Μίξεις Κατανομών και Κρυμμένα Μαρκοβιανά Μοντέλα

Βασίλης Κατσιάνος

Αύγουστος 2023

Περιεχόμενα

1 Πεπερασμένες Μίξεις Κατανομών

Έστω δείγμα ανεξάρτητων παρατηρήσεων $y_1, \dots, y_n$ από μίξη κατανομών με συνάρτηση (πυκνότητας) πιθανότητας:

$$f_\vartheta(y_i) = \sum_{k=1}^{K} p_k f_{\theta_k}(y_i).$$

Η παρατηρούμενη πιθανοφάνεια των δεδομένων για $\vartheta = (p_1, \dots, p_K, \theta_1, \dots, \theta_K)$ γράφεται ως:

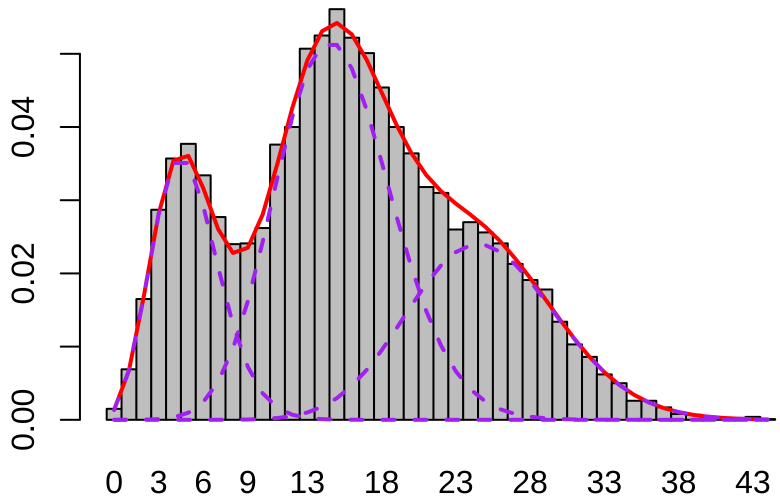
$$\mathcal{L}(\vartheta \mid y) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} p_k f_{\theta_k}(y_i) \right],$$

οπότε είναι αδύνατο να μεγιστοποιηθεί αναλυτικά εξαιτίας του αθροίσματος.

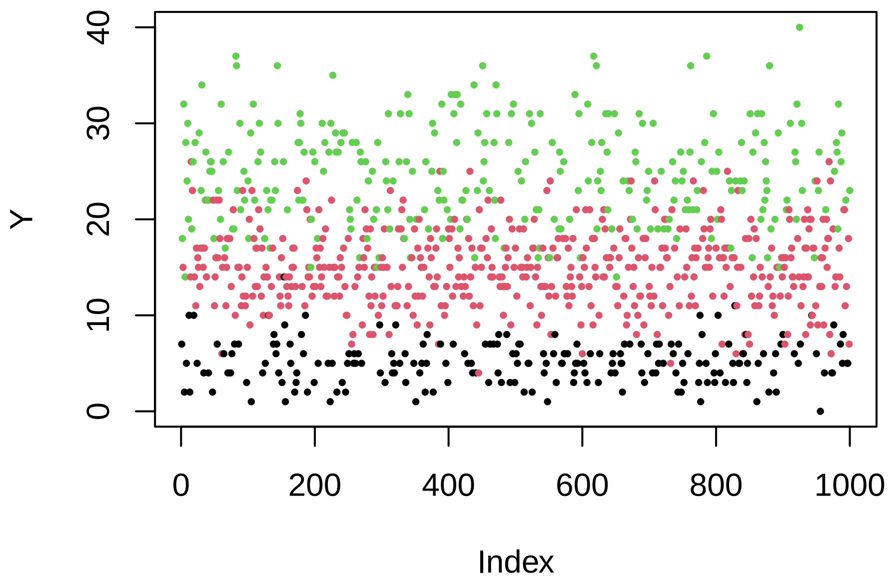
Εισάγουμε τις λανθάνουσες μεταβλητές $X_1, \dots, X_n$ με $\mathbb{P}_p(X_i = k) = p_k$. Τότε, $(Y_i \mid X_i = k) \sim f_{\theta_k}(\cdot)$. Η πλήρης πιθανοφάνεια, δηλαδή η από κοινού πιθανοφάνεια των παρατηρούμενων μεταβλητών y_i και των λανθανουσών μεταβλητών x_i, γράφεται ως:

$$\begin{aligned}\mathcal{L}(\vartheta \mid y, x) &= \prod_{i=1}^{n} f_{\vartheta}(y_i, x_i) = \prod_{i=1}^{n} \left[\mathbb{P}_p(X_i = x_i) f_{\theta}(y_i \mid X_i = x_i) \right] \\ &= \prod_{i=1}^{n} \left[p_{x_i} f_{\theta_{x_i}}(y_i) \right] = \prod_{i=1}^{n} \prod_{k=1}^{K} \left[p_k f_{\theta_k}(y_i) \right]^{\mathbb{1}\{x_i = k\}}.\end{aligned}$$

```r
rpoismix = function(n, theta, p) {
    K = length(theta)
    X = sample(K, n, replace = TRUE, p)
    Y = rpois(n, theta[X])
    return(list(Y = Y, X = X))
}

n = 10000
K = 3
theta = c(5, 15, 25)
p = c(0.2, 0.5, 0.3)
mix = rpoismix(n, theta, p)
Y = mix$Y
X = mix$X
barplot(table(factor(Y, levels = 0:max(Y)))/n, space = 0)
lines(0:max(Y) + 0.5, p[1] * dpois(0:max(Y), theta[1]) + p[2] * dpois(0:max(Y),
    theta[2]) + p[3] * dpois(0:max(Y), theta[3]), col = "red", lwd = 2)
lines(0:max(Y) + 0.5, p[1] * dpois(0:max(Y), theta[1]), col = "purple", lty = 2,
    lwd = 2)
lines(0:max(Y) + 0.5, p[2] * dpois(0:max(Y), theta[2]), col = "purple", lty = 2,
    lwd = 2)
lines(0:max(Y) + 0.5, p[3] * dpois(0:max(Y), theta[3]), col = "purple", lty = 2,
    lwd = 2)
```




```
n = 1000
mix = rpoismix(n, theta, p)
Y = mix$Y
X = mix$X
plot(Y, pch = 16, col = X, cex = 0.5)
```



2 Αλγόριθμος Expectation-Maximization (EM)

Αντί να μεγιστοποιήσουμε άμεσα την παρατηρούμενη πιθανοφάνεια, εργαζόμαστε επαναληπτικά. Ξεκινάμε με κάποια αρχική εκτίμηση $\vartheta^{(0)}$. Υπολογίζουμε τη δεσμευμένη μέση τιμή της πλήρους λογαριθμο-πιθανοφάνειας $\ell(\vartheta \mid y, x) = \log \mathcal{L}(\vartheta \mid y, x)$ δεδομένων των παρατηρούμενων μεταβλητών κάτω από την παράμετρο $\vartheta^{(0)}$, δηλαδή τη συνάρτηση:

$$Q_{\vartheta^{(0)}}(\vartheta) = \mathbb{E}_{\vartheta^{(0)}}\left[\ell(\vartheta \mid y, X) \mid y\right].$$

Αυτή η συνάρτηση ονομάζεται ενδιάμεση ποσότητα του αλγορίθμου EM. Αυτό το βήμα του αλγορίθμου EM ονομάζεται Expectation step (E-step). Έπειτα, μεγιστοποιούμε τη συνάρτηση $Q_{\vartheta^{(0)}}(\vartheta)$ ως προς ϑ και παίρνουμε μία νέα εκτίμηση $\vartheta^{(1)}$. Αυτό το βήμα του αλγορίθμου EM ονομάζεται Maximization step (M-step). Επαναλαμβάνουμε αυτά τα δύο βήματα μέχρι ο αλγόριθμος να συγκλίνει σε κάποιο ϑ^*.

Θεώρημα 1. Αν η παρατηρούμενη πιθανοφάνεια $\mathcal{L}(\vartheta \mid y)$ είναι φραγμένη, τότε η τιμή ϑ^* στην οποία συγκλίνει ο αλγόριθμος EM είναι κάποιο τοπικό μέγιστό της.

Απόδειξη. Αρχικά παρατηρούμε ότι:

$$\mathcal{L}(\vartheta \mid y, x) = f_\vartheta(x, y) = f_\vartheta(y) f_\vartheta(x \mid y) = \mathcal{L}(\vartheta \mid y) f_\vartheta(x \mid y).$$

Λογαριθμίζουμε την παραπάνω ισότητα:

$$\ell(\vartheta \mid y, x) = \ell(\vartheta \mid y) + \log f_\vartheta(x \mid y).$$

Αν η X είναι απόλυτα συνεχής τυχαία μεταβλητή, πολλαπλασιάζουμε και τα δύο μέλη της ισότητας με την πυκνότητα $f_{\vartheta^{(0)}}(x \mid y)$ και ολοκληρώνουμε ως προς x:

$$\int \ell(\vartheta \mid y, x) f_{\vartheta^{(0)}}(x \mid y) dx = \int \ell(\vartheta \mid y) f_{\vartheta^{(0)}}(x \mid y) dx + \int \log f_\vartheta(x \mid y) f_{\vartheta^{(0)}}(x \mid y) dx.$$

