Régression bayésienne sur profils d'exposition : application en épidémiologie des rayonnements ionisants

Marion Belloni[1], Sophie Ancelet[1], Chantal Guihenneuc[2]

[1] IRSN, Laboratoire d'EPIDémiologie des rayonnements ionisants
[2] BioSTM UR 7537, Université de Paris

IRSN
INSTITUT DE RADIOPROTECTION ET DE SÛRETÉ NUCLÉAIRE

AppliBUGS, 11 Juin 2020

UNIVERSITÉ PARIS DESCARTES

Contexte (1/2)

- Intérêt grandissant pour l'exposome (ensemble des expositions environnementales - i.e., non génétiques - reçues par un individu au cours de sa vie (C. Wild (2005))
  - ➔ Comprendre comment une combinaison d'expositions environnementales peut conduire au développement d'une maladie
  - ➔ Mieux comprendre l'étiologie des pathologies multifactorielles
  - ➔ Conduire à de meilleures stratégies de prévention de ces pathologies
- Les cancers : pathologies chroniques et multifactorielles pour lesquelles le concept d'exposome est essentiel
  - ➔ Combinaison d'expositions environnementales et de facteurs génétiques et comportementaux qui peuvent agir simultanément et interagir

Contexte (2/2)

- En épidémiologie, important de prendre en compte correctement ces expositions multiples pour estimer (ou prédire) un risque de cancer au niveau individuel et/ou de la population
  - ➔ Pourtant, historiquement, études épidémiologiques principalement monofactorielles (i.e., caractérisation de l'effet d'un seul facteur d'intérêt après ajustement sur un ou plusieurs facteurs de confusion potentiels)
  - ➔ En épidémiologie des rayonnements ionisants, la question de l'estimation d'un risque de cancer en situation d'exposition à de multiples stresseurs radiologiques n'a jamais été considérée

Une difficulté majeure en situation de co-expositions : la multicolinéarité

- En situation de co-expositions à de multiples stresseurs environnementaux, les covariables d'exposition sont souvent fortement corrélées entre elles ➔ multicolinéarité
  - ➔ Exemple : Cas de travailleurs simultanément exposés à de multiples stresseurs chimiques, radiologiques et/ou physiques (e.g. chaleur) pendant leur activité professionnelle
- Si on utilise un modèle de régression multiple :
  - ➔ Les estimateurs peuvent être instables, donc non-interprétables;
  - ➔ Perte de puissance statistique

Méthodes statistiques pour pallier à un problème de multicolinéarité (1/2)

- EWAS : Estimation par régression séparée de l'association entre l'événement d'intérêt et chaque exposition étudiée
  ➥ Approche exploratoire car ne permet pas de comprendre la relation entre une combinaison d'expositions et la pathologie étudiée
- Sélection de variables en régression (ex : Elastic-net)
  ➥ Permet d'identifier un petit sous-ensemble de variables d'expositions ayant un impact majeur sur le risque d'intérêt. Surtout utile quand le nombre de variables est très important.
- Approches par réduction de dimension (ex, régression sur composantes principales, régression sPLS)
  ➥ Construction des axes et estimation du risque dans 2 étapes disjointes. Quid de la prise en compte des incertitudes sur la construction des axes ?

Méthodes statistiques pour pallier à un problème de multicolinéarité (2/2)

- Algorithmes de machine-learning (ex : k-means, forêts aléatoires)
  ➥ Approches efficaces et pertinentes quand le nombre de variables est très important mais ne permet pas d'estimer les risques d'intérêt
- Approches par classification de données d'expositions corrélées
  ➥ Analyses en classes latentes (LCA) : la variable-réponse n'est pas prise en compte dans la classification
  ➥ Modèles de mélange par régression bayésienne sur profils d'exposition (BPRM pour "Bayesian Profile Regression Mixture") : Consiste à identifier des groupes d'individus ayant des caractéristiques d'exposition proches et associés à un risque proche vis-à-vis de la pathologie d'intérêt

Enjeu spécifique et objectif

- Enjeu spécifique : Estimer un risque de décès par cancer à partir de données de survie (fortement) censurées à droite et de covariables d'exposition disponibles en nombre réduit et fortement corrélées entre elles
- Le package R PReMiuM permet l'estimation des modèles BPRM mais :
  - Le modèle de survie en excès de risque instantané (EHR) - classiquement utilisé en épidémiologie des rayonnements ionisants - n'est pas une option du package (seule option possible actuellement : données de survie censurées suivant une loi de Weibull)
  - Les erreurs de mesure sur les expositions ne peuvent pas être prises en compte (besoin pour nos futures analyses)
- Objectif ➔ Extension des modèles BPRM à des données de survie censurées suivant un modèle en EHR

Cas d'étude en épidémiologie des rayonnements ionisants

Effets sanitaires potentiellement induits par des expositions professionnelles multiples et à faibles doses aux rayonnements ionisants

- Peu étudiés et donc *a fortiori* peu connus
- Élaboration des normes de radioprotection principalement fondée sur un cadre d'exposition **mono-factorielle**

Cas d'étude : Estimation du risque de décès par cancer du poumon chez les mineurs d'uranium français chroniquement et simultanément exposés :

