Inference in Hidden Markov Models

```
library(knitr)
library(dbplyr)
library(ggplot2)
set.seed(42)
```

On rappelle le modèle des chaînes de markov cachés (HMM) on a des observations $x_1, \dots, x_n$ et une suite de variable latentes $z_1, \dots, z_n$ (états cachés) qui prennent des valeurs discrètes dans $\{1, \dots, K\}$. La loi jointe d'un modèle HMM s'écrit

$$p(x_1, \dots, x_n, z_1, \dots, z_n) = p(z_1) \prod_{i=2}^{n} p(z_i \mid z_{i-1}) \prod_{i=1}^{n} p(x_i \mid z_i) \tag{1}$$

La séquence des variables cachées $z_1, \dots, z_n$ est supposée suivre une *chaîne de markov* (d'où la terminologie HMM) avec:

1. Une loi initiale $z_1 \sim \pi$ telle que $\sum_k \pi_k = 1$.
2. Une matrice stochastique de transition $A \in \mathcal{M}_{K \times K}$, *i.e.* $p(z_{i+1} = l \mid z_i = k) = A_{kl}$.

Les *lois d'émissions des observations* $\Psi_i(k) = p(x_i \mid z_i = k)$ sont paramétriques, avec des paramètres θ, et peuvent être continues ou discrètes

- **discrète** $p(x_i = j \mid z_i = k) = B_{kj}$, $\theta = B$
- **continues**, par exemple en Gaussien $p(x_i \mid z_i = k) = \mathcal{N}(x_i \mid \mu_k, \sigma_k^2)$, $\theta = (\mu_k, \sigma_k^2)_k$

On notera donc un modèle HMM par le triplet (π, A, θ) qui le définit entièrement. On s'intéresse à trois problèmes fondamenteux liés aux HMMs:

1. Le calcul de la vraisemblance des observations $p(x_1, \dots, x_n \mid \pi, A, \theta)$.
2. Le calcul du maximum à posteriori ou *décodage* : $\text{argmax}_{z_1, \dots, z_n} p(z_1, \dots, zn \mid x_{1:n}; \pi, A, \theta)$.
3. L'estimation des paramètres par algorithme EM.

Les paramètres π, A et θ du modèles sont connus

C'est un cas idéalisé, mais c'est une première étape intéressante en soit, et nécessaire avant de parler d'estimation. On suppose qu'on observe une séquence $x_{1:n}$ du HMM.

Simulation d'un HMM

Coder une fonction `SimuHMMdisc(n, param)`qui simule une séquence (x_i, z_i) de taille n issue d'un HMM avec loi d'émission discrète paramétrée par une matrice B.

La tester avec le modèle suivant et faire un plot avec i en abscisse et x_i en ordonnée avec la couleur du point données par z_i.

```
n = 1e3
# Chaîne de Markov sur les états cachés
A <- matrix(c(0.95,0.1,0.05,0.9),2,2)

# Loi d'émission
```

```r
B <- t(matrix(c(rep(1/6,6),rep(1/10,5),5/10),6,2))

param = list()
param$pi = c()
param$A = A
param$pi = Re(eigen(t(param$A))$vector[,1]) # take the real-part
param$pi = param$pi / sum(param$pi)

param$B = B
```

Calcul de la vraisemblance : l'algorithme forward

On souhaite calculer $p(x_1, \dots, x_n \mid \pi, A, \theta)$. Comme vu en cours, on peut la réécrire comme une marginalisation de la loi jointe sur les variables cachées (non-observées). On utilise ensuite la forme particulière de la loi jointe dans un modèle HMM données dans la première équation de ce document.

$$p(x_{1:n}) = \sum_{z_{1:n}} p(x_{1:n}, z_{1:n}), \tag{2}$$

$$= \sum_{z_1} \dots \sum_{z_n} p(z_1) \prod_{i=2}^{n} p(z_i \mid z_{i-1}) \prod_{i=1}^{n} p(x_i \mid z_i) \tag{3}$$

Cette méthode "brute" nécessite de sommer sur n variables latentes pouvant prendre K valeurs chacunes, occasionant K^n opérations ! Cela deviendra rapidement infaisable pour des valeurs "raisonnables" de K et n.