Αντίστοιχα θα μπορούσαμε να πολλαπλασιάσουμε με την πιθανότητα $\mathbb{P}_{\vartheta^{(0)}}(X = x \mid y)$ και να αθροίσουμε ως προς x, αν η X ήταν διακριτή τυχαία μεταβλητή. Παρατηρούμε ότι:

$$\int \ell(\vartheta \mid y, x) f_{\vartheta^{(0)}}(x \mid y) dx = \mathbb{E}_{\vartheta^{(0)}}\left[\ell(\vartheta \mid y, X) \mid y\right] = Q_{\vartheta^{(0)}}(\vartheta),$$

$$\int \ell(\vartheta \mid y) f_{\vartheta^{(0)}}(x \mid y) dx = \ell(\vartheta \mid y) \int f_{\vartheta^{(0)}}(x \mid y) dx = \ell(\vartheta \mid y).$$

Επιπλέον, ορίζουμε:

$$\mathcal{H}_{\vartheta^{(0)}}(\vartheta) = -\int \log f_\vartheta(x \mid y) f_{\vartheta^{(0)}}(x \mid y) dx = -\mathbb{E}_{\vartheta^{(0)}}\left[\log f_\vartheta(X \mid y) \mid y\right].$$

Επομένως, έχουμε καταφέρει να αναλύσουμε την παρατηρούμενη πιθανοφάνεια σε δύο κομμάτια:

$$\ell(\vartheta \mid y) = Q_{\vartheta^{(0)}}(\vartheta) + \mathcal{H}_{\vartheta^{(0)}}(\vartheta).$$

Χρησιμοποιώντας την ανισότητα Jensen, μπορούμε να δείξουμε ότι:

$$\begin{aligned}
\mathcal{H}_{\vartheta^{(0)}}\left(\vartheta^{(1)}\right) - \mathcal{H}_{\vartheta^{(0)}}\left(\vartheta^{(0)}\right) &= -\mathbb{E}_{\vartheta^{(0)}}\left[\log f_{\vartheta^{(1)}}(X \mid y) \mid y\right] + \mathbb{E}_{\vartheta^{(0)}}\left[\log f_{\vartheta^{(0)}}(X \mid y) \mid y\right] \\
&= -\mathbb{E}_{\vartheta^{(0)}}\left[\log \frac{f_{\vartheta^{(1)}}(X \mid y)}{f_{\vartheta^{(0)}}(X \mid y)} \middle| y\right] \geqslant -\log \mathbb{E}_{\vartheta^{(0)}}\left[\frac{f_{\vartheta^{(1)}}(X \mid y)}{f_{\vartheta^{(0)}}(X \mid y)} \middle| y\right] \\
&= -\log \int \frac{f_{\vartheta^{(1)}}(x \mid y)}{\cancel{f_{\vartheta^{(0)}}(x \mid y)}} \cancel{f_{\vartheta^{(0)}}(x \mid y)} dx = -\log \int f_{\vartheta^{(1)}}(x \mid y) dx = -\log 1 = 0.
\end{aligned}$$

Η ανισότητα αυτή είναι γνωστή ως η θεμελιώδης ανισότητα του αλγορίθμου EM. Αν $\vartheta^{(0)}$ είναι η τρέχουσα εκτίμησή μας για το ϑ, τότε αυτή η ανισότητα δείχνει ότι οποιαδήποτε κι αν είναι η επόμενη εκτίμηση $\vartheta^{(1)}$, η συνάρτηση $\mathcal{H}_{\vartheta^{(0)}}(\cdot)$ σίγουρα δε θα πέσει κάτω από την τρέχουσα τιμή $\mathcal{H}_{\vartheta^{(0)}}\left(\vartheta^{(0)}\right)$. Εφόσον η συνάρτηση $\mathcal{H}$ αυξάνεται εγγυημένα σε κάθε βήμα του αλγορίθμου EM, μπορούμε να την αγνοήσουμε τελείως και να επικεντρωθούμε στη συνάρτηση Q.

Αν επιλέξουμε οποιαδήποτε τιμή $\vartheta^{(1)}$ που αυξάνει την τιμή της $Q_{\vartheta^{(0)}}(\cdot)$, δηλαδή $Q_{\vartheta^{(0)}}\left(\vartheta^{(1)}\right) > Q_{\vartheta^{(0)}}\left(\vartheta^{(0)}\right)$, τότε θα έχουμε επιτύχει $\ell\left(\vartheta^{(1)}\middle| y,x\right) > \ell\left(\vartheta^{(0)}\middle| y,x\right)$. Επαναλαμβάνοντας αυτή τη διαδικασία, παράγουμε μία ακολουθία εκτιμήσεων η οποία αυξάνει την τιμή της παρατηρούμενης πιθανοφάνειας σε κάθε βήμα του αλγορίθμου και τελικά συγκλίνει σε κάποιο τοπικό μέγιστο.

Προφανώς, αν επιλέξουμε ως $\vartheta^{(1)}$ ακριβώς την τιμή η οποία μεγιστοποιεί τη συνάρτηση $Q_{\vartheta^{(0)}}(\cdot)$, δηλαδή $\vartheta^{(1)} = \operatorname{argmax}_{\vartheta} Q_{\vartheta^{(0)}}(\vartheta)$, τότε ο αλγόριθμος θα έχει τη μέγιστη δυνατή ταχύτητα σύγκλισης. Αυτός ακριβώς είναι ο στόχος του αλγορίθμου EM. Ωστόσο, ακόμα κι αν η αναλυτική μεγιστοποίηση της συνάρτησης $Q_{\vartheta^{(0)}}(\cdot)$ δεν είναι δυνατή, ο αλγόριθμος θα συγκλίνει εγγυημένα σε κάποιο τοπικό μέγιστο της παρατηρούμενης πιθανοφάνειας, αρκεί να επιλέγουμε σε κάθε βήμα μία νέα εκτίμηση η οποία θα αυξάνει έστω και λίγο την τρέχουσα τιμή της συνάρτησης Q. Για τον λόγο αυτό, έχουν αναπτυχθεί διάφορες παραλλαγές του κλασικού αλγορίθμου EM.

3 Εφαρμογή του Αλγορίθμου EM σε Μίξεις Κατανομών

Αρχικά, λογαριθμίζουμε την πλήρη πιθανοφάνεια της μίξης:

$$\ell(\vartheta \mid y, x) = \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{1}\{x_i = k\} \left[\log p_k + \log f_{\theta_k}(y_i)\right].$$

Έπειτα, υπολογίζουμε την ενδιάμεση ποσότητα του EM:

$$\begin{aligned} Q_{\vartheta^{(0)}}(\vartheta) = \mathbb{E}_{\vartheta^{(0)}}\left[\ell(\vartheta \mid y, X) \mid y\right] &= \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{E}_{\vartheta^{(0)}}\left[\mathbb{1}\{X_i = k\} \mid \cancel{y_1}, \dots, y_i, \dots, \cancel{y_n}\right] \left[\log p_k + \log f_{\theta_k}(y_i)\right] \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{P}_{\vartheta^{(0)}}\left(X_i = k \mid y_i\right) \left[\log p_k + \log f_{\theta_k}(y_i)\right]. \end{aligned}$$

Επομένως, το E-step του αλγορίθμου EM για μίξεις κατανομών ανάγεται στον υπολογισμό των δεσμευμένων πιθανοτήτων:

$$w_{ik} = \mathbb{P}_{\vartheta}\left(X_i = k \mid y_i\right).$$

Προφανώς, αυτές οι δεσμευμένες πιθανότητες υπολογίζονται μέσω του θεωρήματος του Bayes:

$$w_{ik} \propto \mathbb{P}_p(X_i = k) f_\theta\left(y_i \mid X_i = k\right) = p_k f_{\theta_k}(y_i).$$

Επιπλέον, ορίζουμε τις σταθερές κανονικοποίησης:

$$c_i(\vartheta) = \sum_{\ell=1}^{K} p_\ell f_{\theta_\ell}(y_i), \quad i = 1, 2, \dots, n.$$

Επιστρέφοντας στη σχέση που δίνει την παρατηρούμενη πιθανοφάνεια της πεπερασμένης μίξης κατανομών, παρατηρούμε ότι:

$$\mathcal{L}(\vartheta \mid y) = \prod_{i=1}^{n} c_i(\vartheta), \quad \ell(\vartheta \mid y) = \log \mathcal{L}(\vartheta \mid y) = \sum_{i=1}^{n} \log c_i(\vartheta).$$

Σύμφωνα με τη θεωρία του αλγορίθμου EM, η παρατηρούμενη πιθανοφάνεια θα πρέπει αναγκαστικά να αυξάνεται σε κάθε βήμα του αλγορίθμου.

Οι κατανομές πιθανότητας και η παρατηρούμενη πιθανοφάνεια, πρέπει να υπολογίζονται αποκλειστικά σε λογαριθμική κλίμακα για λόγους αριθμητικής ευστάθειας. Για τον λόγο αυτό χρησιμοποιούμε το λεγόμενο κόλπο Log-Sum-Exp. Για παράδειγμα, ορίζουμε:

$$v_{ik} = \log p_k + \log f_{\theta_k}(y_i), \quad m_i = \max_{k \in \{1,\dots,K\}} v_{ik}.$$

Τότε, παίρνουμε ότι:

$$w_{ik} = \frac{e^{v_{ik}-m_i}}{\sum_{\ell=1}^{K} e^{v_{i\ell}-m_i}}, \quad \log c_i(\vartheta) = m_i + \log \sum_{\ell=1}^{K} e^{v_{i\ell}-m_i}.$$

Έχοντας υπολογίσει τις πιθανότητες w_{ik} για $\vartheta = \vartheta^{(0)}$, έχουμε προσδιορίσει πλήρως την ενδιάμεση ποσότητα του EM. Επομένως, μπορούμε να προχωρήσουμε στο M-step, το οποίο εξαρτάται από τις συναρτήσεις (πυκνότητας) πιθανότητας $f_{\theta_k}(\cdot)$.

