Une méthode d'échantillonnage pour la reconstitution de séquences de variants de gènes d'antibiorésistance depuis des données métagénomiques

Daniel Bonnéry

Séminaire, 3 juin 2024

Daniel Bonnéry* (Ensae)
Guillaume Kon Kam King (Inrae)
Anne-Laure Abraham (Inrae)
Ouleye Sidibe (Inrae)
Sebastien Leclercq (Inrae)
Nicolas Chopin (Crest-Ensae)

Contents

1. Le modèle paramétrique et la question statistique
2. Desman
3. Stratégies
4. Travaux en cours

1 Le modèle paramétrique et la question statistique
- Observations
- Variants
- Proportions
- Erreur de mesure
- Distribution du tableau de comptages
- Approche bayesienne

2 Desman

3 Stratégies

4 Travaux en cours

1 Le modèle paramétrique et la question statistique
Observations
Variants
Proportions
Erreur de mesure
Distribution du tableau de comptages
Approche bayesienne

2 Desman

3 Stratégies

4 Travaux en cours

Observations

We observe a 3 dimensional array of counts.

$$(n_{v,s,a}) \quad \begin{array}{l} v \in \{1, \ldots, V\} \\ s \in \{1, \ldots, S\} \\ a \in \{1, \ldots, 4\} \end{array}$$

- $(n_{v,s,a})$: comptages des nucléotides
  - à la position v
  - dans l'échantillon s
  - de type a :
    - $a = 1 = A = (1, 0, 0, 0)$
    - $a = 2 = C = (0, 1, 0, 0)$
    - $a = 3 = G = (0, 0, 1, 0)$
    - $a = 4 = T = (0, 0, 0, 1)$

$$n_{.,s=1,.} = \begin{array}{c} \\ v=1 \\ v=2 \end{array} \begin{array}{c} \begin{array}{cccc} a=A & a=C & a=G & a=T \end{array} \\ \begin{pmatrix} 600 & 400 & 0 & 0 \\ 400 & 600 & 1 & 0 \end{pmatrix} \end{array}$$

$$n_{.,s=2,.} = \begin{array}{c} \\ v=1 \\ v=2 \end{array} \begin{array}{c} \begin{array}{cccc} a=A & a=C & a=G & a=T \end{array} \\ \begin{pmatrix} 200 & 800 & 0 & 1 \\ 800 & 200 & 0 & 0 \end{pmatrix} \end{array}$$

1 Le modèle paramétrique et la question statistique
- Observations
- Variants
- Proportions
- Erreur de mesure
- Distribution du tableau de comptages
- Approche bayesienne

2 Desman

3 Stratégies

4 Travaux en cours

Il semblerait qu'il y a deux variants dans chacun des deux échantillons

$$\tau = \begin{matrix} & \overset{g=1}{} & \overset{g=2}{} \\ \overset{v=1}{} & A & C \\ \overset{v=2}{} & C & A \end{matrix} \begin{pmatrix} \\ \\ \end{pmatrix} = \begin{matrix} & \overset{g=1}{} & \overset{g=2}{} \\ \overset{v=1}{} & \mathbf{1}^{0^0} & 0\mathbf{1}^{0^0} \\ \overset{v=2}{} & 0\mathbf{1}^{0^0} & \mathbf{1}^{0^0} \end{matrix} \begin{pmatrix} \\ \\ \end{pmatrix}$$

$$a = 1^{2^{3^4}}$$

- $\tau_{v,g,a}$ indique si
  - à la position v
  - le nucléotide du variant g
  - est a.

1 Le modèle paramétrique et la question statistique
- Observations
- Variants
- Proportions
- Erreur de mesure
- Distribution du tableau de comptages
- Approche bayesienne

2 Desman

3 Stratégies

4 Travaux en cours

$$\pi = \begin{array}{c} \\ g{=}1 \\ g{=}2 \end{array} \begin{array}{c} \begin{array}{cc} s{=}1 & s{=}2 \end{array} \\ \begin{pmatrix} 0.6 & 0.2 \\ .04 & 0.8 \end{pmatrix} \end{array}$$

$\pi_{g,s}$ est la proportion de

- variant g
- dans l'échantillon s.