Décomposition en sous-problèmes

On peut réécrire la vraisemblance comme la marginalisation sur la dernière variable z_n

$$p(x_{1:n}) = \sum_{k=1}^{K} p(x_{1:n}, z_n = k) = \sum_{k=1}^{K} \alpha_n(k) \tag{4}$$

On pose plus généralement pour une observation $i = 1, \dots, n$ et un état $k = 1, \dots, K$ le message:

$$\alpha_i(k) = p(x_{1:i}, z_i = k) \tag{5}$$

L'intérêt de cette réécriture est la récursion suivante qui lie $\alpha_i(k)$ à α_{i-1} :

$$\alpha_i(k) = p(x_{1:i}, z_i = k), \tag{6}$$
$$= \sum_{l=1}^{K} p(x_{1:i}, z_{i-1} = l, z_i = k), \tag{7}$$
$$= \sum_{l=1}^{K} p(x_{1:i-1}, x_i, z_{i-1} = l, z_i = k), \tag{8}$$
$$= \sum_{l=1}^{K} p(x_i, z_i = k \mid x_{1:i-1}, z_{i-1} = l) p(x_{1:i-1}, z_{i-1} = l), \tag{9}$$
$$= \sum_{l=1}^{K} p(x_i \mid z_i = k, x_{1:i-1}, z_{i-1} = l) p(z_i = k \mid x_{1:i-1}, z_{i-1} = l) p(x_{1:i-1}, z_{i-1} = l), \tag{10}$$
$$= \sum_{l=1}^{K} p(x_i \mid z_i = k) p(z_i = k \mid z_{i-1} = l) p(x_{1:i-1}, z_{i-1} = l), \quad \text{(HMM model)} \tag{11}$$
$$= \sum_{l=1}^{K} \Psi_i(k) A_{lk} \alpha_{i-1}(l), \tag{12}$$
$$\tag{13}$$

Note: Les calculs jusqu'à l'avant dernière ligne sont toujours valables, sans aucune hypothèse sur le modèle. A l'avant dernière ligne on utilise le fait que dans un HMM, connaitre (conditionner par rapport à) z_{i-1} rend les x_i et z_i indépendant du passé.

Note2: On initialise la récursion avec $\alpha_1(l) = p(x_1, z_1 = l) = \pi_l \Psi_i(l), \forall l$.

En pratique, les $\alpha_i(k)$ sont des probas qui peuvent devenir très petites et on privilégie plutôt un codage en log, la récursion s'écrit alors :

$$\log \alpha_i(k) = \log \left(\sum_{l=1}^{K} \exp\{\log A_{lk} + \log \Psi_i(k) + \log \alpha_{i-1}(l)\} \right). \tag{14}$$

Un aparté: le "log-sum-exp" trick **R** n'a pas une précision de calcul infinie, par exemple

```
cat("En R, exp(-1000) = ", exp(-1000), '\n')
```

```
## En R, exp(-1000) =  0
```

```
cat("Le test d'égalité exp(-1000) == 0 renvoie", exp(-1000) == 0, '\n')
```

```
## Le test d'égalité exp(-1000) == 0 renvoie TRUE
```

```
cat("Le passage au log renvoie donc log(exp(-1000)) =", log(exp(-1000)), '\n')
```

```
## Le passage au log renvoie donc log(exp(-1000)) = -Inf
```

Pourtant, on sait que $\log(e^{-1000}) = -1000$!

Dans la récursion ci-dessus, on va calculer à chaque itération des quantités de type `log(sum(exp(a)))`. Si à l'itération $i-1$, les $\log \alpha_{i-1}(k)$ sont très négatifs (*i.e* les $\alpha_{i-1}(k)$ sont petits), alors à l'itération i il y a un risque de propagation des *underflow* numériques. En effet, on va calculer les $\sum_k \exp(\log \alpha_{i-1}(k) + y_{ik})$ où y_{ik} est lui-même négatif (log de probas). Chaque terme de la somme peut finir par dépasser la résolution numérique de **R**. Alors on aura la somme qui vaudra 0 et le log de la somme qui vaudra `-Inf`. Cela se propagera ensuite à toute les itération suivantes avec des `log(sum(exp(-Inf, ..., -Inf))) = -Inf`.