4 Εφαρμογή του Αλγορίθμου EM σε Μίξεις Κατανομών Poisson

Έστω ότι $(Y_i \mid X_i = k) \sim \text{Poisson}(\theta_k)$. Σύμφωνα με την προηγούμενη παράγραφο, υπολογίζουμε ότι:

$$Q_{\vartheta^{(0)}}(\vartheta) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \left[\log p_k - \theta_k + y_i \log \theta_k - \log(y_i!)\right].$$

Πρώτα μεγιστοποιούμε την ενδιάμεση ποσότητα του EM ως προς το διάνυσμα πιθανοτήτων p, λαμβάνοντας υπόψιν τον περιορισμό $\sum_{k=1}^{K} p_k = 1$. Ορίζουμε τη Λαγκραντζιανή συνάρτηση:

$$\mathcal{L}(p,\lambda) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \log p_k - \lambda \left(\sum_{k=1}^{K} p_k - 1 \right).$$

Παραγωγίζοντας ως προς p_k, υπολογίζουμε ότι:

$$\frac{\partial \mathcal{L}(p,\lambda)}{\partial p_k} = \frac{1}{p_k} \sum_{i=1}^{n} w_{ik} - \lambda \quad \Rightarrow \quad p_k^{(1)} = \frac{1}{\lambda} \sum_{i=1}^{n} w_{ik}.$$

Παρατηρούμε ότι $\sum_{k=1}^{K} w_{ik} = 1$. Εφαρμόζοντας τον περιορισμό $\sum_{k=1}^{K} p_k^{(1)} = 1$, υπολογίζουμε το λ:

$$1 = \sum_{k=1}^{K} p_k^{(1)} = \frac{1}{\lambda} \sum_{k=1}^{K} \sum_{i=1}^{n} w_{ik} = \frac{1}{\lambda} \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} = \frac{1}{\lambda} \sum_{i=1}^{n} 1 = \frac{n}{\lambda} \quad \Rightarrow \quad \lambda = n.$$

Επομένως, καταλήγουμε ότι:

$$p_k^{(1)} = \frac{1}{n} \sum_{i=1}^{n} w_{ik},$$

δηλαδή η νέα εκτίμηση του p_k είναι ο μέσος όρος της δεσμευμένης posterior πιθανότητας κάθε παρατήρησης να προέρχεται από την κατηγορία k. Προφανώς, η ίδια μεγιστοποίηση ισχύει για οποιαδήποτε πεπερασμένη μίξη κατανομών.

Έπειτα, παραγωγίζουμε την ενδιάμεση ποσότητα του EM ως προς θ_k:

$$\frac{\partial Q_{\vartheta^{(0)}}(\vartheta)}{\partial \theta_k} = \sum_{i=1}^{n} w_{ik} \left(-1 + \frac{y_i}{\theta_k}\right) \quad \Rightarrow \quad \theta_k^{(1)} = \frac{\sum_{i=1}^{n} w_{ik} y_i}{\sum_{i=1}^{n} w_{ik}},$$

δηλαδή η νέα εκτίμηση του θ_k είναι ο σταθμισμένος μέσος όρος όλων των παρατηρήσεων, όπου κάθε παρατήρηση σταθμίζεται με τη δεσμευμένη posterior πιθανότητα να ανήκει στην κατηγορία k.

Έχοντας υπολογίσει την εκτίμηση μέγιστης πιθανοφάνειας $\widehat{\vartheta}$ μέσω του αλγορίθμου EM, μπορούμε επιπλέον να υπολογίσουμε τις κατανομές πιθανότητας w_{ik} για $\vartheta = \widehat{\vartheta}$. Τότε, μπορούμε να υπολογίσουμε τις maximum a posteriori εκτιμήσεις των λανθανουσών μεταβλητών $X_1, \ldots, X_n$ ως εξής:

$$\widehat{X}_i = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}} \; w_{ik}, \quad i = 1, 2, \ldots, n.$$

```
loglikpoismix = function(Y, theta, p) {
    n = length(Y)
    K = length(theta)
    w = matrix(0, n, K)
    loglik = 0
    for (i in 1:n) {
        logw = log(p) + dpois(Y[i], theta, log = TRUE)
        maximum = max(logw)
        unnormalized = exp(logw - maximum)
        c = sum(unnormalized)
        w[i, ] = unnormalized/c
        loglik = loglik + maximum + log(c)
    }
    return(list(w = w, loglik = loglik))
}

EMpoismix = function(Y, theta, p, tol = 1e-05) {
    steps = 1
    poismix = loglikpoismix(Y, theta, p)
    w = poismix$w
    loglik = poismix$loglik
    err = Inf
    while (err > tol) {
        steps = steps + 1
        p = colMeans(w)
        theta = colMeans(w * Y)/p
        poismix = loglikpoismix(Y, theta, p)
        w = poismix$w
        loglik[steps] = poismix$loglik
        err = loglik[steps] - loglik[steps - 1]
```

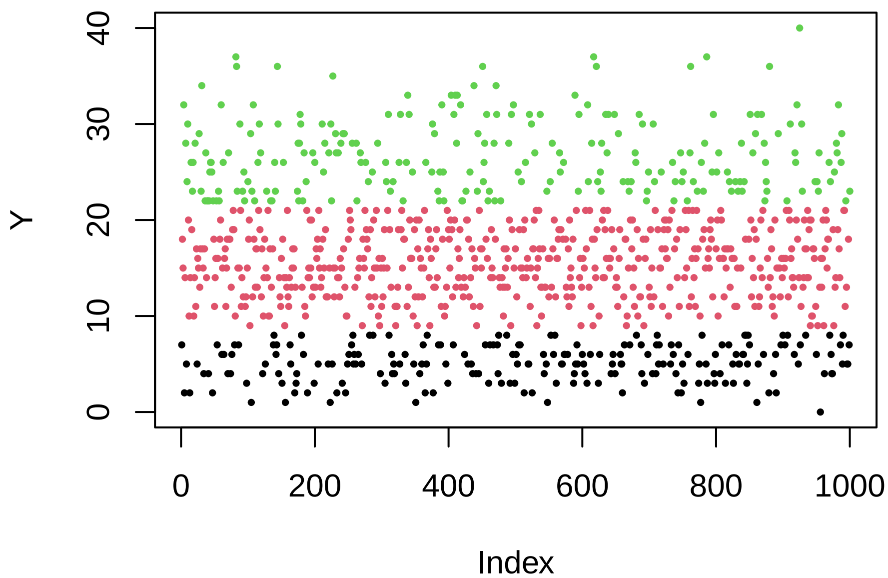
```r
    }
    return(list(theta = theta, p = p, w = w, loglik = loglik))
}

MLE = EMpoismix(Y, mean(Y) + sd(Y) * ((1 - K)/2):((K - 1)/2), rep(1, K)/K)
print(MLE$theta)
```

```
## [1]  5.281834 15.588634 25.617278
```

```r
print(MLE$p)
```

```
## [1] 0.2171178 0.5270992 0.2557829
```

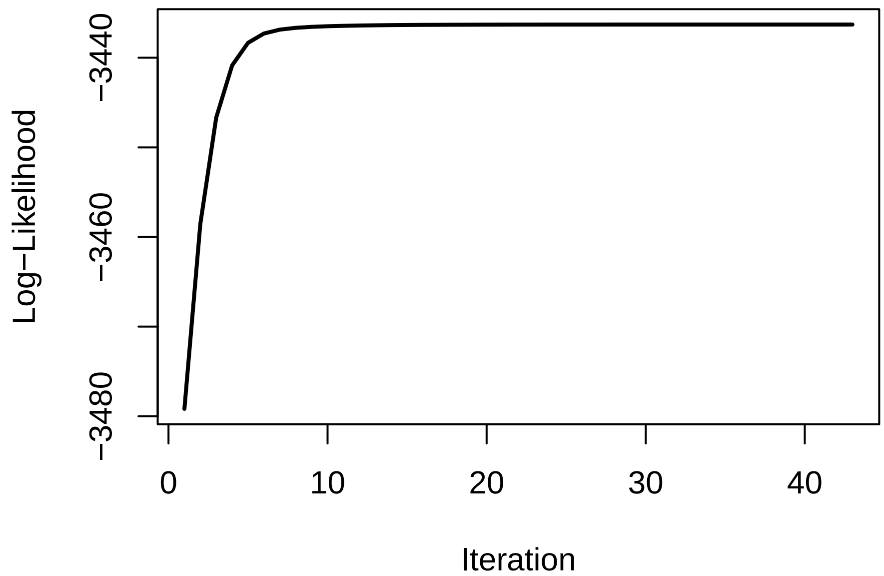
```r
XMAP = apply(MLE$w, 1, which.max)
plot(Y, col = XMAP, pch = 16, cex = 0.5)
```




```r
table(X, XMAP)
```

```
##    XMAP
## X     1   2   3
##    1 191  13   0
##    2  24 447  28
##    3   0  86 211
```

```r
plot(MLE$loglik[-1], type = "l", xlab = "Iteration", ylab = "Log-Likelihood",
    lwd = 2)
```



5 Εφαρμογή του Δειγματολήπτη Gibbs σε Μίξεις Κατανομών

Θεωρούμε τη δεσμευμένα συζυγή Dirichlet(α) prior για το διάνυσμα πιθανοτήτων p με συνάρτηση πυκνότητας πιθανότητας:

$$f(p) = \frac{\Gamma(K\alpha)}{[\Gamma(\alpha)]^K} \prod_{k=1}^{K} p_k^{\alpha-1} \propto \prod_{k=1}^{K} p_k^{\alpha-1}.$$

Τότε, η δεσμευμένη posterior κατανομή του διανύσματος p δίνεται από τη σχέση:

$$f(p \mid x) \propto f(p) \prod_{i=1}^{n} \mathbb{P}(X_i = x_i \mid p) \propto \prod_{k=1}^{K} \left[p_k^{\alpha-1} \prod_{i=1}^{n} p_k^{\mathbb{1}_{\{x_i=k\}}} \right] = \prod_{k=1}^{K} p_k^{N_k+\alpha-1},$$

δηλαδή $p \mid x \sim$ Dirichlet$\left(N_1+\alpha, N_2+\alpha, \dots, N_K+\alpha\right)$, όπου $N_k = \sum_{i=1}^{n} \mathbb{1}_{\{x_i=k\}}$ για $k = 1, 2, \dots, K$. Ως ειδική περίπτωση μπορούμε να θεωρήσουμε την prior του Jeffreys για το διάνυσμα p, η οποία προκύπτει για $\alpha = 0.5$. Για να προσομοιώσουμε από τη δεσμευμένη posterior κατανομή του διανύσματος p, μπορούμε πρώτα να προσομοιώσουμε τυχαίες μεταβλητές $W_k \sim$ Gamma$(N_k+\alpha, 1)$ για $k = 1, 2, \dots, K$ και μετά να θέσουμε $p = \frac{1}{W}\left(W_1, W_2, \dots, W_K\right)$, όπου $W = \sum_{k=1}^{K} W_k$.