1 Le modèle paramétrique et la question statistique
- Observations
- Variants
- Proportions
- Erreur de mesure
- Distribution du tableau de comptages
- Approche bayesienne

2 Desman

3 Stratégies

4 Travaux en cours

Erreur de mesure

La variable $\epsilon_{b,a}$ est la probabilité que la mesure d'un nucleotide de type b donne a

$$\epsilon = \begin{array}{c} \\ b{=}1 \\ b{=}2 \\ b{=}3 \\ b{=}4 \end{array} \begin{array}{c} \begin{array}{cccc} a{=}1 & a{=}2 & a{=}3 & a{=}4 \end{array} \\ \begin{pmatrix} \mathbf{0.91} & 0.03 & 0.03 & 0.03 \\ 0.03 & \mathbf{0.91} & 0.03 & 0.03 \\ 0.02 & 0.02 & \mathbf{0.94} & 0.02 \\ 0.05 & 0.04 & 0.01 & \mathbf{0.90} \end{pmatrix} \end{array}$$

1 Le modèle paramétrique et la question statistique
- Observations
- Variants
- Proportions
- Erreur de mesure
- Distribution du tableau de comptages
- Approche bayesienne

2 Desman

3 Stratégies

4 Travaux en cours

Distribution du tableau de comptages

$$\mathcal{L}\left(n|\pi,\tau,\epsilon\right)$$
$$= \prod_{v=1}^{V}\prod_{s=1}^{S}(n_{v,s,+})! \times \frac{\prod_{a=1}^{4}\left(\sum_{g=1}^{G}\sum_{b=1}^{4}\tau_{v,g,b}\epsilon_{b,a}\pi_{g,s}\right)^{n_{v,s,a}}}{\prod_{a=1}^{4}n_{v,s,a}!}$$

1 Le modèle paramétrique et la question statistique
- Observations
- Variants
- Proportions
- Erreur de mesure
- Distribution du tableau de comptages
- Approche bayesienne

2 Desman

3 Stratégies

4 Travaux en cours

Quelle est la loi de probabilité du résultat du lancé d'un dé sachant que le résultat est pair ? **A priori**, le dé est non pipé, et la loi η du résultat est uniforme.

Règle de Bayes:

$$\eta(\{x\}|\{2,4,6\}) = \frac{\eta(\{x\} \cap \{2,4,6\})}{\eta(\{2\}) + \eta(\{4\}) + \eta(\{6\})}$$

Notre problème est similaire. Une loi sur le vecteur de variables aléatoires (n, π, τ, ϵ) est construit à partir de lois a priori sur τ,π et ϵ, et à partir de la loi de n sachant ces paramêtres.

$$\eta(n, \pi, \tau, \epsilon) = \eta(n \mid \tau, \pi, \epsilon) \times \eta(\tau) \times \eta(\pi) \times \eta(\epsilon)$$

La solution de notre problème est:

$$\eta(\pi, \tau, \epsilon \mid n) = \frac{\eta(n, \pi, \tau, \epsilon)}{\int \eta(n, \pi, \tau, \epsilon) \mathrm{d}\tau \ \mathrm{d}\pi \ \mathrm{d}\epsilon}$$

1 Le modèle paramétrique et la question statistique

2 Desman
- Présentation
- Algorithme de Gibbs
- Les étapes de Gibbs dans Desman
- Identification d'un problème
- Diagnostics

3 Stratégies

4 Travaux en cours

1 Le modèle paramétrique et la question statistique

2 Desman
- Présentation
- Algorithme de Gibbs
- Les étapes de Gibbs dans Desman
- Identification d'un problème
- Diagnostics

3 Stratégies

4 Travaux en cours

Quince, C., Delmont, T. O., Raguideau, S., Alneberg, J., Darling, A. E., Collins, G., and Eren, A. M. (2017).
Desman: a new tool for de novo extraction of strains from metagenomes.
Genome biology, 18:1–22.

- Desman est un algorithme d'échantillonage qui permet d'obtenir un échantillon dont les propriétés sont proches d'un échantillon iid de la loi a posteriori de (τ,π,ϵ).
- Il est basé sur des méthodes de Monte Carlo par chaînes de Markov
- obtenues par échantillonnage de Gibbs
- après introduction de variables latentes pour obtenir un maximum de lois conjuguées et éviter de devoir tirer avec Metropolis Hastings

1 Le modèle paramétrique et la question statistique

2 Desman
- Présentation
- Algorithme de Gibbs
- Les étapes de Gibbs dans Desman
- Identification d'un problème
- Diagnostics