- au radon
- aux poussières d'uranium
- aux rayonnements gamma

dans le cadre de leur activité professionnelle

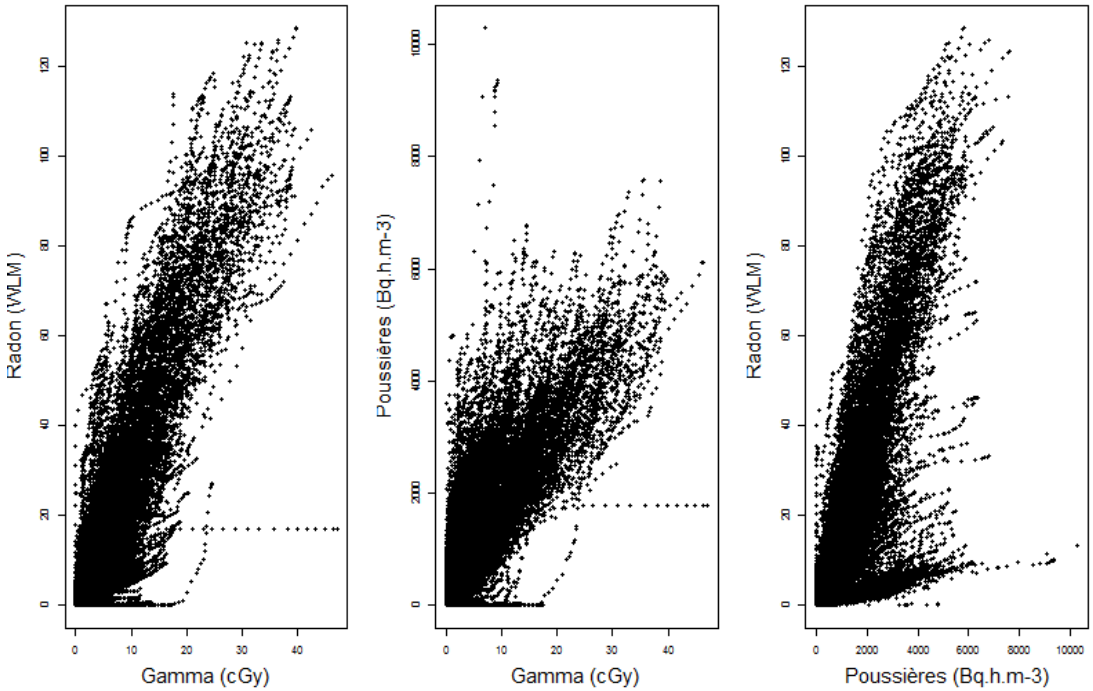
➔ Risque non estimé jusqu'à présent – Focus sur l'exposition au radon

➔ Population d'étude : Mineurs français ayant travaillé au moins un an en tant que mineur d'uranium et embauchés après 1955 dans le groupe CEA-COGEMA

Cohorte post-55 des mineurs d'uranium français

Mineurs, n	3377
Âge à l'entrée dans l'étude, moyenne [min-max]	28.3 [16.9-57.7]
Durée de travail en années, moyenne [min-max]	16.7 [1.0-40.9]
Durée de suivi en années, moyenne [min-max]	32.8 [0.1-51.0]
Statut vital n (%)	
Vivant $<$ 85 ans	2412 (71.4)
Vivant $\geq$ 85 ans	74 (2.2)
Mort par cancer du poumon	94 (2.8)
Mort d'une autre cause	777 (23.0)
Perdu de vue	20 (0.6)
Exposition au radon	
Mineurs exposés, n (%)	2,910 (86.2)
Durée d'exposition en années, moyenne [min-max]	12.9 (1.0–35.0)
Exposition cumulée (en WLM), moyenne [min-max]	17.8 (0.01–128.4)
Exposition aux rayonnements gamma	
Mineurs exposés, n (%)	3,240 (95.9)
Durée d'exposition en années, moyenne [min-max]	13.2 (1.0–36.0)
Exposition cumulée (en mGy), moyenne [min-max]	54.9 (0.20–470.1)
Exposition aux poussières d'uranium	
Mineurs exposés, n (%)	2,746 (81.3)
Durée d'exposition en années, moyenne [min-max]	12.9 (1.0–35.0)
Exposition cumulée (en $kBq.m^{-3}.h$), moyenne [min-max]	1.64 (0.01–10.4)

Corrélation entre les expositions radiologiques cumulées



Coefficients de corrélation de Pearson:

	Gamma	Radon	Poussiere
Gamma	1.	0.90	0.81
Radon	0.90	1.	0.75
Poussiere	0.81	0.75	1.

Structure générale d'un modèle de régression sur profils d'exposition

- Ce modèle consiste à définir des groupes de mineurs d'uranium ayant des profils d'exposition proches et associés à un risque proche de décès par cancer du poumon.
- Modèle hiérarchique incluant 3 sous-modèles:
  - Le *sous-modèle de maladie* : il modélise le lien entre l'âge au décès par cancer du poumon et le risque associé pour chaque groupe
  - Le *sous-modèle d'exposition* : il modélise les caractéristiques d'exposition de chaque groupe
  - Le *sous-modèle d'attribution* : il modélise la probabilité d'appartenance d'un mineur à chaque groupe

Sous-modèle de maladie

- La variable réponse : l'âge au décès par cancer du poumon (variable censurée à droite et tronquée à gauche) de chaque mineur
- Modèle de survie en excès de risque instantané (EHR)
- Risque instantané de décès par cancer du poumon du mineur i au temps t défini par :

$$h_i(t) = h_0(t)(1 + \beta_{C_i})$$

- $h_0(t)$: risque instantané de base au temps t, supposé constant par morceaux - 4 classes d'âge considérées : < 40 ans, 40-55 ans, 55-70 ans, > 70 ans ➔ 4 paramètres $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$
- C_i est le label de groupe du mineur i

➔ Si deux individus sont dans le même groupe, ils ont le même risque de décès par cancer du poumon

Sous-modèle d'exposition

Le profil d'expositions $Z_i = (Z_{i,1}, ..., Z_{i,P})$ d'un mineur i, avec P le nombre de caractéristiques d'exposition, se compose :

- De variables continues $Z_i^{Cont,q} | C_i = c \sim LogNormale(\mu_c^q, \sigma_c^q)$
  - ➔ Les mesures d'expositions professionnelles cumulées sur toute la période de suivi et laggées de 5 ans au radon, aux rayonnements gamma et aux poussières d'uranium du mineur i
  - ➔ L'âge à la première exposition du mineur i
- De variables catégorielles $Z_i^{Cat,q} | C_i = c \sim Multinomiale(p_c^q)$
  - ➔ Le type de poste (5 catégories) qui est un proxy sur les conditions d'exposition
  - ➔ La mine dans laquelle le mineur a travaillé (2 catégories : Hérault (roche sédimentaire) ou autre (roche granitique))
  - ➔ La durée d'exposition (4 catégories)