Έπειτα, θεωρούμε τη δεσμευμένα συζυγή Gamma(β, λ) prior για την παράμετρο θ_k με συνάρτηση πυκνότητας πιθανότητας:

$$f(\theta_k) = \frac{\lambda^\beta}{\Gamma(\beta)} \theta_k^{\beta-1} e^{-\lambda\theta_k} \propto \theta_k^{\beta-1} e^{-\lambda\theta_k}.$$

Τότε, η δεσμευμένη posterior κατανομή της παραμέτρου θ_k δίνεται από τη σχέση:

$$f(\theta_k \mid x, y) \propto f(\theta_k) \prod_{i=1}^{n} f(y_i \mid x_i, \theta_k) \propto \theta_k^{\beta-1} e^{-\lambda\theta_k} \prod_{i=1}^{n} \left(e^{-\theta_k} \frac{\theta_k^{y_i}}{y_i!} \right)^{\mathbb{1}_{\{x_i=k\}}} \propto \theta_k^{S_k+\beta-1} e^{-(N_k+\lambda)\theta_k},$$

δηλαδή $\theta_k \mid x, y \sim$ Gamma$(S_k+\beta, N_k+\lambda)$, όπου $S_k = \sum_{i=1}^{n} y_i \mathbb{1}_{\{x_i=k\}}$ για $k = 1, 2, \dots, K$. Ως ειδική περίπτωση μπορούμε να θεωρήσουμε τη μη-ορθή prior του Jeffreys για την παράμετρο θ_k, η οποία δίνεται από τη σχέση $f(\theta_k) \propto \theta_k^{-0.5}$ και προκύπτει για $\beta = 0.5$, $\lambda = 0$.

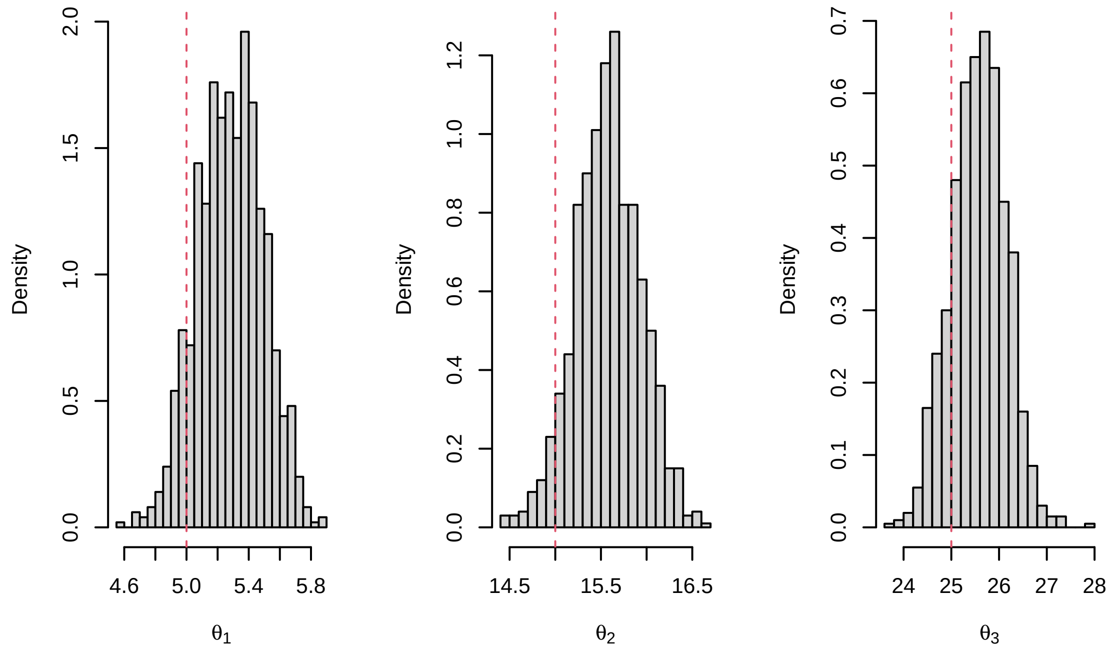
Τέλος, γνωρίζουμε ότι η δεσμευμένη posterior κατανομή της λανθάνουσας μεταβλητής X_i δίνεται από τη σχέση:

$$\mathbb{P}(X_i = k \mid y_i, \vartheta) = w_{ik} \propto p_k f_{\theta_k}(y_i).$$

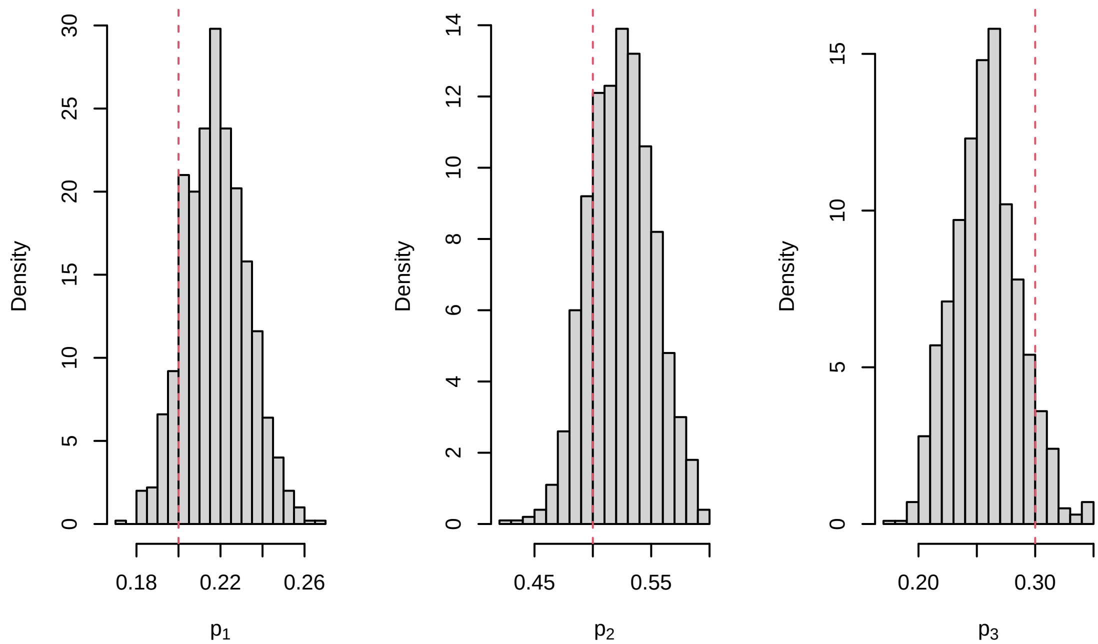
Επειδή η πιθανοφάνεια του μοντέλου μίξης κατανομών και οι prior που επιλέγουμε για τις παραμέτρους του αντιμετωπίζουν συμμετρικά τις παραμέτρους που αντιστοιχούν σε διαφορετικές κατηγορίες του μοντέλου, πολλές φορές εμφανίζεται το φαινόμενο του label switching στην εφαρμογή των αλγορίθμων MCMC για την προσομοίωση από την από κοινού posterior κατανομή των παραμέτρων του μοντέλου. Για να προσπαθήσουμε να αντιμετωπίσουμε αυτό το φαινόμενο, μπορούμε να επιβάλουμε έναν περιορισμό ταυτοποιησιμότητας όπως $\theta_1 < \theta_2 < \cdots < \theta_K$ σε κάθε βήμα του αλγορίθμου MCMC.

```
MCMCpoismix = function(Y, theta0, p0, alpha, beta, lambda, niter, nburn) {
    K = length(theta)
    theta = matrix(0, niter, K)
    p = matrix(0, niter, K)
    X = matrix(0, niter, n)
    theta[1, ] = theta0
    p[1, ] = p0
    w = loglikpoismix(Y, theta[1, ], p[1, ])$w
    X[1, ] = apply(w, 1, function(x) {
        sample(K, 1, prob = x)
    })
    for (j in 2:niter) {
        x = factor(X[j - 1, ], levels = 1:K)
        N = table(x)
        p[j, ] = rgamma(K, N + alpha)
        p[j, ] = p[j, ]/sum(p[j, ])
        S = aggregate(Y ~ x, FUN = sum, drop = FALSE)[, 2]
        S[is.na(S)] = 0
        theta[j, ] = rgamma(K, S + beta, N + lambda)
        w = loglikpoismix(Y, theta[j, ], p[j, ])$w
        X[j, ] = apply(w, 1, function(x) {
            sample(K, 1, prob = x)
        })
        I = order(theta[j, ])
        theta[j, ] = theta[j, I]
        p[j, ] = p[j, I]
        X[j, ] = I[X[j, ]]
    }
    return(list(theta = theta[-(1:nburn), ], p = p[-(1:nburn), ], X = X[-(1:nburn),
        ]))
}

posterior = MCMCpoismix(Y, mean(Y) + sd(Y) * ((1 - K)/2):((K - 1)/2), rep(1,
```

