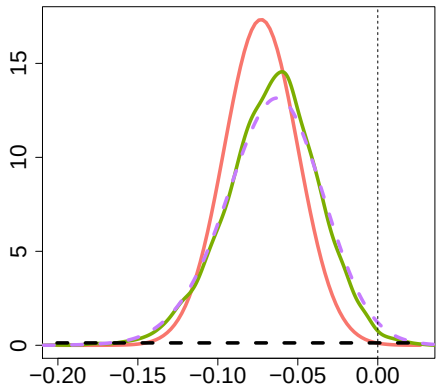


An ecological example ($p = 51$ species)

Posterior distribution for β:

- p_{VEM}
- posterior with $\widehat{K}$
- (posterior with model averaging)

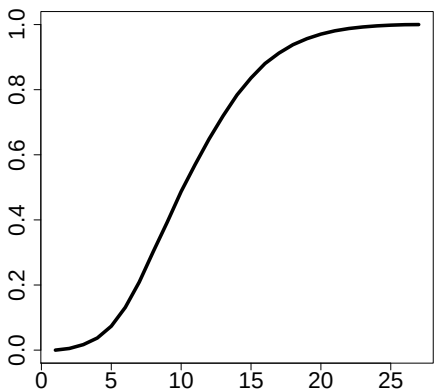




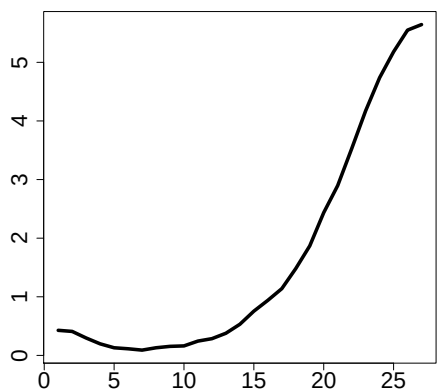
Path to the posterior:

- 25 steps to reach the posterior
- Mostly to recover the dependency structure between the Z_i

ρ_h



$MI(Z)$



Monte-Carlo EM for PLN

Monte-Carlo EM (MCEM). Replace the E step with a sampling step

$$(Z^m)_{1 \leqslant m \leqslant M} \overset{\text{iid}}{\sim} p_\theta(\cdot \mid Y) \qquad \rightarrow \quad \widehat{\mathbb{E}}[f(Z)] = M^{-1} \sum_{m=1}^{M} f(Z^m).$$

Monte-Carlo EM for PLN

Monte-Carlo EM (MCEM). Replace the E step with a sampling step

$$(Z^m)_{1\leqslant m\leqslant M} \text{ iid}\sim p_\theta(\cdot \mid Y) \qquad \rightarrow \quad \widehat{\mathbb{E}}[f(Z)] = M^{-1}\sum_{m=1}^{M} f(Z^m).$$

Importance sampling. When $p_\theta(Z \mid Y)$ not available, sample from a surrogate proposal:

$$(Z^m)_{1\leqslant m\leqslant M} \text{ iid}\sim q(\cdot) \qquad \rightarrow \quad \widehat{\mathbb{E}}[f(Z)] = \sum_{m=1}^{M} w^m f(Z^m),$$

where $w^m \propto p_\theta(Z^m \mid Y)/q(Z^m)$.

Monte-Carlo EM for PLN

Monte-Carlo EM (MCEM). Replace the E step with a sampling step

$$(Z^m)_{1\leqslant m\leqslant M} \text{ iid}\sim p_\theta(\cdot \mid Y) \qquad \rightarrow \quad \widehat{\mathbb{E}}[f(Z)] = M^{-1}\sum_{m=1}^{M} f(Z^m).$$

Importance sampling. When $p_\theta(Z \mid Y)$ not available, sample from a surrogate proposal:

$$(Z^m)_{1\leqslant m\leqslant M} \text{ iid}\sim q(\cdot) \qquad \rightarrow \quad \widehat{\mathbb{E}}[f(Z)] = \sum_{m=1}^{M} w^m f(Z^m),$$

where $w^m \propto p_\theta(Z^m \mid Y)/q(Z^m)$.

Monte-Carlo EM.

- Regular M step to update $\theta^{(h)}$,
- Monte-Carlo E step, using a Gaussian q fitted to estimated the moments of $p_{\theta^{(h)}}(Z \mid Y)$,
- Starting with q_ψ from VEM.

Composite likelihood

Importance sampling has poor efficiency (low ESS) in 'large' dimension ($p \geqslant 5, 10$).

Composite likelihood

Importance sampling has poor efficiency (low ESS) in 'large' dimension ($p \geqslant 5, 10$).

Composite likelihood = linear combination of marginal likelihoods:

- Spread the p species into B overlapping blocks $\mathcal{C}_1, \ldots \mathcal{C}_B$ of size k;
- Define

$$c\ell_\theta(Y) = \sum_{b=1}^{B} \log p_\theta(Y^{\mathcal{C}_b});$$

- Get $\widehat{\theta}_{c\ell} = \arg\max_\theta c\ell_\theta(Y)$ using EM (which still applies);
- $\widehat{\theta}_{c\ell}$ is asymptotically normal, with known asymptotic variance (Godambe information replaces Fisher information).