Sous modèle d'attribution

Le nombre de groupes est supposé inconnu et donc estimé. Seul le nombre maximal C_{max} est fixé.

$$C_i \overset{iid}{\sim} Multinomial(\psi_1, \psi_2, ..., \psi_{C_{max}})$$

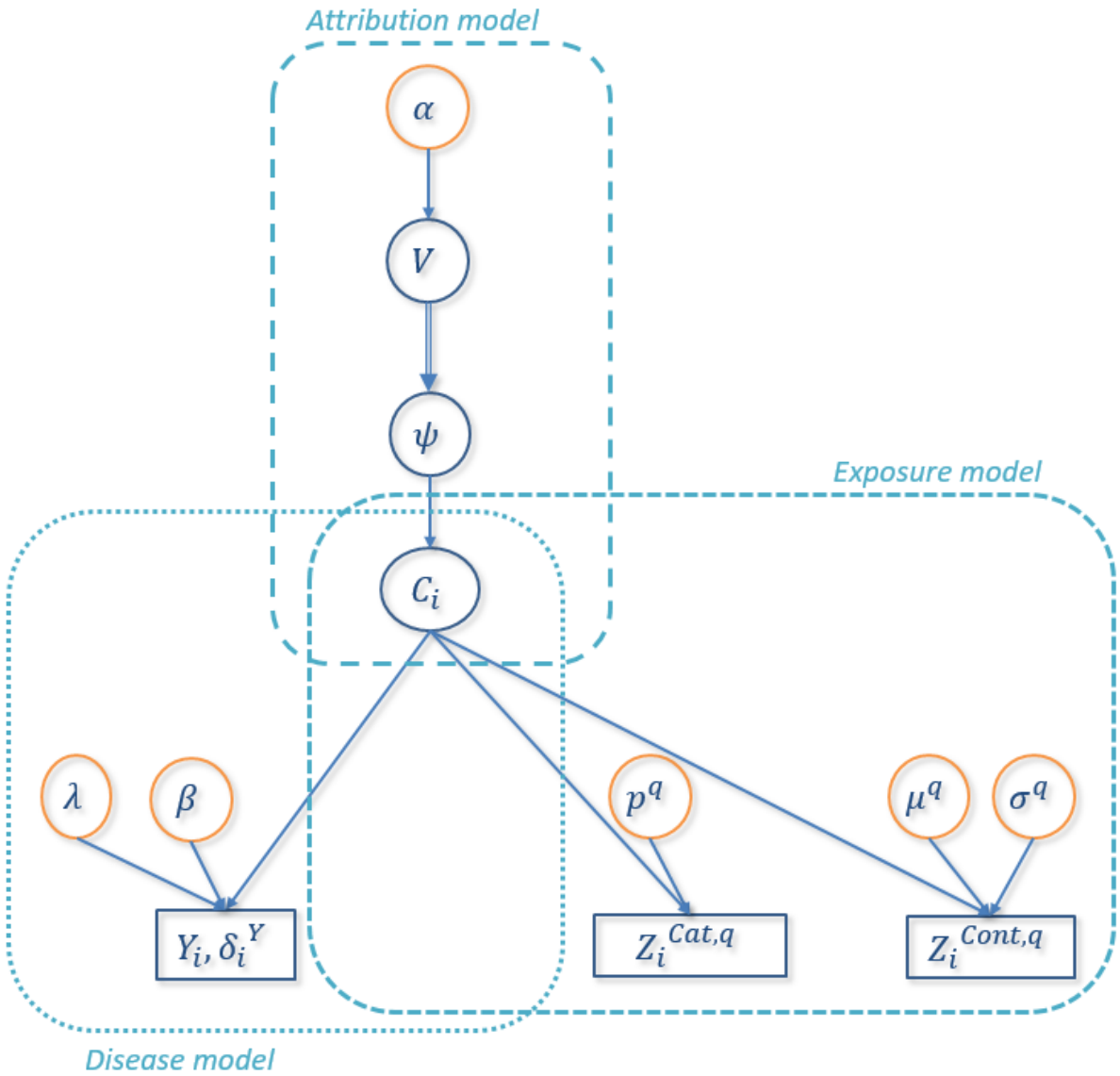
$$\psi_c = P(C = c)$$

➥ Le vecteur de probabilités $\psi = (\psi_1, \psi_2, ..., \psi_{C_{max}})$ suit un processus de Dirichlet tronqué construit comme suit:

$$\begin{cases} V_c|\alpha & \sim & Beta(1, \alpha), c = 1, ..., C_{max} - 1 \\ \psi_c & = & V_c(1 - \sum_{k=1}^{c-1} \psi_k), c = 1, ..., C_{max} - 1 \\ \psi_{C_{max}} & = & 1 - \sum_{k=1}^{C_{max}} \psi_k \end{cases}$$

Rôle du paramètre α: plus α est grand, plus le nombre de groupes non vides est grand

Diagramme acyclique orienté du modèle



Choix des distributions *a priori*

- Sous-modèle de maladie:
  - $\beta_c \sim Normale(0; 1000^2)$ tronqué pour que $h_i(t) > 0 \ \forall i \ \forall t$: Excès de risque instantané du groupe c
  - $\lambda_j \sim Gamma(a; b)$: Taux de mortalité de base de la classe d'âge $]s_{j-1}, s_j]$; a et b estimés à partir de données nationales
- Sous-modèle d'exposition:

  Expositions continues ($LogN(\mu_c^q, \sigma_c^q)$):
  - $\mu_c^q \sim N(0; 1000^2)$ si $q =$ âge à la première exposition
  - $\mu_c^q \sim N(c; d)$ si $q =$ radon ou gamma ou poussière; c et d estimés à partir des expositions de la cohorte allemande des mineurs d'uranium
  - $\sigma_c^q \sim Uniform(0; 100)$

  Expositions catégorielles ($Multinomial(p_c^q)$):
  - $p_c^q \sim Dirichlet(1/2; ...; 1/2)$
- Sous-modèle d'attribution:
  - $\alpha \sim Uniform(0.3; 10)$ selon Molitor *et al.* (2010)