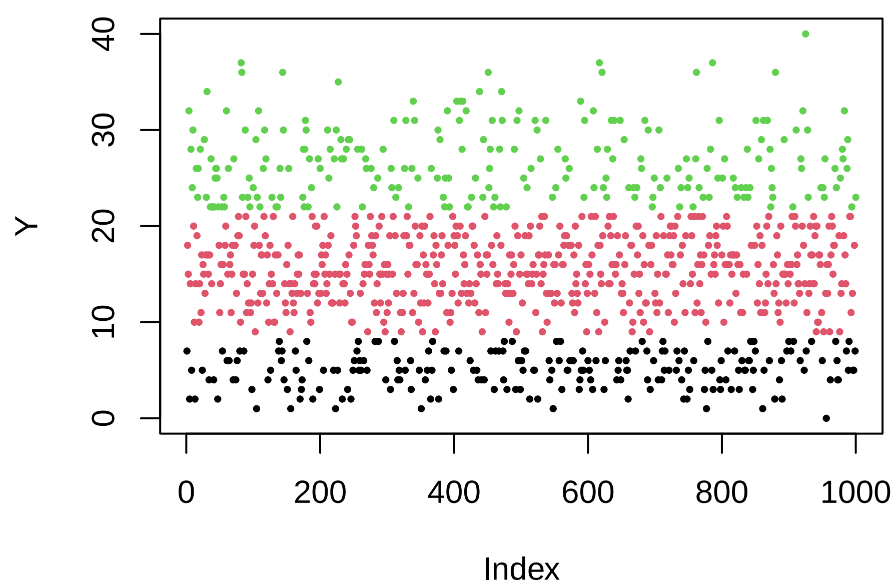
```
    K)/K, 0.5, 0.5, 0, 2000, 1000)
par(mfrow = c(1, 3))
hist(posterior$theta[, 1], "FD", freq = FALSE, main = NA, xlab = expression(theta[1]))
abline(v = theta[1], col = 2, lty = 2)
hist(posterior$theta[, 2], "FD", freq = FALSE, main = NA, xlab = expression(theta[2]))
abline(v = theta[2], col = 2, lty = 2)
hist(posterior$theta[, 3], "FD", freq = FALSE, main = NA, xlab = expression(theta[3]))
abline(v = theta[3], col = 2, lty = 2)
```




```
hist(posterior$p[, 1], "FD", freq = FALSE, main = NA, xlab = expression(p[1]))
abline(v = p[1], col = 2, lty = 2)
hist(posterior$p[, 2], "FD", freq = FALSE, main = NA, xlab = expression(p[2]))
abline(v = p[2], col = 2, lty = 2)
hist(posterior$p[, 3], "FD", freq = FALSE, main = NA, xlab = expression(p[3]))
abline(v = p[3], col = 2, lty = 2)
```




```r
XMAP = apply(posterior$X, 2, function(x) {
    which.max(table(factor(x, levels = 1:K)))
})
plot(Y, col = XMAP, pch = 16, cex = 0.5)
```




```r
table(X, XMAP)
```

```
##    XMAP
## X     1   2   3
##   1 191  13   0
##   2  24 447  28
##   3   0  86 211
```

6 Εφαρμογή του Γενικευμένου Λόγου Πιθανοφανειών σε Μίξεις Κατανομών

Έστω ότι θέλουμε να συγκρίνουμε δύο εμφωλευμένα μοντέλα μίξεων κατανομών με διαφορετικό πλήθη κατηγοριών. Θεωρούμε το μηδενικό μοντέλο $\mathcal{M}_0$ με $K-1$ κατηγορίες και το εναλλακτικό μοντέλο $\mathcal{M}_1$ με K κατηγορίες. Υπολογίζουμε την εκτίμηση μέγιστης πιθανοφάνειας $\hat{\vartheta}_0$ του μοντέλου $\mathcal{M}_0$ και την εκτίμηση μέγιστης πιθανοφάνειας $\hat{\vartheta}_1$ του μοντέλου $\mathcal{M}_1$ με χρήση του αλγορίθμου EM. Τότε, παίρνουμε την παρατηρούμενη τιμή του κριτηρίου γενικευμένου λόγου πιθανοφανειών ως εξής:

$$\mathrm{LR}^{\mathrm{obs}} = -2\left[\ell\left(\hat{\vartheta}_0 \mid y\right) - \ell\left(\hat{\vartheta}_1 \mid y\right)\right].$$

Γενικά γνωρίζουμε ότι η στατιστική συνάρτηση του κριτηρίου γενικευμένου λόγου πιθανοφανειών ακολουθεί την κατανομή χ^2_ν υπό την ισχύ του μοντέλου $\mathcal{M}_0$, όπου ν είναι το πλήθος των περιορισμών που πρέπει να τεθούν στο μοντέλο $\mathcal{M}_1$ ώστε να προκύψει το μοντέλο $\mathcal{M}_0$ ως ειδική περίπτωσή του. Όμως, στην περίπτωση των μοντέλων μίξεων κατανομών με διαφορετικό πλήθος κατηγοριών, το πλήθος βαθμών ελευθερίας ν δεν ορίζεται μονοσήμαντα, οπότε δεν ισχύουν οι προϋποθέσεις του θεωρήματος Wilks.

Για τον λόγο αυτόν μπορούμε να χρησιμοποιήσουμε τη μέθοδο του παραμετρικού bootstrap για να εκτιμήσουμε την κατανομή της στατιστικής συνάρτησης του κριτηρίου γενικευμένου λόγου πιθανοφανειών υπό την ισχύ του μοντέλου $\mathcal{M}_0$. Με άλλα λόγια, προσομοιώνουμε δείγματα $y^{(1)}, y^{(2)}, \dots, y^{(n_{\mathrm{boot}})}$ από το μοντέλο $\mathcal{M}_0$ με διάνυσμα παραμέτρων $\hat{\vartheta}_0$. Έπειτα, υπολογίζουμε την εκτίμηση μέγιστης πιθανοφάνειας $\hat{\vartheta}_0^{(j)}$ του μοντέλου $\mathcal{M}_0$ και την εκτίμηση μέγιστης πιθανοφάνειας $\hat{\vartheta}_1^{(j)}$ του μοντέλου $\mathcal{M}_1$ με χρήση του αλγορίθμου EM βάσει του τυχαίου δείγματος $y^{(j)}$ για $j = 1, 2, \dots, n_{\mathrm{boot}}$. Τέλος, υπολογίζουμε την τιμή $\mathrm{LR}_j^{\mathrm{boot}}$ του bootstrapped κριτηρίου γενικευμένου λόγου πιθανοφανειών, όπως κάναμε για την τιμή $\mathrm{LR}^{\mathrm{obs}}$. Τότε, μπορούμε να εκτιμήσουμε το p-value του κριτηρίου γενικευμένου λόγου πιθανοφανειών ως εξής:

$$\text{p-value} = \frac{1 + \sum_{j=1}^{n_{\mathrm{boot}}} \mathbb{1}_{\{\mathrm{LR}_j^{\mathrm{boot}} > \mathrm{LR}^{\mathrm{obs}}\}}}{1 + n_{\mathrm{boot}}}.$$

```r
LRpoismix = function(n, theta, p, LRobs, nboot, tol = 1e-05) {
    K = length(theta)
    LRboot = numeric(nboot)
    for (i in 1:nboot) {
        Y = rpoismix(n, theta, p)$Y
        loglik0 = EMpoismix(Y, mean(Y) + sd(Y) * ((1 - K)/2):((K - 1)/2), rep(1,
            K)/K, tol)$loglik
        loglik1 = EMpoismix(Y, mean(Y) + sd(Y) * (-K/2):(K/2), rep(1, K + 1)/(K +
            1), tol)$loglik
        LRboot[i] = -2 * (tail(loglik0, 1) - tail(loglik1, 1))
    }
    pval = (1 + sum(LRboot > LRobs))/(1 + nboot)
    return(pval)
}
```

```
MLE0 = EMpoismix(Y, mean(Y) + sd(Y) * ((1 - K/2):(K/2 - 1)), rep(1, K - 1)/(K -
    1))
LRobs = -2 * (tail(MLE0$loglik, 1) - tail(MLE$loglik, 1))
print(LRobs)
```

```
## [1] 257.3839
```

```
LRpoismix(n, MLE0$theta, MLE0$p, LRobs, 100)
```

```
## [1] 0.00990099
```

7 Κρυμμένα Μαρκοβιανά Μοντέλα (ΚΜΜ)

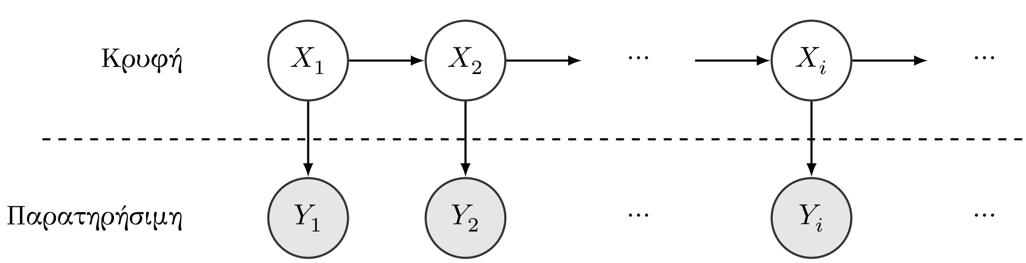
Έστω $\{X_i\}$ χρονικά ομογενής Μαρκοβιανή αλυσίδα διακριτού χρόνου με αρχική κατανομή $p_k = \mathbb{P}(X_1 = k)$ και πιθανότητες μετάβασης πρώτης τάξης $p_{k,\ell} = \mathbb{P}\left(X_{i+1} = \ell \mid X_i = k\right)$ για $k, \ell = 1, 2, \ldots, K$.