Il existe une façon de se prémunir de cette propagation. L'idée est d'utiliser l'identité suivant qui est vraie pour tout $m \in \mathbb{R}$

$$\log(\sum_k e^{a_k}) = m + \log(\sum_k e^{a_k - m}),$$

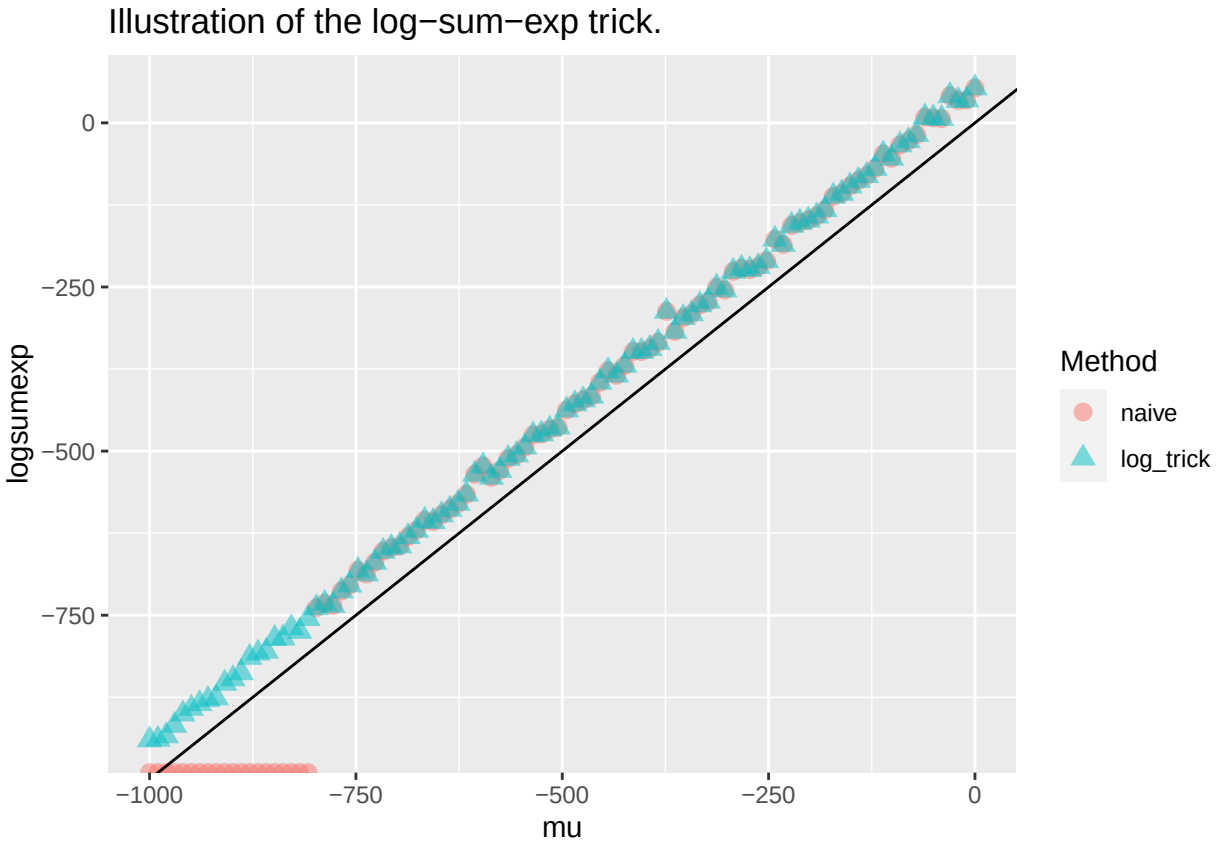
avec $m := \max_k a_k$. De cette manière, on s'assure de toujours avoir un $e^0 = 1$ quelque par dans la somme à l'intérieur du log.

```
logsumexp <- function(logx) {
  # compute \log(\sum exp(logx)) by rescaling it by m = \max(logx)
  # indeed : \log(\sum exp(logx)) = m + \log(\sum exp(logx - m))
  # This ensures an exp(0) somewhere in the sum
  m = max(logx)
  return(m + log(sum(exp(logx - m))))
}
```