Inférence bayésienne

- Algorithme Monte-Carlo par Chaînes de Markov (MCMC) de type Metropolis-Hastings within Gibbs adaptatif codé sous Python 3.4
- 1 chaîne :
  - 100 phases adaptatives de 100 itérations chacune pour mettre à jour les variances des lois de proposition afin de viser un taux d'acceptation de 40% pour les paramètres seuls et 20% pour les vecteurs
  - 10.000 itérations de temps de chauffe
  - 150.000 itérations (en plus des phases adaptatives et du temps de chauffe)
  - Pas de de stockage (thin) de 20

  → 7500 itérations dans l'échantillon *a posteriori*.
- 3 mouvements de changement de label de groupe: Recommandé pour éviter la convergence vers un mode local
- Analyse de convergence de l'algorithme et des corrélations intra-chaînes

Post-traitement

Post-traitement nécessaire afin d'obtenir la meilleure partition de mineurs d'uranium selon leurs profils:

- Calcule des matrices S_k pour toutes les itérations k telles que $S_k(i,j) = 1$ si i et j appartiennent au même groupe à l'itération k, et 0 sinon. $\mathbf{S}$ est la moyenne des S_k où l'élément S_{ij} est la probabilité que les individus i et j soient dans le même groupe
- Choix de la partition générée par l'algorithme MCMC qui minimise la distance des moindres carrés entre $\mathbf{S}$ et S_k

L'algorithme ne converge pas...

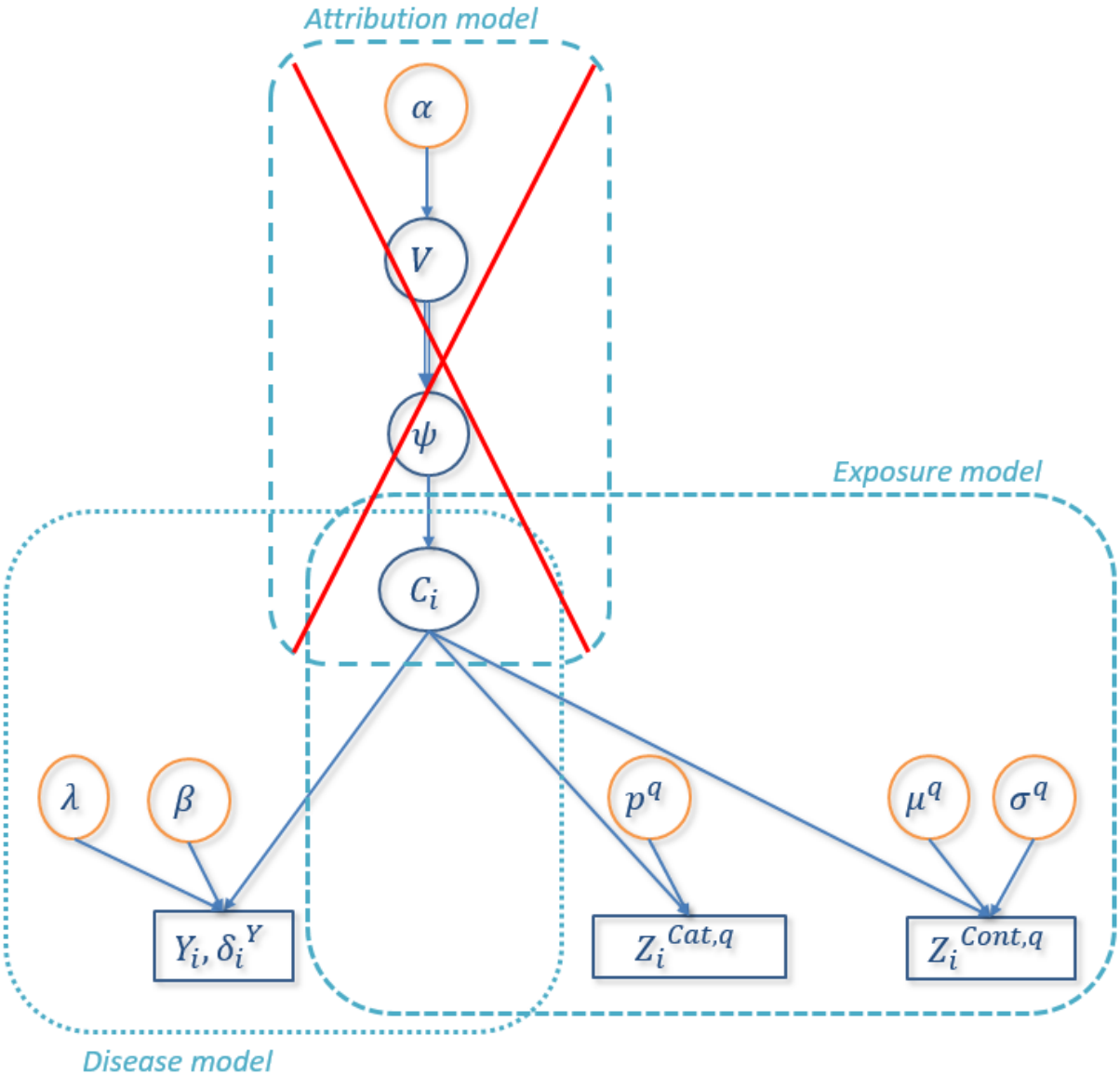
Nombre de groupes non-vides estimé en fonction de la valeur initiale du paramètre α:

Valeur initiale α	Nombre de groupes non-vides estimé
0.5	5
1.5	6
2.5	8
3.5	8
4.5	7
5.5	7
6.5	7
7.5	8
8.5	7
9.5	8