Θεωρούμε μία δεύτερη στοχαστική διαδικασία $\{Y_i\}$. Έστω ότι η στοχαστική διαδικασία $\{Y_i\}$ είναι δεσμευμένα ανεξάρτητη δεδομένης της Μαρκοβιανής αλυσίδας $\{X_i\}$, ότι η κατανομή της τυχαίας μεταβλητής Y_i δεδομένης της Μαρκοβιανής αλυσίδας $\{X_i\}$ εξαρτάται μόνο από την τυχαία μεταβλητή X_i και ότι $(Y_i \mid X_i = k) \sim f_{\theta_k}(\cdot)$. Με άλλα λόγια,

$$
\begin{aligned}
f_\theta\left(y_1, \ldots, y_n \mid X_1 = x_1, \ldots, X_n = x_n\right) &= \prod_{i=1}^{n} f_\theta\left(y_i \mid X_1 = x_1, \ldots, X_n = x_n\right) \\
&= \prod_{i=1}^{n} f_\theta\left(y_i \mid X_i = x_i\right) = \prod_{i=1}^{n} f_{\theta_{x_i}}(y_i).
\end{aligned}
$$

Τότε, η δισδιάστατη στοχαστική διαδικασία $\{(X_i, Y_i)\}$ είναι ένα χρονικά ομογενές κρυμμένο Μαρκοβιανό μοντέλο διακριτού χρόνου με πεπερασμένο χώρο καταστάσεων.

Η Μαρκοβιανή αλυσίδα $\{X_i\}$ είναι κρυφή, δηλαδή μη-παρατηρήσιμη. Την εισάγουμε για να μοντελοποιήσουμε τη σειριακή εξάρτηση της παρατηρήσιμης στοχαστικής διαδικασίας $\{Y_i\}$. Η γραφική αναπαράσταση της δομής εξάρτησης ενός ΚΜΜ φαίνεται στο παρακάτω γράφημα.



Σε αντίθεση με τη Μαρκοβιανή αλυσίδα $\{X_i\}$ της οποίας κάθε τυχαία μεταβλητή X_i εξαρτάται μόνο από την τιμή της αμέσως προηγούμενης τυχαίας μεταβλητής X_{i-1}, κάθε τυχαία μεταβλητή Y_i της στοχαστικής διαδικασίας $\{Y_i\}$ εξαρτάται από τις τιμές όλων των παρελθοντικών τυχαίων μεταβλητών $Y_1, \ldots, Y_{i-1}$. Δηλαδή, παρότι η στοχαστική διαδικασία $\{Y_i\}$ είναι δεσμευμένα ανεξάρτητη δεδομένης της Μαρκοβιανής αλυσίδας $\{X_i\}$, χωρίς αυτήν τη δέσμευση έχει "άπειρη μνήμη".

Σε όλους τους υπολογισμούς στα ΚΜΜ παίζει κεντρικό ρόλο η έννοια της δεσμευμένης ανεξαρτησίας. Λέμε ότι μία τυχαία μεταβλητή X και μία τυχαία μεταβλητής Y είναι δεσμευμένα ανεξάρτητες δεδομένης μίας τυχαίας μεταβλητής Z και συμβολίζουμε με $(X \perp\!\!\!\perp Y) \mid Z$ όταν η δεσμευμένη κατανομή της X δεδομένης της Z και η δεσμευμένη κατανομή της Y δεδομένης της Z είναι ανεξάρτητες.

Μπορούμε εύκολα να συμεράνουμε τη δεσμευμένη ανεξαρτησία από το γράφημα της δομής εξάρτησης του ΚΜΜ. Συγκεκριμένα, αν δεσμεύσουμε σε μία τυχαία μεταβλητή Z η οποία "κόβει" το μονοπάτι μεταξύ των τυχαίων μεταβλητών X και Y, τότε $(X \perp\!\!\!\perp Y) \mid Z$. Για παράδειγμα,

$$((Y_1, \dots, Y_i) \perp\!\!\!\perp (Y_{i+1}, \dots, Y_n)) \mid X_i,$$

$$((Y_1, \dots, Y_i) \perp\!\!\!\perp X_{i+1}) \mid X_i,$$

$$((Y_i, \dots, Y_n) \perp\!\!\!\perp X_{i-1}) \mid X_i.$$

Έστω δείγμα $y_1, \dots, y_n$ από μία στοχαστική διαδικασία. Εισάγουμε την κρυφή Μαρκοβιανή αλυσίδα $\{X_i\}$. Η πλήρης πιθανοφάνεια, δηλαδή η από κοινού πιθανοφάνεια των παρατηρούμενων μεταβλητών y_i και των κρυφών μεταβλητών x_i, για $\vartheta = \left(p_1, \dots, p_K, p_{1,1}, \dots, p_{K,K}, \theta_1, \dots, \theta_K\right)$ γράφεται ως:

$$\begin{aligned}
\mathcal{L}(\vartheta \mid y, x) &= f_\vartheta(x, y) = \mathbb{P}_\vartheta(X = x) f_\vartheta(y \mid x) \\
&= \mathbb{P}_p\left(X_1 = x_1\right) \cdot \prod_{i=2}^{n} \mathbb{P}_p\left(X_i = x_i \mid X_{i-1} = x_{i-1}\right) \cdot \prod_{i=1}^{n} f_\theta\left(y_i \mid X_i = x_i\right) \\
&= p_{x_1} \cdot \prod_{i=2}^{n} p_{x_{i-1}, x_i} \cdot \prod_{i=1}^{n} f_{\theta_{x_i}}(y_i) \\
&= \prod_{k=1}^{K} p_k^{\mathbb{1}\{x_i = k\}} \cdot \prod_{i=2}^{n} \prod_{k=1}^{K} \prod_{\ell=1}^{K} p_{k,\ell}^{\mathbb{1}\{x_{i-1} = k, x_i = \ell\}} \cdot \prod_{i=1}^{n} \prod_{k=1}^{K} f_{\theta_k}(y_i)^{\mathbb{1}\{x_i = k\}}.
\end{aligned}$$

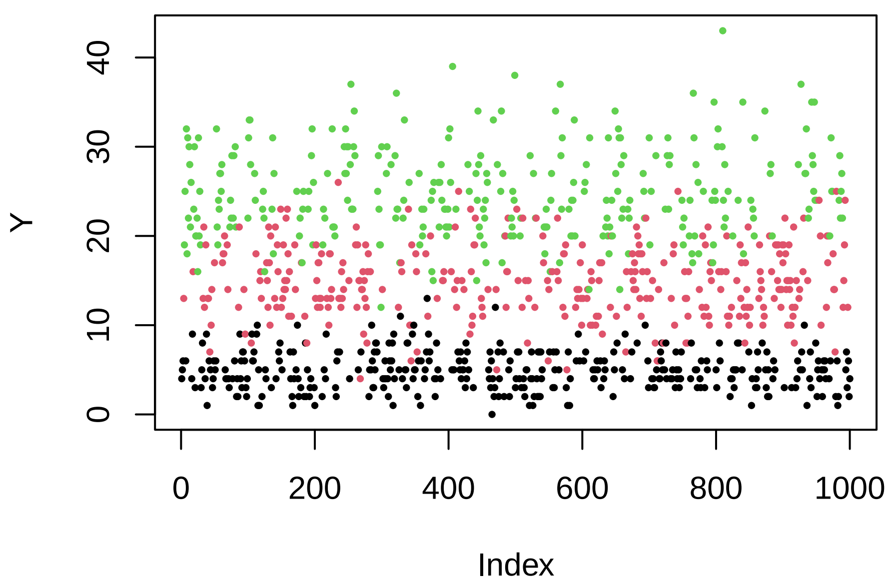
Από την άλλη, η παρατηρούμενη πιθανοφάνεια δίνεται από τον πολλαπλασιαστικό νόμο:

$$\begin{aligned}
\mathcal{L}(\vartheta \mid y) &= f_\vartheta(y_1) \prod_{i=2}^{n} f_\vartheta(y_i \mid y_1, \dots, y_{i-1}) \\
&= \left[\sum_{k=1}^{K} \mathbb{P}_p\left(X_1 = k\right) f_\theta\left(y_1 \mid X_1 = k\right)\right] \\
&\quad \times \prod_{i=2}^{n} \left[\sum_{k=1}^{K} \mathbb{P}_\vartheta\left(X_i = k \mid y_1, \dots, y_{i-1}\right) f_\theta\left(y_i \mid \cancel{y_1, \dots, y_{i-1}}, X_i = k\right)\right] \\
&= \left[\sum_{k=1}^{K} p_k f_{\theta_k}(y_1)\right] \prod_{i=2}^{n} \left[\sum_{k=1}^{K} \mathbb{P}_\vartheta\left(X_i = k \mid y_1, \dots, y_{i-1}\right) f_{\theta_k}(y_i)\right].
\end{aligned}$$

Η παρατηρούμενη πιθανοφάνεια είναι αδύνατον να μεγιστοποιηθεί αναλυτικά και σε αυτήν την περίπτωση εξαιτίας των αθροισμάτων που εμπλέκονται.