3 Stratégies

4 Travaux en cours

Algorithme de Gibbs

Algorithm 1 Algorithme de Gibbs

1: Initialiser $\boldsymbol{\Theta}^{(0)} = (\Theta_1^{(0)}, \Theta_2^{(0)}, \ldots, \Theta_I^{(0)})$
2: **for** $t = 1$ to T **do**
3: **for** $i = 1$ to I **do**
4: Échantillonner
5: $\Theta_i^{(t)} \sim \eta\left(\Theta_i \mid \Theta_1^{(t)}, \ldots, \Theta_{i-1}^{(t)}, \Theta_{i+1}^{(t-1)}, \ldots, \Theta_n^{(t-1)}\right)$
6: **end for**
7: **end for**

On appelle noyau de transition la distribution $P^{\Theta^{(t+1)}|\Theta^{(t)}}$

Explications

- L'algorithme commence par une initialisation de $\mathbf{\Theta}^{(0)}$.
- À chaque étape t, chaque variable Θ_i est mise à jour en échantillonnant de sa distribution conditionnelle.
- Le processus est répété pour un nombre d'itérations T.
- En fin de compte, les échantillons $\mathbf{\Theta}^{(T)}$ sont utilisés pour estimer la distribution cible.

$$\Theta_i^{(t)} \sim \eta\left(\Theta_i \mid \Theta_1^{(t)}, \ldots, \Theta_{i-1}^{(t)}, \Theta_{i+1}^{(t-1)}, \ldots, \Theta_n^{(t-1)}\right)$$

Comment faire ?

1. Lois conjuguées
2. Metropolis Hasting

1 Le modèle paramétrique et la question statistique

2 Desman
- Présentation
- Algorithme de Gibbs
- Les étapes de Gibbs dans Desman
- Identification d'un problème
- Diagnostics

3 Stratégies

4 Travaux en cours

Algorithm 2 Noyau MCMC

procedure $M(n, \Theta = (\tau, \pi, \epsilon); \alpha_\pi, \alpha_\epsilon, e)$

$(\nu, \mu) \leftarrow \text{sample}_{(\nu,\mu)}(n, \Theta)$

$\pi \leftarrow \text{sample}_\pi(\mu; \alpha_\pi)$

$\tau \leftarrow \text{sample}_\tau(n, \tau, \pi, \epsilon)$

$\epsilon \leftarrow \text{sample}_\epsilon(\nu; \alpha_\epsilon)$

Return (τ, π, ϵ)

end procedure

1 Le modèle paramétrique et la question statistique

2 Desman
- Présentation
- Algorithme de Gibbs
- Les étapes de Gibbs dans Desman
- Identification d'un problème
- Diagnostics

3 Stratégies

4 Travaux en cours

L'étape de Gibbs correspondante à τ semble problématique: τ ne varie pas avec t.

1 Le modèle paramétrique et la question statistique

2 Desman
- Présentation
- Algorithme de Gibbs
- Les étapes de Gibbs dans Desman
- Identification d'un problème
- Diagnostics

3 Stratégies

4 Travaux en cours

$$n_{.,s=1,.} = \begin{array}{c} \\ v=1 \\ v=2 \end{array} \begin{array}{c} \begin{array}{cccc} a=A & a=C & a=G & a=T \end{array} \\ \begin{pmatrix} 600 & 400 & 0 & 0 \\ 400 & 600 & 1 & 0 \end{pmatrix} \end{array}$$

$$n_{.,s=2,.} = \begin{array}{c} \\ v=1 \\ v=2 \end{array} \begin{array}{c} \begin{array}{cccc} a=A & a=C & a=G & a=T \end{array} \\ \begin{pmatrix} 200 & 800 & 0 & 1 \\ 800 & 200 & 0 & 0 \end{pmatrix} \end{array}$$

$$\begin{array}{ccc} t=1 & & t=2 \end{array} \qquad \begin{array}{ccc} & & t=3 \end{array}$$

$$\begin{array}{c} \\ v=1 \\ v=2 \end{array} \begin{array}{c} \begin{array}{cc} g=1 & g=2 \end{array} \\ \begin{pmatrix} A & C \\ A & C \end{pmatrix} \end{array} \rightarrow \begin{array}{c} \\ v=1 \\ v=2 \end{array} \begin{array}{c} \begin{array}{cc} g=1 & g=2 \end{array} \\ \begin{pmatrix} A & C \\ C & C \end{pmatrix} \end{array} \rightarrow \begin{array}{c} \\ v=1 \\ v=2 \end{array} \begin{array}{c} \begin{array}{cc} g=1 & g=2 \end{array} \\ \begin{pmatrix} A & C \\ C & A \end{pmatrix} \end{array}$$