Pour illustrer l'intérêt de cette méthode, on peut générer des points $a_1, \ldots, a_K$ selon une distribution $\mathcal{N}(\mu_e, 20)$, avec $\mu_1 = 0 < \ldots < \mu_E = -1000$. Au bout d'un certain seuil, les e^{a_k} vont underflow et l'intérêt du log-sum-exp trick apparait.

```
K = 200
nexp = 100 # taille de la sequence des $\mu_e$
mus = seq(0, -1000, length.out=nexp)
res = data.frame(naive=NA, log_trick=NA, mu = mus)
for (i in 1:length(mus)){
  a = rnorm(K, mean = mus[i], sd = 20)
  res$naive[i] = log(sum(exp(a)))
  res$log_trick[i] = logsumexp(a)
}

library(reshape2)
df = reshape2::melt(res, id='mu', variable.name = "Method", value.name = "logsumexp")
gg = ggplot(df) +
    geom_point(aes(x=mu, y=logsumexp, color=Method, shape=Method), size=3, alpha=0.5) +
    geom_abline(slope=1, intercept = 0) + # add y=x line
    ggtitle("Illustration of the log-sum-exp trick.")
print(gg)
```



Sur le graphique, on voit clairement que la méthode log trick donne toujours un nombre flottant tandis que la méthode naïve underflow un peu après $\mu_e < -750$.

```
tail(res) %>% kable()
```

	naive	log_trick	mu
95	-Inf	-891.9045	-949.4949
96	-Inf	-900.6153	-959.5960
97	-Inf	-917.8089	-969.6970
98	-Inf	-933.8207	-979.7980
99	-Inf	-939.0995	-989.8990
100	-Inf	-940.0257	-1000.0000

Bonus (en exercice) On remarque également que le résultat semble toujours au dessus de la droite $y = x$. Montrer que c'est vrai en espérance, *i.e.* $\mathbb{E}_{\mathbf{a}}\left[\log(\sum_k \exp(a_k))\right] > \mu_e$. (Une inégalité de Jensen pourrait s'avérer utile ici).

Questions Dans le cas d'une loi d'émission discrète avec les paramètres `param` défini au debut. Coder

1. une fonction `forward(x, i, param)` qui prend en entrée les observations x, un temps i et les paramètres π, A, θ, et calcule le vecteur des $\log \alpha_i$.

L'illustrer dans le cadre suivant. Que remarquez vous entre les valeurs de $\log \alpha_{10}$ et le vrai z_{10} (accessible ici puisqu'on a simulé) ?

```
logalpha_10 = forward(hmmsim$X, i=10, param)
print(logalpha_10)
## [1] -18.53122 -19.77959
```

```
print(hmmsim$Z[10])
## [1] 2
```

2. Une fonction `HMMloglik(x, param)` qui calcule la log-vraisemblance de `x` sous un modèle HMM discret avec paramètres `param`. On rappelle :

$$\log p(x_{1:n}) = \log \sum_{k=1}^{K} p(x_{1:n}, z_n = k) = \log \sum_{k=1}^{K} \exp(\log \alpha_n(k))$$

Decoding : the Viterbi algorithm

Note de cours

Le décodage consiste à trouver la séquence de variables cachées $\hat{z}_{1:n}$ la plus probable étant donnés les paramètres et les observations $x_{1:n}$.

$$\hat{z}_{1:n} = \arg\max_{z_{1:n}} \log p(z_{1:n} \mid x_{1:n}; \theta, \pi, A)$$

C'est donc du maximum a posteriori ! Comme l'espace de recherche des $z_{1:n}$ est discret, une énumération serait possible mais trop coûteuse : K^n possibilités ! Encore une fois, une décomposition en sous problème simple va s'avérer utile. L'algorithme de Vitberbi permet de résoudre le problème de décodage à l'aide de manière séquentielle. Si on connaissait la meilleure séquence $\hat{z}_{1:(n-1)}$ on pourrait facilement résoudre $\arg\max_{z_n} p(\hat{z}_{1:(n-1)}, z_n, x_{1:n} | \theta, \pi, A)$.