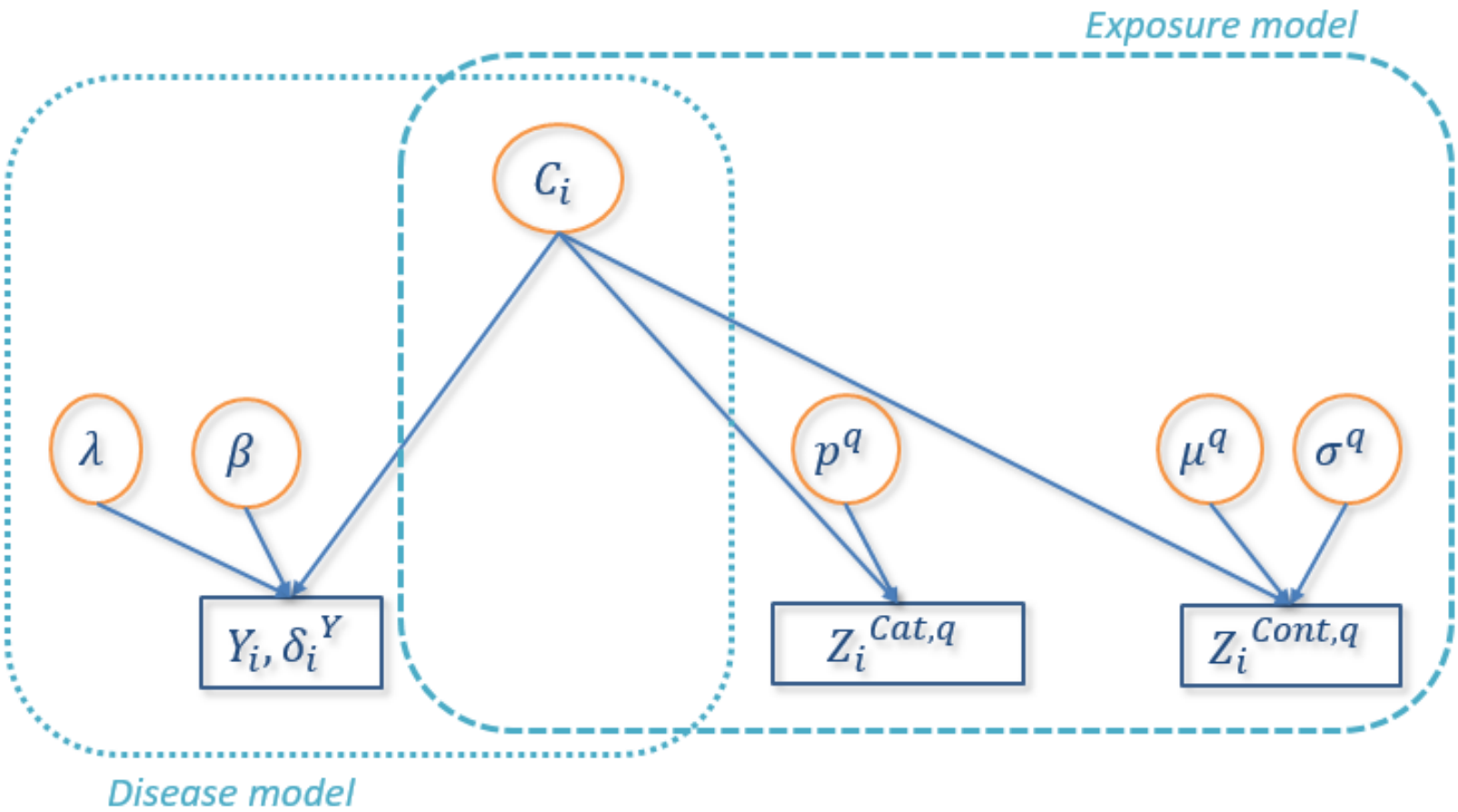
➔ Au sein de chaque partition, aucun indice de non-convergence : Convergence vers des modes locaux

➔ Estimation à nombre de groupes non-vides fixé entre 5 et 8

Estimation à nombre de groupes non-vides fixé (1/2)



Estimation à nombre de groupes non-vides fixé (2/2)



$C_i \sim Multinomiale(1/C_{max}, ..., 1/C_{max})$

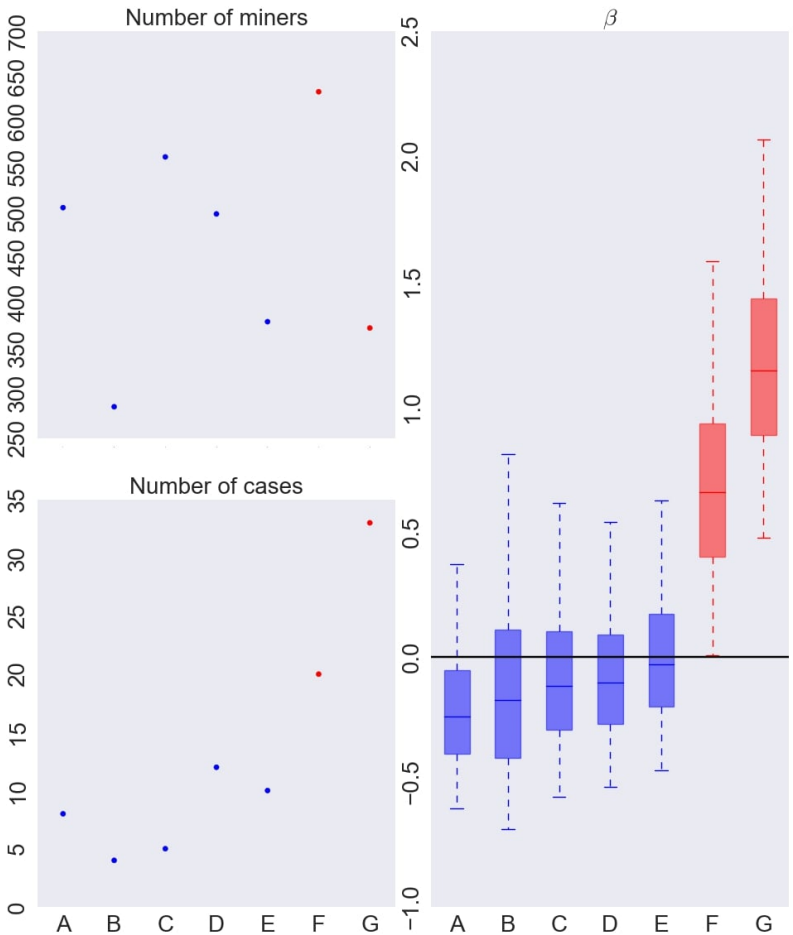
Comparaison des modèles selon le nombre de groupes non-vides

Nombre de groupes non-vides	DIC	WAIC[1]
5	146345	110872
6	136714	108773
7	118602	107004
8	104566	105704

➔ On présente les résultats du modèle à 8 groupes non-vides

[1]Watanabe S. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory.Journal of Machine Learning Research11(2010) 3571–3594.