1 Le modèle paramétrique et la question statistique

2 Desman

3 Stratégies
- Block sampling
- Fixed variants
- Cibler ρ
- SMC and Tempering

4 Travaux en cours

1 Le modèle paramétrique et la question statistique

2 Desman

3 Stratégies
- Block sampling
- Fixed variants
- Cibler ρ
- SMC and Tempering

4 Travaux en cours

Il s'agit d'échantillonner en block $\tau_{v,.,.}$:

$$\begin{array}{c} \\ v{=}1 \\ v{=}2 \end{array} \begin{array}{c} t{=}1 \\ \begin{array}{cc} g{=}1 & g{=}2 \end{array} \\ \begin{pmatrix} A & C \\ A & C \end{pmatrix} \end{array} \rightarrow \begin{array}{c} \\ v{=}1 \\ v{=}2 \end{array} \begin{array}{c} t{=}2 \\ \begin{array}{cc} g{=}1 & g{=}2 \end{array} \\ \begin{pmatrix} A & C \\ C & A \end{pmatrix} \end{array}$$

Problème:

4^1	4
4^2	16
4^3	64
4^4	256
4^5	1 024
4^{10}	1 048 576
4^{15}	1 073 741 824
4^{20}	1 099 511 627 776

1 Le modèle paramétrique et la question statistique

2 Desman

3 Stratégies
- Block sampling
- Fixed variants
- Cibler ρ
- SMC and Tempering

4 Travaux en cours

Il s'agit de remplacer la distribution a priori de τ par $\mathrm{Uniform}(\mathcal{T})$

- Metropolis Hasting est évité.
- Satisfaisant lorsque les variants sont connus
- Ne répond pas à l'objectif initial de détection.

1 Le modèle paramétrique et la question statistique

2 Desman

3 Stratégies
- Block sampling
- Fixed variants
- Cibler ρ
- SMC and Tempering

4 Travaux en cours

$$\mathcal{L}\left(n|\pi,\tau,\epsilon\right)$$

$$= \prod_{v=1}^{V}\prod_{s=1}^{S}(n_{v,s,+})! \times \frac{\prod_{a=1}^{4}\left(\sum_{g=1}^{G}\sum_{b=1}^{4}\tau_{v,g,b}\epsilon_{b,a}\pi_{g,s}\right)^{n_{v,s,a}}}{\prod_{a=1}^{4}n_{v,s,a}!}$$

$$= \prod_{v=1}^{V}\prod_{s=1}^{S}(n_{v,s,+})! \times \frac{\prod_{a=1}^{4}\left(\sum_{g=1}^{G}\rho_{v,g,a}\pi_{g,s}\right)^{n_{v,s,a}}}{\prod_{a=1}^{4}n_{v,s,a}!}$$

avec

$$\rho_{v,g,a} = \sum_{b=1}^{4}\tau_{vgb}\epsilon_{ba}$$

$$\rho_{.,g,.} = \begin{matrix} \\ v=1 \\ \vdots \\ v=V \end{matrix} \begin{matrix} \begin{matrix} A & C & G & T \end{matrix} \\ \begin{pmatrix} 0.03 & \mathbf{0.91} & 0.03 & 0.03 \\ \vdots & \vdots & \vdots & \vdots \\ 0.05 & 0.04 & \mathbf{0.90} & 0.01 \end{pmatrix} \end{matrix}$$

- On utilise un hyperparamêtre de forme très petit pour la loi de Dirichlet, ce qui assure la concentration des $\rho_{v,g,.}$ autour des sommets A,C,G, ou T.
- On garde des lois conjuguées pour ρ.
- La mise à jour se fait naturellement par block.
- Rajout d'un a priori non conjugué sur l'hyperparamêtre de forme.