Σε αντίθεση με τις πεπερασμένες μίξεις κατανομών, στις οποίες ο υπολογισμός της παρατηρούμενης πιθανοφάνειας για συγκεκριμένες τιμές των αγνώστων παραμέτρων είναι άμεσος, ο υπολογισμός της για ένα ΚΜΜ προϋποθέτει τον υπολογισμό των δεσμευμένων πιθανοτήτων $\phi_{i|i-1}(k) = \mathbb{P}_\vartheta\left(X_i = k \mid y_1, \dots, y_{i-1}\right)$. Αυτές οι δεσμευμένες πιθανότητες καλούνται predictive, καθώς προβλέπουν την τρέχουσα κατάσταση της κρυφής Μαρκοβιανής αλυσίδας δεδομένων των τιμών των παρελθοντικών παρατηρούμενων μεταβλητών. Ο αναλυτικός υπολογισμός τους απαιτεί ένα αναδρομικό σχήμα. Ένα τέτοιο αναδρομικό σχήμα εμπλέκεται σε κάθε μέθοδο εκτίμησης των αγνώστων παραμέτρων του ΚΜΜ, οπότε ο υπολογισμός της παρατηρούμενης πιθανοφάνειας είναι άμεση απόρροια της εκτίμησης των παραμέτρων του ΚΜΜ.

```
rpoisHMM = function(n, theta, p, P) {
    K = length(theta)
    X = numeric(n)
    X[1] = sample(K, 1, prob = p)
    for (i in 2:n) {
        X[i] = sample(K, 1, prob = P[X[i - 1], ])
    }
    Y = rpois(n, theta[X])
    return(list(Y = Y, X = X))
}

n = 1000
K = 3
theta = c(5, 15, 25)
p = c(1, 0, 0)
P = matrix(c(0.5, 0.3, 0.2, 0.3, 0.6, 0.1, 0.2, 0.1, 0.7), 3)
HMM = rpoisHMM(n, theta, p, P)
Y = HMM$Y
X = HMM$X
plot(Y, col = X, pch = 16, cex = 0.5)
```



8 Εφαρμογή του Αλγορίθμου EM σε ΚΜΜ

Αρχικά, λογαριθμίζουμε την πλήρη πιθανοφάνεια του ΚΜΜ:

$$\begin{aligned}\ell(\vartheta \mid y, x) = \sum_{k=1}^{K} \mathbb{1}\{x_1 = k\} \log p_k &+ \sum_{i=2}^{n} \sum_{k=1}^{K} \sum_{\ell=1}^{K} \mathbb{1}\{x_{i-1} = k, x_i = \ell\} \log p_{k,\ell} \\ &+ \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{1}\{x_i = k\} \log f_{\theta_k}(y_i).\end{aligned}$$

Έπειτα, υπολογίζουμε την ενδιάμεση ποσότητα του EM:

$$\begin{aligned}Q_{\vartheta^{(0)}}(\vartheta) = \sum_{k=1}^{K} \mathbb{P}_{\vartheta^{(0)}}\left(X_1 = k \mid y_1, \dots, y_n\right) \log p_k &+ \sum_{i=2}^{n} \sum_{k=1}^{K} \sum_{\ell=1}^{K} \mathbb{P}_{\vartheta^{(0)}}\left(X_{i-1} = k, X_i = \ell \mid y_1, \dots, y_n\right) \log p_{k,\ell} \\ &+ \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{P}_{\vartheta^{(0)}}\left(X_i = k \mid y_1, \dots, y_n\right) \log f_{\theta_k}(y_i).\end{aligned}$$

Επομένως, το E-step του αλγορίθμου EM στα ΚΜΜ ανάγεται στον υπολογισμό των δεσμευμένων πιθανοτήτων $\phi_{i|n}(k) = \mathbb{P}_{\vartheta}\left(X_i = k \mid y_1, \dots, y_n\right)$ και $\phi_{i-1,i|n}(k, \ell) = \mathbb{P}_{\vartheta}\left(X_{i-1} = k, X_i = \ell \mid y_1, \dots, y_n\right)$. Αυτές οι δεσμευμένες πιθανότητες καλούνται smoothing, επειδή εξομαλύνουν τον θόρυβο που προκύπτει από την από κοινού κατανομή όλων των παρατηρούμενων μεταβλητών για να συνάγουν την κατανομή της κρυφής Μαρκοβιανής αλυσίδας. Αυτές οι δεσμευμένες πιθανότητες δεν μπορούν να υπολογιστούν άμεσα μέσω του θεωρήματος του Bayes, αλλά απαιτούν ένα αναδρομικό σχήμα, το οποίο περιλαμβάνει και τον υπολογισμό των predictive πιθανοτήτων.

9 Αλγόριθμος Forward-Backward

Ο αλγόριθμος Forward-Backward έχει ως τελικό στόχο τον υπολογισμό των περιθωρίων smoothing κατανομών $\phi_{i|n}(\cdot)$ και των δισδιάστατων smoothing κατανομών $\phi_{i-1,i|n}(\cdot, \cdot)$, οπότε ενσωματώνεται στο E-step του αλγορίθμου EM για την εκτίμηση των αγνώστων παραμέτρων του ΚΜΜ. Αποτελείται από δύο διαδοχικά αναδρομικά σχήματα.

Το πρώτο αναδρομικό σχήμα του αλγορίθμου στοχεύει στον υπολογισμό των δεσμευμένων πιθανοτήτων $\phi_{i|i}(k) = \mathbb{P}_{\vartheta}\left(X_i = k \mid y_1, \dots, y_i\right)$. Αυτές οι δεσμευμένες πιθανότητες καλούνται filtering, επειδή φιλτράρουν τις τιμές των έως τώρα παρατηρούμενων μεταβλητών για να συνάγουν την κατανομή της τρέχουσας κατάστασης της κρυφής Μαρκοβιανής αλυσίδας. Αυτό το αναδρομικό σχήμα διατρέχει από την αρχή μέχρι το τέλος όλες τις καταστάσεις της κρυφής Μαρκοβιανής αλυσίδας. Για τον λόγο αυτό ονομάζεται forward filtering.

Το δεύτερο αναδρομικό σχήμα χρησιμοποιεί τις filtering πιθανότητες $\phi_{i|i}(\cdot)$ για τον υπολογισμό των smoothing πιθανοτήτων. Αυτό το αναδρομικό σχήμα διατρέχει από το τέλος μέχρι την αρχή όλες τις καταστάσεις της κρυφής Μαρκοβιανής αλυσίδας. Για τον λόγο αυτό ονομάζεται backward smoothing.

Forward Filtering

Στόχος μας είναι να υπολογίσουμε τις filtering κατανομές $\phi_{i|i}(\cdot)$ για $i = 1, 2, \ldots, n$. Εφαρμόζοντας το θεώρημα του Bayes, παίρνουμε ότι:

$$\begin{aligned}\phi_{1|1}(k) &= \mathbb{P}_{\vartheta}\left(X_1 = k \mid y_1\right) \\ &\propto \mathbb{P}_{p}\left(X_1 = k\right) f_{\theta}\left(y_1 \mid X_1 = k\right) \\ &= p_k f_{\theta_k}(y_1),\end{aligned}$$

$$\begin{aligned}\phi_{i|i}(k) &= \mathbb{P}_{\vartheta}\left(X_i = k \mid y_1, \ldots, y_{i-1}, y_i\right) \\ &\propto \mathbb{P}_{\vartheta}\left(X_i = k \mid y_1, \ldots, y_{i-1}\right) f_{\theta}\left(y_i \mid \cancel{y_1, \ldots, y_{i-1}}, X_i = k\right) \\ &= \phi_{i|i-1}(k) f_{\theta_k}(y_i), \quad i = 2, 3, \ldots, n.\end{aligned}$$

Βλέπουμε ότι ο υπολογισμός κάθε filtering κατανομής για $i = 2, 3, \ldots, n$ προϋποθέτει τον υπολογισμό της τρέχουσας predictive κατανομής $\phi_{i|i-1}(\cdot)$. Εφαρμόζοντας το θεώρημα ολικής πιθανότητας, παίρνουμε ότι:

$$\begin{aligned}\phi_{i|i-1}(k) &= \mathbb{P}_{\vartheta}\left(X_i = k \mid y_1, \ldots, y_{i-1}\right) \\ &= \sum_{\ell=1}^{K} \mathbb{P}_{\vartheta}\left(X_{i-1} = \ell \mid y_1, \ldots, y_{i-1}\right) \underbrace{\mathbb{P}_{p}\left(X_i = k \mid X_{i-1} = \ell, \cancel{y_1, \ldots, y_{i-1}}\right)}_{((Y_1, \ldots, Y_{i-1}) \perp\!\!\!\perp X_i) \mid X_{i-1}} \\ &= \sum_{\ell=1}^{K} \phi_{i-1|i-1}(\ell) p_{\ell,k}, \quad i = 2, 3, \ldots, n.\end{aligned}$$

Βλέπουμε ότι υπολογισμός κάθε predictive κατανομής $\phi_{i|i-1}(\cdot)$ προϋποθέτει τον υπολογισμό της αμέσως προηγούμενης filtering κατανομής $\phi_{i-1|i-1}(\cdot)$. Επομένως, παίρνουμε ένα αναδρομικό σχήμα σύμφωνα με το οποίο υπολογίζουμε εναλλάξ τις filtering και τις predictive κατανομές της κρυφής Μαρκοβιανής αλυσίδας από την αρχή μέχρι το τέλος.