Identification des groupes - Partition en 8 groupes

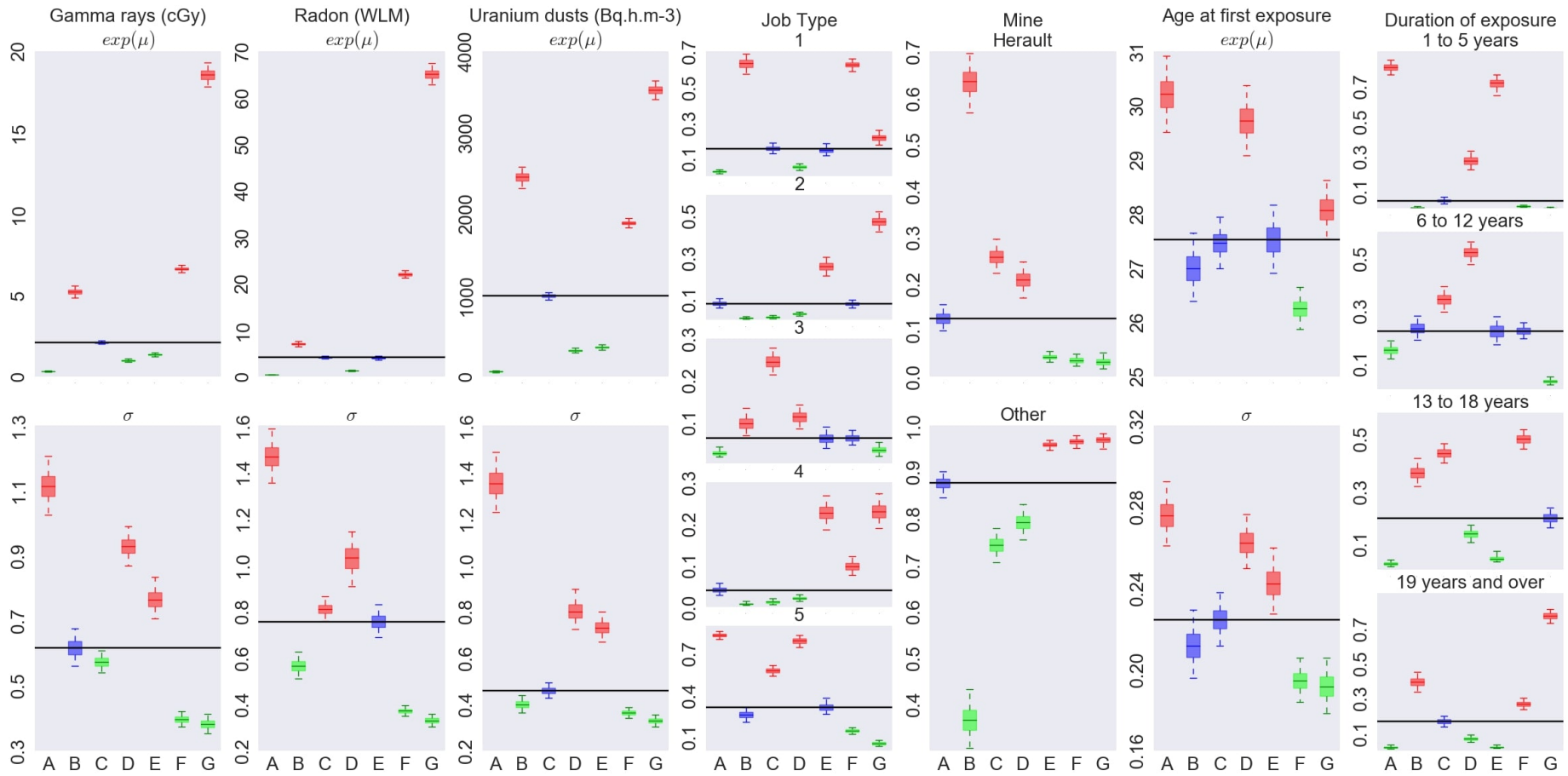


Médianes et Intervalles de crédibilité à 95%

8 groupes dont:

- 2 groupes à risques significativement plus élevés
- 5 groupes à risques non significativement différents de 0
- 1 groupe de référence : les mineurs d'uranium non-exposés

Caractérisation des groupes - Partition en 8 groupes



Job type : 1) foreurs après mécanisation 2) foreurs avant mécanisation 3) autre souterrain après mécanisation 4) autre souterrain avant mécanisation 5) surface

Analyses par simulations: Objectifs

Simulation avec nombre de groupes non-vides fixé:

- ➔ Vérifier l'algorithme MCMC;
- ➔ Discuter la qualité des estimations de risques fournies en fonction de "vraies" valeurs d'excès de risque et des caractéristiques d'exposition;
- ➔ Évaluer la qualité des classifications.

Simulation à nombre de groupes non-vides estimé:

- ➔ Vérifier l'algorithme MCMC;
- ➔ Discuter la qualité des estimations de risques fournies en fonction de "vraies" valeurs d'excès de risque et des caractéristiques d'exposition;
- ➔ Évaluer la qualité des classifications ;
- ➔ Mieux comprendre les difficultés d'inférence rencontrées : Convergence vers un mode local : dans quel(s) cas?