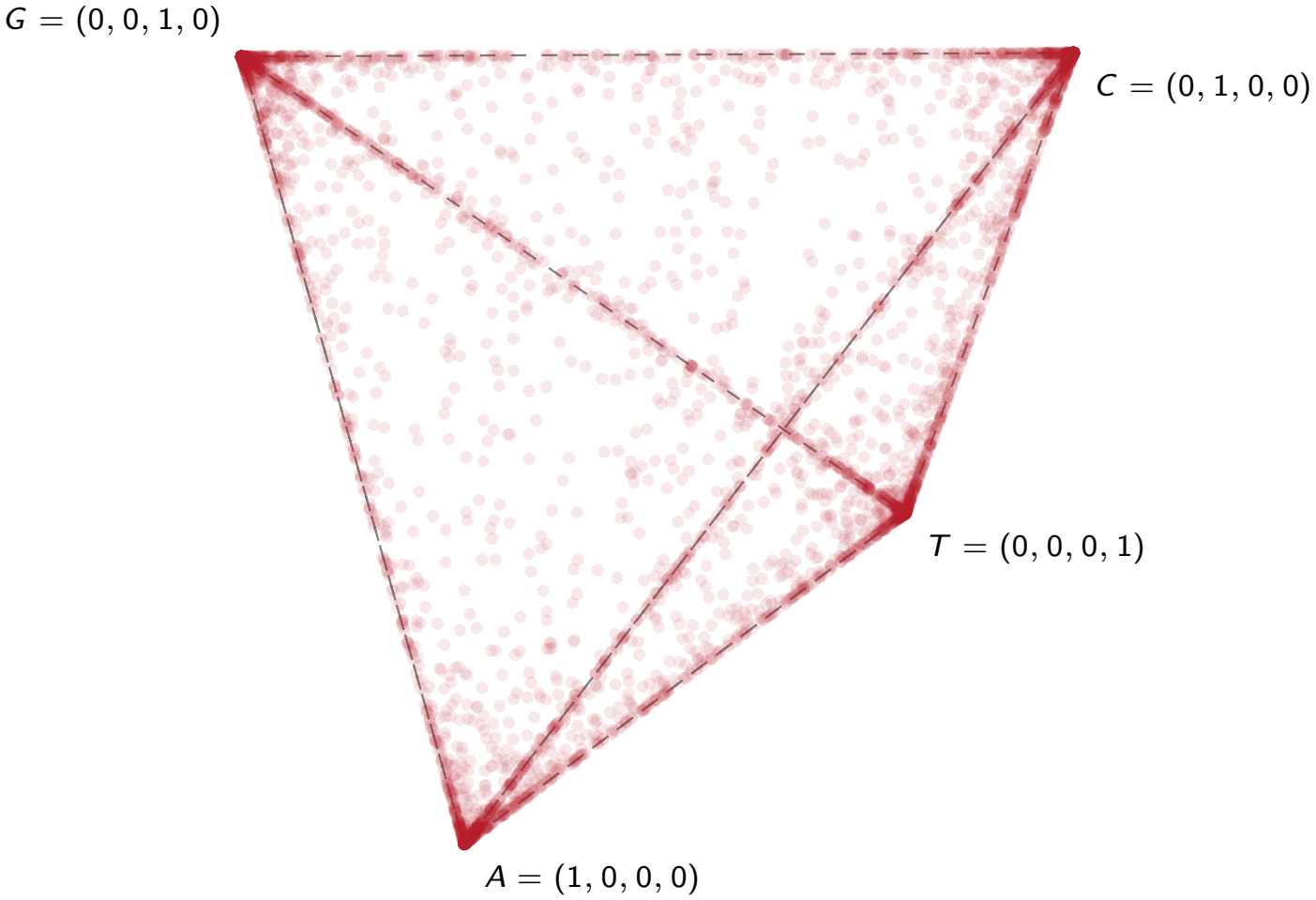
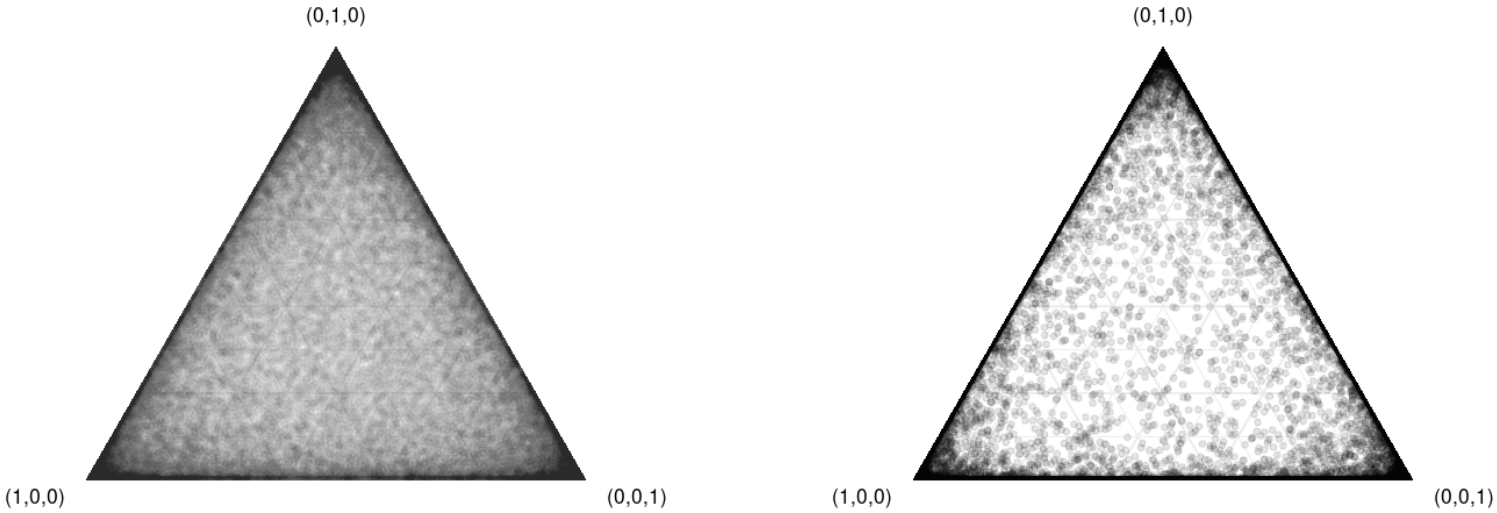