Composite likelihood

Importance sampling has poor efficiency (low ESS) in 'large' dimension ($p \geqslant 5, 10$).

Composite likelihood = linear combination of marginal likelihoods:

- Spread the p species into B overlapping blocks $\mathcal{C}_1, \ldots \mathcal{C}_B$ of size k;
- Define

$$c\ell_\theta(Y) = \sum_{b=1}^{B} \log p_\theta(Y^{\mathcal{C}_b});$$

- Get $\widehat{\theta}_{c\ell} = \arg\max_\theta c\ell_\theta(Y)$ using EM (which still applies);
- $\widehat{\theta}_{c\ell}$ is asymptotically normal, with known asymptotic variance (Godambe information replaces Fisher information).

MCEM for composite likelihood. Same as before

- replacing the E step with e Monte Carlo step
- where importance sampling is made in dimension $k \ll p$.

Outline

Mixture models

More complex latent structure

Too complex latent structure

Summary

Latent variable model = generic and flexible modelling tool.

Inference mostly depends on the complexity of the conditional distribution $p_\theta(Z \mid Y)$.

Summary

Latent variable model = generic and flexible modelling tool.

Inference mostly depends on the complexity of the conditional distribution $p_\theta(Z \mid Y)$.

Variational approximations provide efficient algorithms to manage too complex hidden structure, but with few statistical guaranties.

Monte Carlo-based methods can take advantage of variational approximations, but at a computational cost.

Some alternatives

Classical MCMC-based techniques often come at a prohibitive computational cost in genomics.

Some alternatives

Classical MCMC-based techniques often come at a prohibitive computational cost in genomics.

Variational auto-encoders

- provide more flexibility to fit the variational parameters ($\rightarrow$ better approximation)
- still relying on a Gaussian assumption

Some alternatives

Classical MCMC-based techniques often come at a prohibitive computational cost in genomics.

Variational auto-encoders

- provide more flexibility to fit the variational parameters ($\rightarrow$ better approximation)
- still relying on a Gaussian assumption

Learning some basic useful transforms.
'Learn' a smooth 1D transform

$$\psi = \psi_{\mu,\sigma^2,y},$$

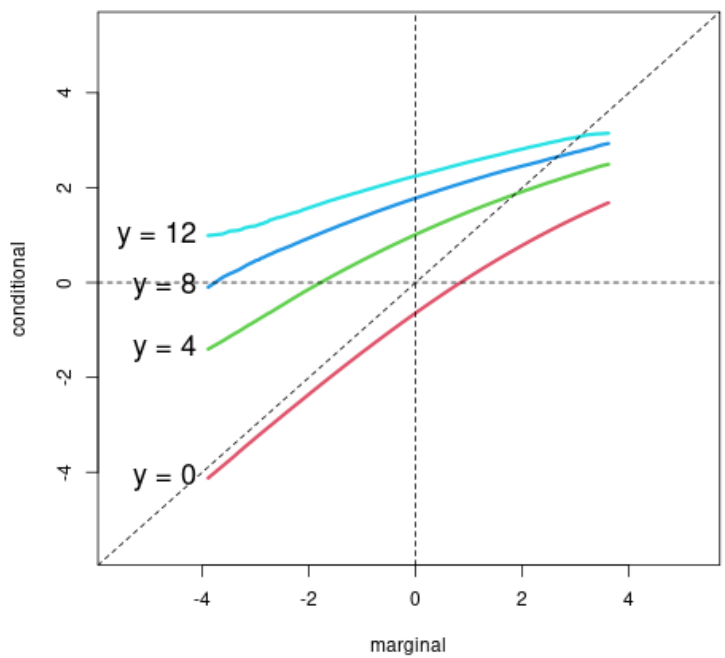
such that, if

$$Z \sim \mathcal{N}(\mu, \sigma^2)$$

then

$$\psi(Z) \sim p_{PLN(\mu,\sigma^2)}(Z \mid Y = y)$$

$\mu = 0, \quad \sigma^2 = 1$



References I

I. E. Auger and C. E. Lawrence. Algorithms for the optimal identification of segment neighborhoods. *Bull. Math. Biol.*, 51(1):39–54, 1989.

E.A. Allman, C. Matias, and J.A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, pages 3099–3132, 2009.

P. Bastide, M. Mariadassou, and S. R. Detection of adaptive shifts on phylogenies by using shifted stochastic processes on a tree. *Journal of the Royal Statistical Society. Series B*, page *to appear*, 2016.

S. Chaiken. A combinatorial proof of the all minors matrix tree theorem. *SIAM Journal on Algebraic Discrete Methods*, 3(3):319–329, 1982.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38, 1977.

S. Donnet and S. R. Accelerating Bayesian estimation for network Poisson models using frequentist variational estimates. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2021.

J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17(6):368–376, 1981.