Analyses par simulations: Paramétrages

- 4 groupes (dont le groupe de référence) simulés en proportions égales d'individus
- Covariables : 3 expositions aux rayonnements ionisants (radon, gamma, poussières) + âge à la première exposition
- Valeurs des paramètres $\boldsymbol{\mu} = [\mu_1, \mu_2, \mu_3]$ et $\boldsymbol{\sigma} = [\sigma_1, \sigma_2, \sigma_3]$ des distributions lognormales inspirées des médianes estimées dans les groupes C, F et G obtenus sur données réelles avec nombre de groupes non-vides fixé à 8
- 4 scénarios testés:
  - $\boldsymbol{\mu}$ estimés, $\boldsymbol{\sigma}$ estimés, $\beta = [0, 5, 10]$
  - $\boldsymbol{\mu}$ estimés, $\boldsymbol{\sigma}$ estimés * 0.3, $\beta = [0, 5, 10]$
  - $\boldsymbol{\mu}$ estimés, $\boldsymbol{\sigma}$ estimés, $\beta = [0, 50, 100]$
  - $\boldsymbol{\mu}$ estimés, $\boldsymbol{\sigma}$ estimés * 0.3, $\beta = [0, 2.5, 5]$

Premiers résultats

- 1 jeu de données simulées pour chaque scénario
- Estimation à nombre de groupes non-vides fixé

	Individus bien classés	Paramètres bien estimés[2]
Scénario 1	3142/3233 (97,2%)	22/27 (81.5%)
Scénario 2 (σ réduit)	3377/3377 (100%)	27/27 (100%)
Scénario 3 (β augmenté)	3153/3233 (97.5%)	21/27 (77.8%)
Scénario 4 (σ et β réduits)	3377/3377 (100%)	27/27 (100%)

[2] Intervalle de crédibilité à 95% contient la vraie valeur

Valeurs de β simulées et estimées pour chaque scénario

	Groupe 1	Groupe 2	Groupe 3
Scénario 1	**0**	**5**	**10**
	-0.28[3] [-0.64; 0.29][4]	6.01 [4.41; 8.07]	9.02 [6.77; 11.88]
Scénario 2	**0**	**5**	**10**
(σ réduit)	0.44 [-0.13; 1.26]	5.25 [3.71; 7.11]	9.69 [7.40; 12.50]
Scénario 3	**0**	**50**	**100**
(β augmenté)	*1.30 [0.63; 2.23]*	46.6 [38.23; 56.91]	102.67 [85.06; 124.93]
Scénario 4	**0**	**2.5**	**5**
(σ et β réduits)	0.16 [-0.32; 0.87]	3.12 [1.98; 4.56]	5.91 [4.29; 7.91]

[3] Médiane *a posteriori*

[4] Intervalle de crédibilité à 95%

Conclusions (1/2)

- Extension des modèles BPRM au contexte de données de survie censurées, suivant un modèle de survie en excès de risque instantané
- Première application en épidémiologie des rayonnements ionisants de cette classe de modèles
  - Loi a posteriori multimodale ? Convergence vers des modes locaux de notre algorithme MCMC (si oui, dans quel(s) cas ?) ou problème d'estimation sur le paramètre α ?
    - ➔ Manque de signal dans les données pour mettre en évidence, si elle existe, une partition optimale
    - ➔ Problèmes d'ajustement des modèles de mélange avec des algorithmes MCMC standard : Quid de l'utilisation d'un algorithme MCMC plus avancé basé sur des dynamiques Hamiltoniennes?
  - Estimation à nombre de groupes fixé (ex : 8 groupes): approche prometteuse car interprétation originale et riche de l'association potentielle entre le risque de décès par cancer du poumon et les profils d'exposition radiologiques des mineurs d'uranium français

Conclusions (2/2)

- Analyses par simulations : premiers résultats
  - L'inférence semble globalement correcte sur le jeu de données simulé
  - Impact plus important - sur les estimations de risque - de l'écart entre les groupes en termes de caractéristiques d'exposition (σ) qu'en termes d'excès de risque (β)

Perspectives

- Augmenter le nombre de jeux de données simulés pour chaque scénario
- Faire varier le nombre de covariables d'exposition inclus dans la constitution des groupes ainsi que le niveau d'information apporté
- Étudier l'impact potentiel de la proportion de données censurées sur l'estimation du risque d'intérêt
- Considérer pour modèle d'estimation le modèle BPRM complet (avec estimation du nombre de groupes non-vides) afin, notamment, de mieux comprendre les difficultés d'inférence rencontrées sur les données de la cohorte post-55 des mineurs d'uranium
- Étendre le modèle BPRM afin de prendre en compte les erreurs de mesure sur les expositions

Merci de votre attention!

Lois a priori des poids $\Psi = \{\Psi_c = P(C{=}c)\ c{=}1, ..., C_{max}\}$

- α : $[\alpha] \sim$ Unif[0.3 , 10]
- $V_1, ..., V_{C_{max}-1}$: (C_{max}-1) v.a. iid, $[V_c|\alpha] \sim$ Beta(1,α)

$$\Downarrow$$

$$
\begin{array}{lcl}
\Psi_1 & = & V_1 \\
\Psi_2|\Psi_1 & = & V_2(1-\Psi_1) \\
\Psi_3|\Psi_1,\Psi_2 & = & V_3(1-(\Psi_1+\Psi_2)) \\
... & & \\
\Psi_{C_{max}-1}|\Psi_1,...,\Psi_{C_{max}-2} & = & V_{C_{max}-1}(1-(\Psi_1+...+\Psi_{C_{max}-2}))
\end{array}
$$

$\Rightarrow \Psi_{C_{max}} = 1 - \sum_{k=1}^{C_{max}-1} \Psi_k$ *Processus de Dirichlet*