Figure 1: Échantillon d'une loi de Dirichlet, $\alpha_\tau = 0.1\mathbb{1}_4$



(a) $\alpha = 0.1\ \mathbb{1}_3$ (b) $\alpha = 0.01\ \mathbb{1}_3$

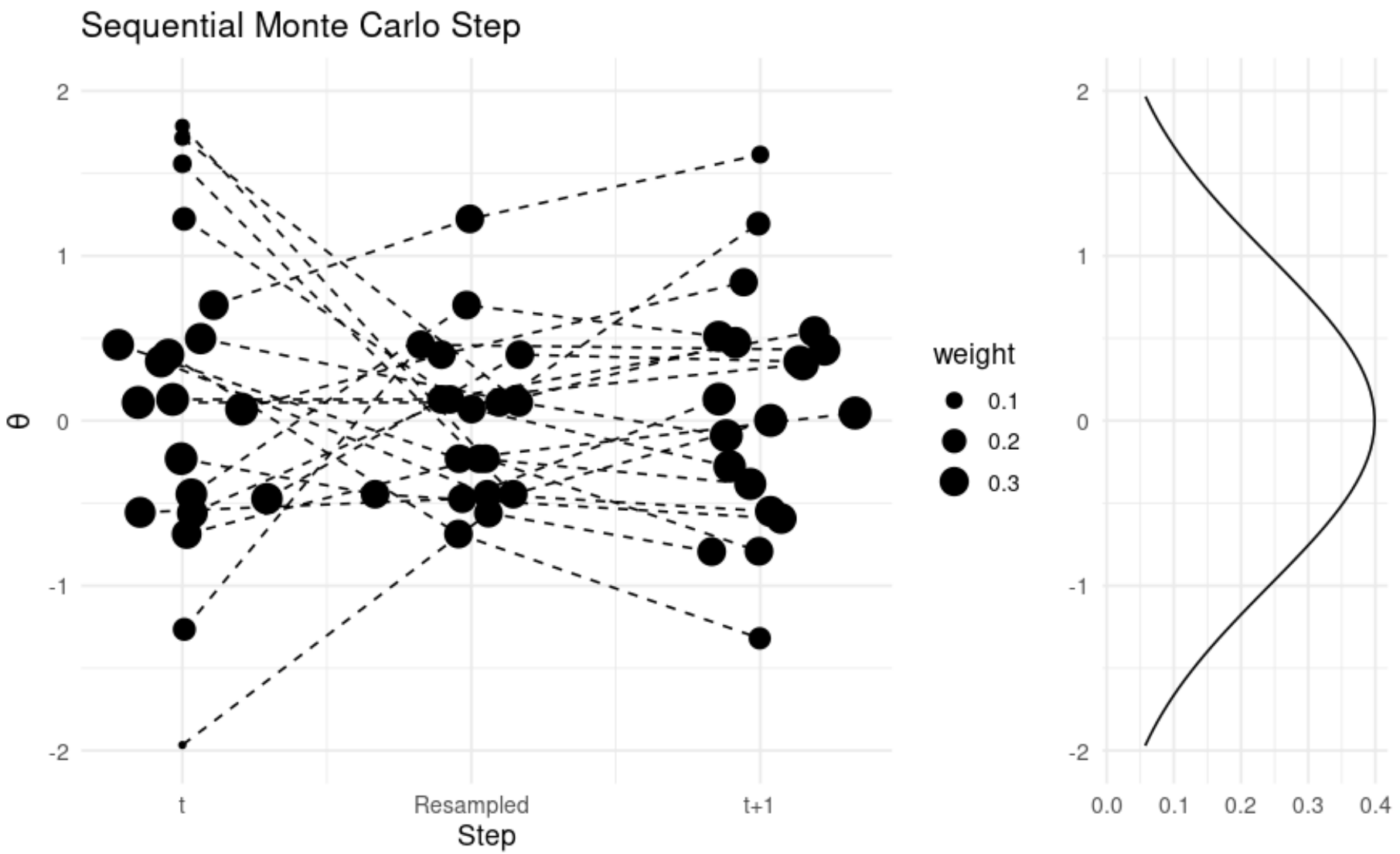
Figure 2: Échantillon de distribution de Dirichlet sur un 3-simplexe