E. Gassiat, A. Cleynen, and S. R. Inference in finite state space non parametric hidden Markov models and applications. *Statistics and Computing*, 26(1-2):61–71, 2016.

M. Guedj, S. R., A. Célisse, and G. Nuel. Kerfdr: a semi-parametric kernel-based approach to local false discovery rate estimation. *BMC Bioinformatics*, 10(1):84, 2009.

Ph. Hupé. *Biostatistical algorithms for omics data in oncology: Application to DNA copy number microarray experiments*. PhD thesis, AgroParisTech, 2008.

Joseph B. Kruskal. On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. *Proceedings of the American Mathematical Society*, 7(1):48–50, February 1956.

Nicolas Lartillot. A Phylogenetic Kalman Filter for Ancestral Trait Reconstruction Using Molecular Data. *Bioinformatics*, 30(4):488–496, 2014.

References II

G. J. McLachlan, R. W. Bean, and L. Ben-Tovim Jones. A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics*, 22(13):1608–1615, 2006.

P. Nicolas, L. Bize, F. Muri, M. Hoebeke, F. Rodolphe, S. D. Ehrlich, B. Prum, and Ph. Bessières. Mining bacillus subtilis chromosome heterogeneities using hidden markov models. *Nucleic acids research*, 30(6):1418–1426, 2002.

S. R., A. Bar-Hen, J.-J. Daudin, and L. Pierre. A semi-parametric approach for mixture models: Application to local false discovery rate estimation. *Computational statistics & data analysis*, 51(12):5483–5493, 2007.

G. Rigaill, E. Lebarbier, and S. R. Exact posterior distributions over the segmentation space and model selection for multiple change-point detection problem. *Stat. Comp.*, 22:917–29, 2011.

L. Schwaller, S. R., and M. Stumpf. Bayesian inference of graphical model structures using trees. *J. Soc. Franc. Stat.*, 160(2):1–23, 2019.

J. D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

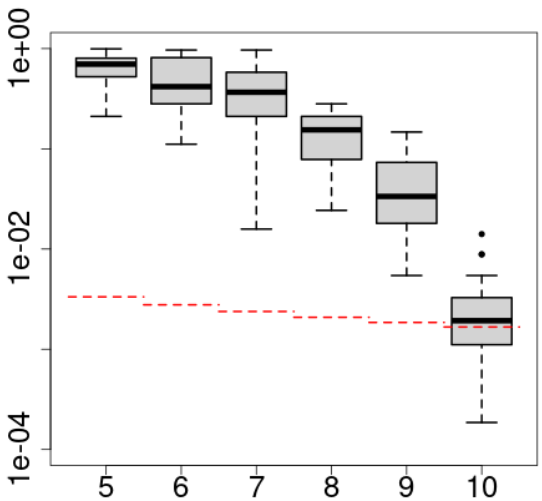
Monte-Carlo EM for PLN

p species, blocks of size k

Log-likelihood $\log p_\theta(Y)$

Synthetic data

Normality

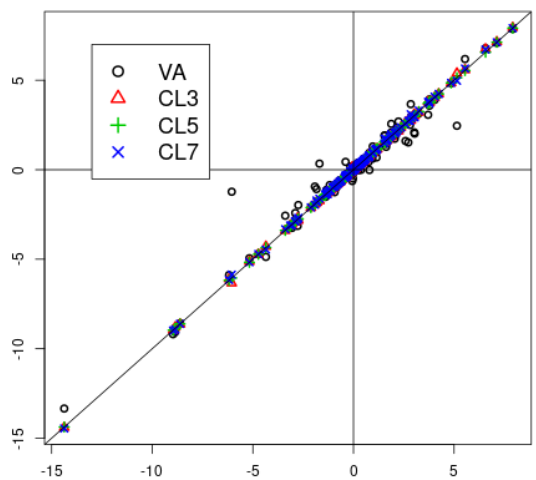


KS p-val = f(p)

Composite log-likelihood $c\ell_\theta(Y)$

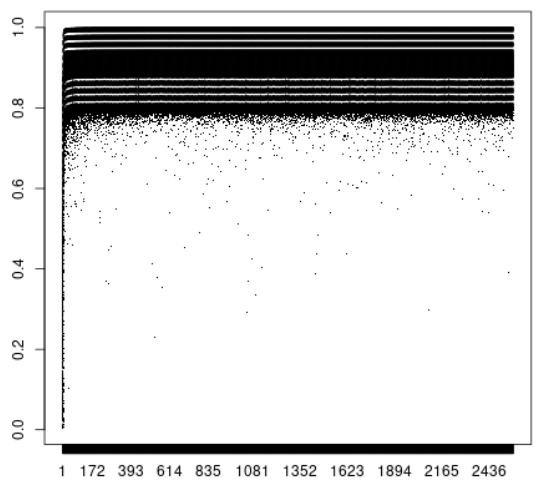
Barents fish ($n = 89$, $p = 30$)

Estimates $\widehat{B}$



Color = k

ESS ($k = 5$)



ESS = f(iteration)