Label switching moves

La vraisemblance du modèle de mélange est insensible à l'ordre des groupes MAIS la construction stick-breaking l'est :

$$E(\psi_c) > E(\psi_{c+1}), \forall c$$

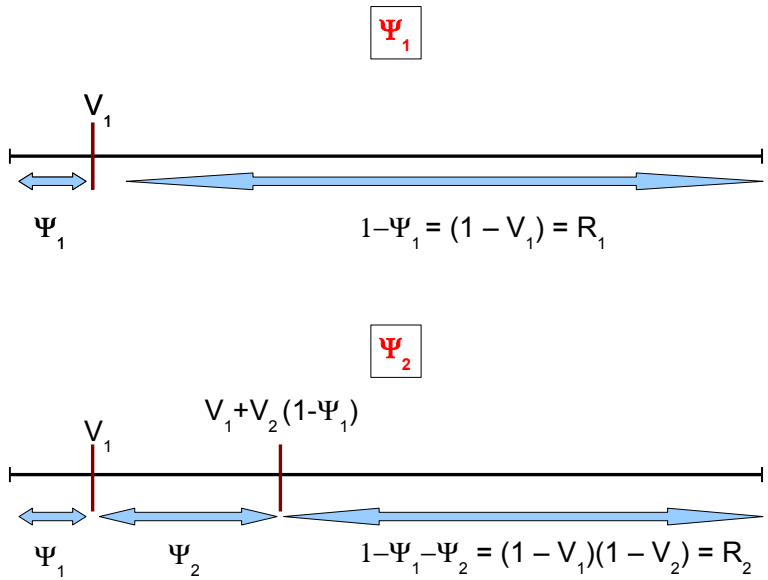
Il faut introduite un changement dans l'ordre des groupes, un "label switching move" afin d'éviter de converger vers un maximum local.

Inférence bayésienne (3/3)

Variable	Algorithme	Proposal
C_i	Gibbs	.
α	RWMH[5]	$\alpha^{cand} \sim N(\alpha^{curr}, \sigma_{proposal})$
V^c	RWMH	$V^c_{cand} \sim N(V^c_{curr}, \sigma_{proposal})$
μ^c_p	RWMH	$\mu^c_{p,cand} \sim N(\mu^c_{p,curr}, \sigma_{proposal})$
σ^c_p	RWMH	$\sigma^c_{p,cand} \sim N(\sigma^c_{p,curr}, \sigma_{proposal})$
p^c_p	Gibbs	.
λ_j	RWMH	$\lambda_{j,cand} \sim N(\lambda_{j,curr}, \sigma_{proposal})$
β^c	RWMH	$\beta^c_{cand} \sim N(\beta_{curr^c}, \sigma_{proposal})$

[5] Random Walk Metropolis-Hastings

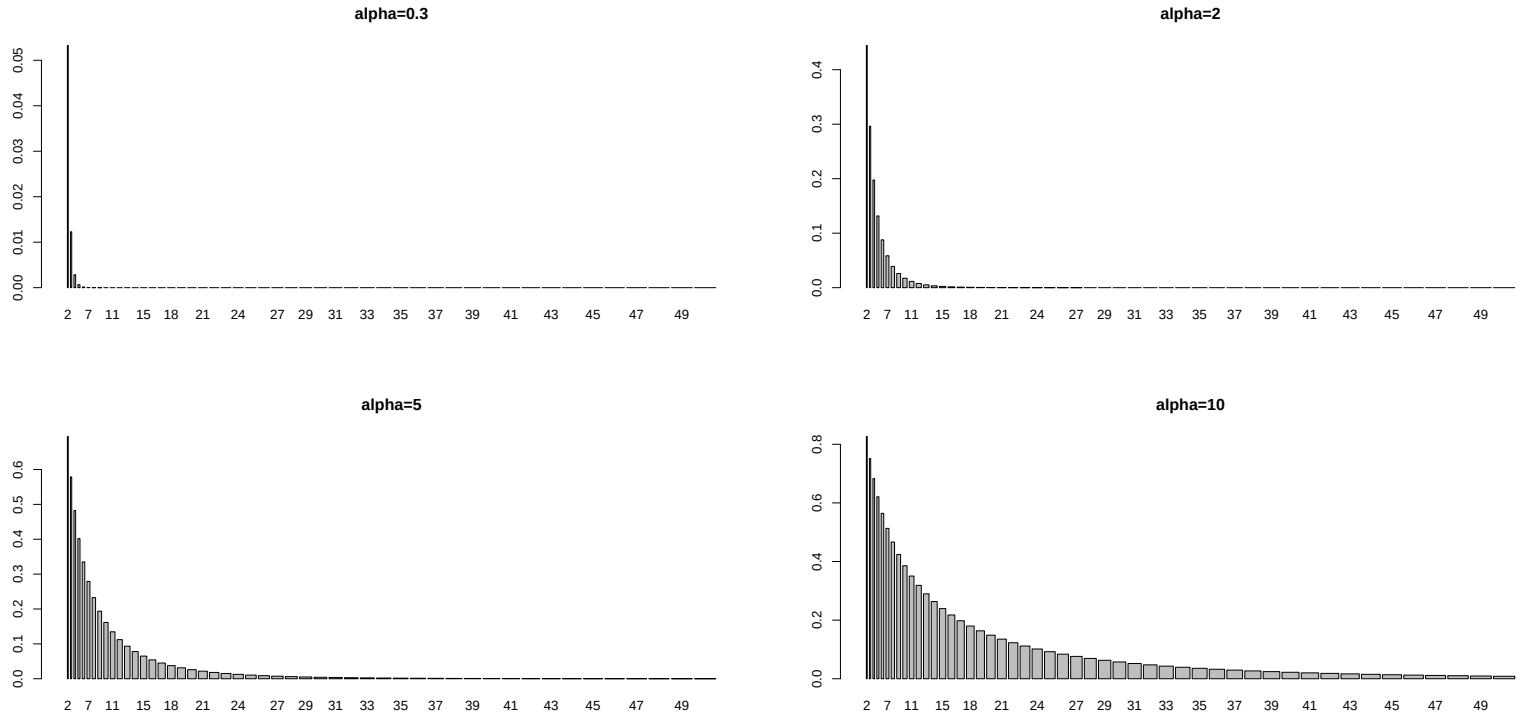
Lois a priori des poids $\Psi = \{\Psi_c = P(C{=}c)\ c{=}1, \ldots, C_{max}\}$



Le "reste" $R_c = \prod_{j=1}^{c}(1 - V_j)$

Comme $V_j \sim$ Beta(1,α) alors $E(R_c) = \{\frac{\alpha}{1+\alpha}\}^c$

$E(R_c)$ pour c=1 à 50



Plus $\alpha \uparrow$, plus un grand nombre de groupes est probable

Choix de C_{max}?

Pas trop grand (calculs inutiles),
Pas trop petit (ne pas oublier des groupes)