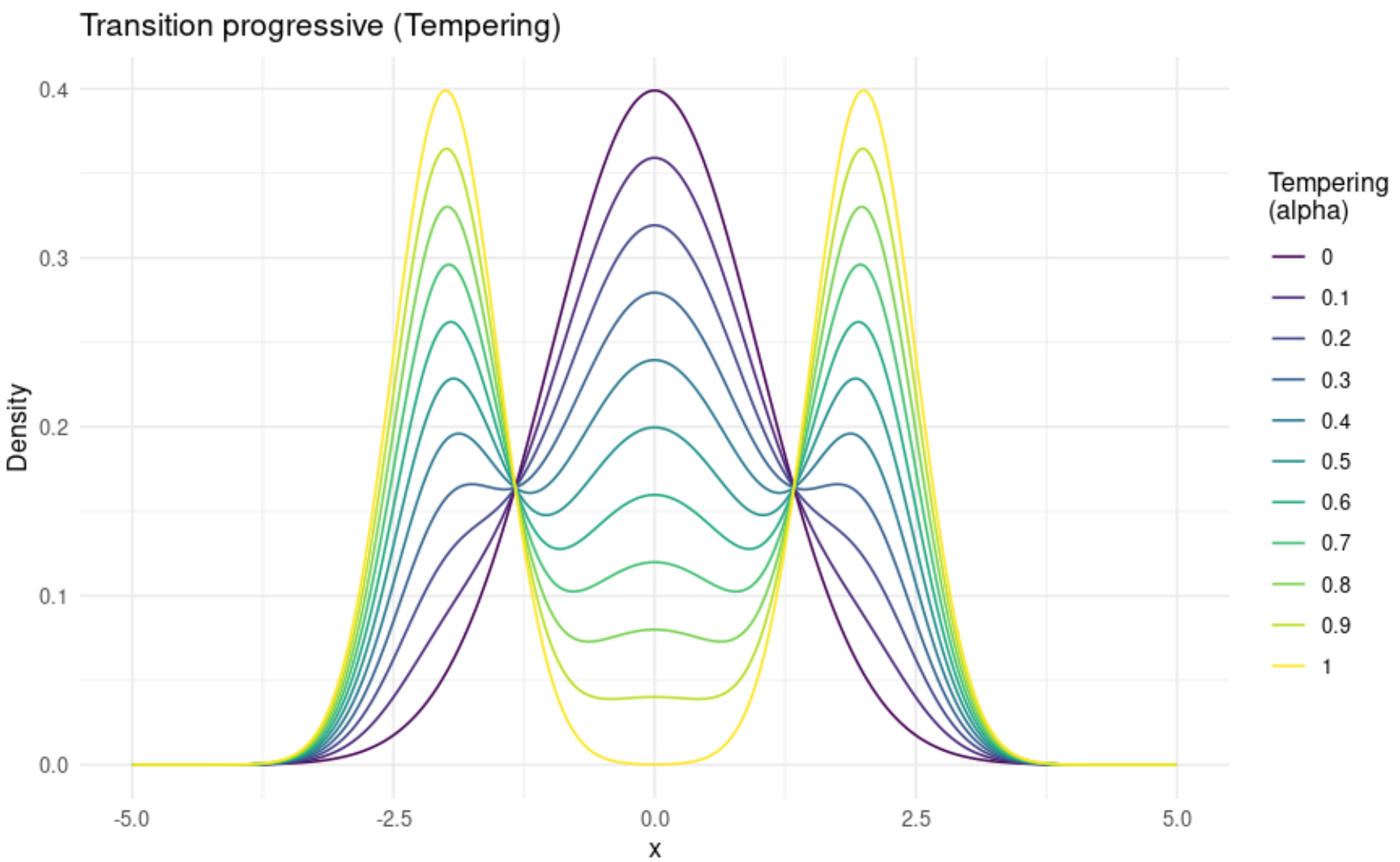
1 Le modèle paramétrique et la question statistique