Définissons $V_i(k) = \max_{z_{1:(i-1)}} p(x_{1:i}, z_i = k, z_{1:(i-1)} \mid \theta, \pi, A)$, alors on peut montrer que V_i vérifie la récursion suivante :

$$\begin{aligned} V_1(k) &= \pi_1 \Psi_i(k), \\ V_i(k) &= \Psi_i(k) \cdot \max_{l=1,\dots,K} V_{i-1}(l) A_{lk}, \end{aligned} \tag{15}$$

On calcule les $V_i(k)$ pour tout les $i = 1, \dots, n$, et on trouve $\hat{z}_n$ grace à $\hat{z}_n = \arg\max_k V_n(k)$. On peut ensuite reconstruire la séquence $\hat{z}_{n-1}$ jusqu'à $\hat{z}_1$ de proche en proche. Pour retrouver $\hat{z}_{i-1}$ on prend l'état qui a donné la transition $\hat{z}_{i-1} \to \hat{z}_i$ la plus probable (conditionnellement aux observations $x_{1:n}$). C'est-à-dire l'argmax sur les transitions possibles : $S_i(k) = \arg\max_{l=1,\dots,K} V_{i-1}(l) A_{lk}$

$$\hat{z}_{i-1} := S_i(\hat{z}_i)$$

> **Note** : ici, l'opérateur max joue un rôle analogue à l'opérateur $\sum$ pour la marginalisation sur les variables z. En réalité, l'algorithme de Viterbi est un cas particulier de l'algorithme "max-product" pour calculer les modes qui est exacte pour les arbres. Plus d'information ici par exemple

Questions

1. Montrer que le problème de maximisation de Viterbi peut se réécrire en fonction de la loi jointe

$$\arg\max_{z_{1:n}} \log p(z_{1:n} \mid x_{1:n}; \theta, \pi, A) = \arg\max_{z_{1:n}} \log p(z_{1:n}, x_{1:n} \mid \theta, \pi, A)$$

2. Montrer que dans un modèle HMM on a bien l'identité :

$$\max_{z_{1:(i-1)}} p(x_{1:i}, z_i = k, z_{1:(i-1)} \mid \theta, \pi, A) = \Psi_i(k) \cdot \max_{l=1,\dots,K} V_{i-1}(l) A_{lk}$$

3. Programmer une fonction `Viterbi(x, param)` qui renvoie l'estimateur du MAP $\hat{z}_{1:n}$.

Astuce: le log est une fonction croissante donc $\log(\max f(x)) = \max \log(f(x))$, on peut (et il faut) donc coder les quantités précédentes en log-space.

4. Tester votre fonction sur le code suivant

```
z_hat = Viterbi(hmmsim$X, param)
n_wrong = sum(z_hat != hmmsim$Z)
cat('Proportion of wrong states : ', n_wrong / length(hmmsim$X))
```

```
## Proportion of wrong states :  0.204
```

Estimation : l'algorithme de Baum-Welch (EM pour les HMMs)

Jusqu'ici on a supposé les paramètres $\eta = (\pi, A, \theta)$ connus, ce qui n'est pas très réaliste en pratique. Dans cette dernière partie on s'intéresse à l'estimation des paramètres du modèle à l'aide des observations $x_1, \dots, x_n$.

Etant donné que les HMMs sont des modèles à variables latentes, on peut utiliser l'algorithme EM, qui s'appelle historiquement l'algorithme de *Baum-Welch*.

- Initialisation de $\eta^{(0)} = (\pi^{(0)}, A^{(0)}, \theta^{(0)})$
- Pour $t = 0, \dots, T$ itérer entre
  - **E-step** (lissage ou *smoothing*) Calcul de borne inférieure

$$\begin{aligned} Q_t(\eta) &= E_{z_{1:n} \sim p(\cdot \mid x_{1:n}, \eta^{(t)})}[\log p(x, z \mid \eta)], \\ &= \sum_k \tau_1(k) \log \pi_k + \sum_{i=1}^{n-1} \sum_{k,l=1}^{K} \xi_{i,i+1}(k,l) \log A_{k,l} + \sum_{i=1}^{n} \tau_i(k) \log \Psi_i(k) \end{aligned} \tag{16}$$