2 Desman

3 Stratégies
- Block sampling
- Fixed variants
- Cibler ρ
- SMC and Tempering

4 Travaux en cours


Sequential Monte Carlo Step
θ
2
1
0
-1
-2
t
Resampled
t+1
Step
weight
0.1
0.2
0.3
0.0
0.1
0.2
0.3
0.4


Transition progressive (Tempering)
Density
0.4
0.3
0.2
0.1
0.0
-5.0
-2.5
0.0
2.5
5.0
x
Tempering (alpha)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Stratégies de tempering:

- Puissance de la vraisemblance
- Data tempering
- Paramêtre de tempering naturel.

Stratégie retenue

- Cibler τ, ϵ et π
- Relaxer τ
- Le paramêtre de forme sur τ devient le paramêtre de tempering naturel.

Variable	Models				
	Desman	Model 1 Block Sampling	Model 2 Fixed variants	Model 4 Tempering	Model 3 Relaxation
G	unobserved,fixed				
V	observed,fixed				
S	observed,fixed				
κ_τ	-			Dirac$(0.1, 0.1)$	-
$\alpha_\tau \mid \kappa_\tau$	-			Beta(κ_τ)	-
$\tau \mid \alpha_\tau$	$\tau_{v,g,.} \sim$ Uniform$(\{A, C, G, T\})$		Dirac$_\mathcal{T}$	$\forall v, g,\ \tau_{v,g,.} \overset{\text{iid}}{\sim}$ Dirichlet$(\alpha_\tau \mathbb{1}_4)$	-
κ_ϵ	Dirac$(0.1\mathbb{1}_4)$	Dirac$(0.01,1)$			-
$\tilde{\epsilon} \mid \kappa_\epsilon$	-	Beta(κ_ϵ)			-
$\epsilon \mid \kappa_\epsilon, \alpha_\epsilon$	$\forall b, \epsilon_{b,.} \sim$ Dirichlet(κ_ϵ)	Dirac $(\tilde{\epsilon}/3(J_4 - I_4) + (1 - \tilde{\epsilon})I_4)$			-
κ_ρ	-				Dirac$(1, 10)$
α_ρ	-				Gamma(κ_ρ)
$\rho \mid \alpha_\rho$	-				$\forall v, g,\ \rho_{v,g,.}$ Dirichlet$(\alpha_\rho$
κ_π	-	Dirac$(0.1, 1)$			
α_π	Dirac(0.1)	Beta(κ_π)			
π	$\forall g,\ \pi_{.,s} \sim$ Dirichlet$(\alpha_\pi \mathbb{1}_G)$				
$n_{.,.,.,+} \mid$	observed,fixed				
$n \mid n_{.,.,.,+}, \rho$	observed,$\forall v, s, n_{v,s,.} \sim$ Multinomial $\left(n_{v,s,+}, \sum_{g=1}^{G} \rho_{v,g,.} \pi_{g,s}\right)$				

1 Le modèle paramétrique et la question statistique

2 Desman

3 Stratégies

4 Travaux en cours

- Développement d'algorithmes en R basés sur jags.
- SMC
- tempering.
- Choix de modèle (G)
- ESS
- Exploitation des trajectoires de particules obtenues.

Dau, H.-D. and Chopin, N. (2022). Waste-free sequential monte carlo. Journal of the Royal Statistical Society Series B: Statistical Methodology, 84(1):114–148.

Malsiner-Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2017). Identifying mixtures of mixtures using bayesian estimation. Journal of Computational and Graphical Statistics, 26(2):285–295.

Quince, C., Delmont, T. O., Raguideau, S., Alneberg, J., Darling, A. E., Collins, G., and Eren, A. M. (2017). Desman: a new tool for de novo extraction of strains from metagenomes. Genome biology, 18:1–22.