  - Cela revient à calculer les quantités
    1. $\tau_i(k) = p(z_i = k \mid x_{1:n}; \eta^{(t)})$,
    2. $\xi_{i,i+1}(k,l) = p(z_i = k, z_{i+1} = l \mid x_{1:n}; \eta^{(t)})$

    à l'aide de l'algorithme forward-backward.
  - **M-step** Maximisation de la borne en η

$$\eta^{(t+1)} := \arg\max_{\pi, A, \theta} Q_t(\pi, A, \theta)$$

    Cette étape dépend de la loi des émissions !
  - STOP si $l(\eta^{(t+1)}) - l(\eta^{(t)}) < \epsilon$

Algorithme forward-backward pour le lissage (E-step)

L'algorithme forward-backward permet le calcul des lois marginales à posteriori des $z_i \mid x_{1:n}$ et des jointes $(z_i, z_{i+1}) \mid x_{1:n}$. Il procède en deux phases qui consistent à parcourir la chaîne dans le sens temporel (*forward*), puis dans le sens inverse (*backward*).

L'idée principale est que l'on peut séparer la chaîne d'un HMM en 2 parties (passé & futur) conditionellement à z_i :

$$p(z_i = k \mid x_{1:n}) \propto p(z_i = k, x_{1:i})\, p(x_{(i+1):n} \mid z_i = k) = \frac{\alpha_i(k)\beta_i(k)}{p(x_{1:n})}.$$

On définit donc les 2 quantités

- **Forward** $\alpha_i(k) = p(z_i = k, z_{1:i}) = \Psi_i(k) \sum_l A_{lk} \alpha_{i-1}(l)$ (cf. première partie)
- **Backward** $\beta_i(k) := p(x_{(i+1):n} \mid z_i = k) = \sum_{l=1}^{K} A_{kl} \Psi_{i+1}(l) \beta_{i+1}(l)$ (cf. slides de cours) avec l'initialisation $\beta_n(k) = 1, \forall k$.

On pourra calculer les $\tau_i(k)$ et $\xi_{i,i+1}(k,l)$ grâce aux formules

$$\begin{aligned} \tau_i(k) &\propto \alpha_i(k)\beta_i(k), \\ \xi_{i,i+1}(k,l) &\propto \alpha_i(k)\Psi_{i+1}\beta_{i+1}(l)A_{kl}. \end{aligned} \tag{17}$$

Note On va encore coder en log-space et pouvoir utiliser le log-sum-exp pour assurer la stabilité numérique au fil de la récursion !

Questions

0. **(facultatif)** Vérifier la formule de récursion pour $\beta_i(k)$ et les formule liant τ_i et $\xi_{i,i+1}$ à α et β (cf. slides).

Nous avons déjà implémenté la partie forward de l'algorithme qui calcule les $\log \alpha_i(k)$ pour un i donné. On voit qu'on aura maintenant besoin des $\log \alpha_i$ pour tout i.

1. Modifier `forward(x, i, param)` en une fonction `forward_estep(x, param)` qui retourne la matrice $K \times n$ des $(\log \alpha_i(k))_{k,i}$.
2. Implémenter `backward(x, param)` qui renvoie la matrice $K \times n$ des $(\log \beta_i(k))_{k,i}$.
3. Implémenter une fonction `estep(x, param)` qui fait le lissage et renvoie les $\tau_i(k)$ (matrice $K \times V$ et $\xi_{i,i+1}(k,l)$ (tenseur $K \times K \times (n-1)$). Renvoyer la log-vraisemblance du modèle également.
4. Programmer une fonction `mstep(x, smoothed_post)` qui calcule $\eta^{(t+1)}$ dans un HMM à émission discrète.
5. Programmer une fonction `baum_welch(x, init_param, max.it, atol)` qui retourne un estimateur $\hat{\eta}$ et le vecteur stockant l'évolution de la log-vraisemblance dans l'algorithme.

Pour aller plus loin

Refaire le TP avec des lois d'émissions Gaussiennes (variance $\sigma^2 = 1$ connu et fixée, on n'estime que les moyennes). Que faudra-t-il modifier finalement ?