Επιπλέον, ορίζουμε τις σταθερές κανονικοποίησης των filtering κατανομών:

$$c_1(\vartheta) = \sum_{\ell=1}^{K} p_\ell f_{\theta_\ell}(y_1),$$

$$c_i(\vartheta) = \sum_{\ell=1}^{K} \phi_{i|i-1}(\ell) f_{\theta_\ell}(y_i), \quad i = 2, 3, \ldots, n.$$

Επιστρέφοντας στη σχέση που μάς δίνει την παρατηρούμενη πιθανοφάνεια του ΚΜΜ, παρατηρούμε ότι:

$$\mathcal{L}(\vartheta \mid y) = c_1(\vartheta) \cdot \prod_{i=2}^{n} c_i(\vartheta) = \prod_{i=1}^{n} c_i(\vartheta), \quad \ell(\vartheta \mid y) = \log \mathcal{L}(\vartheta \mid y) = \sum_{i=1}^{n} \log c_i(\vartheta).$$

Επομένως, ως απόρροια του αλγορίθμου forward filtering, μπορούμε να υπολογίσουμε την παρατηρούμενη πιθανοφάνεια του ΚΜΜ για την τρέχουσα εκτίμηση του ϑ. Σύμφωνα με τη θεωρία του αλγορίθμου EM, η παρατηρούμενη πιθανοφάνεια θα πρέπει αναγκαστικά να αυξάνεται σε κάθε βήμα του αλγορίθμου.

Όπως πάντα, οι filtering κατανομές και η παρατηρούμενη πιθανοφάνεια, πρέπει να υπολογίζονται αποκλειστικά σε λογαριθμική κλίμακα για λόγους αριθμητικής ευστάθειας. Για τον λόγο αυτό χρησιμοποιούμε το κόλπο Log-Sum-Exp. Για παράδειγμα, ορίζουμε:

$$v_k = \log p_k + \log f_{\theta_k}(y_1), \quad m_1 = \max_{k \in \{1,\dots,K\}} v_k.$$

Τότε, παίρνουμε ότι:

$$\phi_{1|1}(k) = \frac{e^{v_k - m_1}}{\sum_{\ell=1}^{K} e^{v_\ell - m_1}}, \quad \log c_1(\vartheta) = m_1 + \log \sum_{\ell=1}^{K} e^{v_\ell - m_1}.$$

Το ίδιο εφαρμόζουμε για τις υπόλοιπες filtering κατανομές. Το αναδρομικό σχήμα forward filtering συνοψίζεται παρακάτω.

Algorithm 1 Forward Filtering

Είσοδος: Παρατηρήσεις $y_1, \dots, y_n$ και τρέχουσα εκτίμηση ϑ.

1: Υπολογίζουμε τη σταθερά κανονικοποίησης $\log c_1(\vartheta)$.

2: Υπολογίζουμε τη filtering κατανομή $\phi_{1|1}(\cdot)$.

3: Για $i = 2, 3, \dots, n$ υπολογίζουμε:

i: την predictive κατανομή $\phi_{i|i-1}(\cdot)$,

ii: τη σταθερά κανονικοποίησης $\log c_i(\vartheta)$,

iii: τη filtering κατανομή $\phi_{i|i}(\cdot)$.

4: Υπολογίζουμε την παρατηρούμενη λογαριθμο-πιθανοφάνεια $\ell(\vartheta \mid y)$.

Έξοδος: Filtering κατανομές $\phi_{i|i}(\cdot)$ και παρατηρούμενη λογαριθμο-πιθανοφάνεια $\ell(\vartheta \mid y)$.

Backward Smoothing

Στόχος μας είναι να υπολογίσουμε τις smoothing κατανομές $\phi_{i|n}(\cdot)$ και $\phi_{i-1,i|n}(\cdot,\cdot)$. Αρχικά, ορίζουμε τις backward μεταβλητές:

$$b_i(k) = f_\vartheta\left(y_{i+1}, \dots, y_n \mid X_i = k\right), \quad i = 1, 2, \dots, n-1.$$

Εφαρμόζοντας το θεώρημα του Bayes, παίρνουμε ότι:

$$\begin{aligned}
\phi_{i|n}(k) &= \mathbb{P}_\vartheta\left(X_i = k \mid y_1, \dots, y_i, y_{i+1}, \dots, y_n\right) \\
&\propto \mathbb{P}_\vartheta\left(X_i = k \mid y_1, \dots, y_i\right) \underbrace{f_\vartheta\left(y_{i+1}, \dots, y_n \mid \cancel{y_1, \dots, y_i}, X_i = k\right)}_{((Y_1,\dots,Y_i) \perp\!\!\!\perp (Y_{i+1},\dots,Y_n))|X_i} \\
&= \phi_{i|i}(k) b_i(k), \quad i = 1, 2, \dots, n-1,
\end{aligned}$$

$$
\begin{aligned}
\phi_{i-1,i|n}(k,\ell) &= \mathbb{P}_\vartheta\left(X_{i-1}=k, X_i=\ell \mid y_1,\ldots,y_{i-1},y_i,\ldots,y_n\right)\\
&\propto \underbrace{\mathbb{P}_\vartheta\left(X_{i-1}=k, X_i=\ell \mid y_1,\ldots,y_{i-1}\right)}_{\text{πολλαπλασιαστικός νόμος}} \underbrace{f_\vartheta\left(y_i,\ldots,y_n \mid \cancel{y_1,\ldots,y_{i-1}}, \cancel{X_{i-1}=k}, X_i=\ell\right)}_{((Y_i,\ldots,Y_n)\perp\!\!\!\perp X_{i-1})|X_i}\\
&= \mathbb{P}_\vartheta\left(X_{i-1}=k \mid y_1,\ldots,y_{i-1}\right) \underbrace{\mathbb{P}_p\left(X_i=\ell \mid X_{i-1}=k, \cancel{y_1,\ldots,y_{i-1}}\right)}_{((Y_1,\ldots,Y_{i-1})\perp\!\!\!\perp X_i)|X_{i-1}}\\
&\quad\times \underbrace{f_\theta\left(y_i \mid X_i=\ell\right) f_\vartheta\left(y_{i+1},\ldots,y_n \mid X_i=\ell\right)}_{(Y_i\perp\!\!\!\perp(Y_{i+1},\ldots,Y_n))|X_i}\\
&= \phi_{i-1|i-1}(k)p_{k,\ell}f_{\theta_\ell}(y_i)b_i(\ell), \quad i=2,3,\ldots,n-1,
\end{aligned}
$$

$$
\begin{aligned}
\phi_{n-1,n|n}(k,\ell) &= \mathbb{P}_\vartheta\left(X_{n-1}=k, X_n=\ell \mid y_1,\ldots,y_{n-1},y_n\right)\\
&\propto \underbrace{\mathbb{P}_\vartheta\left(X_{n-1}=k, X_n=\ell \mid y_1,\ldots,y_{n-1}\right)}_{\text{πολλαπλασιαστικός νόμος}} f_\theta\left(y_n \mid \cancel{y_1,\ldots,y_{n-1}}, \cancel{X_{n-1}=k}, X_n=\ell\right)\\
&= \mathbb{P}_\vartheta\left(X_{n-1}=k \mid y_1,\ldots,y_{n-1}\right) \underbrace{\mathbb{P}_p\left(X_n=\ell \mid X_{n-1}=k, \cancel{y_1,\ldots,y_{n-1}}\right)}_{((Y_1,\ldots,Y_{n-1})\perp\!\!\!\perp X_n)|X_{n-1}} f_{\theta_\ell}(y_n)\\
&= \phi_{n-1|n-1}(k)p_{k,\ell}f_{\theta_\ell}(y_n).
\end{aligned}
$$

Βλέπουμε ότι ο υπολογισμός όλων των smoothing κατανομών καθορίζεται πλήρως από τον υπολογισμό των filtering κατανομών και των τρεχουσών backward μεταβλητών. Οι backward μεταβλητές υπολογίζονται αναδρομικά σύμφωνα με το θεώρημα ολικής πιθανότητας:

$$
\begin{aligned}
b_i(k) &= f_\vartheta\left(y_{i+1},\ldots,y_n \mid X_i=k\right)\\
&= \sum_{\ell=1}^{K} \mathbb{P}_p\left(X_{i+1}=\ell \mid X_i=k\right) \underbrace{f_\vartheta\left(y_{i+1},\ldots,y_n \mid \cancel{X_i=k}, X_{i+1}=\ell\right)}_{((Y_{i+1},\ldots,Y_n)\perp\!\!\!\perp X_i)|X_{i+1}}\\
&= \sum_{\ell=1}^{K} p_{k,\ell} \underbrace{f_\theta\left(y_{i+1} \mid X_{i+1}=\ell\right) f_\vartheta\left(y_{i+2},\ldots,y_n \mid X_{i+1}=\ell\right)}_{(Y_{i+1}\perp\!\!\!\perp(Y_{i+2},\ldots,Y_n))|X_{i+1}}\\
&= \sum_{\ell=1}^{K} p_{k,\ell}f_{\theta_\ell}(y_{i+1})b_{i+1}(\ell), \quad i=1,2,\ldots,n-2,
\end{aligned}
$$

$$
\begin{aligned}
b_{n-1}(k) &= f_\vartheta\left(y_n \mid X_{n-1}=k\right)\\
&= \sum_{\ell=1}^{K} \mathbb{P}_p\left(X_n=\ell \mid X_{n-1}=k\right) f_\theta\left(y_n \mid \cancel{X_{n-1}=k}, X_n=\ell\right)\\
&= \sum_{\ell=1}^{K} p_{k,\ell}f_{\theta_\ell}(y_n).
\end{aligned}
$$

Συνεπώς, παίρνουμε ένα αναδρομικό σχήμα σύμφωνα με το οποίο υπολογίζουμε εναλλάξ τις backward μεταβλητές και τις smoothing κατανομές της κρυφής Μαρκοβιανής αλυσίδας από το τέλος μέχρι την αρχή. Η τελευταία περιθώρια smoothing κατανομή ταυτίζεται με την τελευταία filtering κατανομή, η οποία έχει ήδη υπολογιστεί. Οι υπόλοιπες smoothing κατανομές υπολογίζονται με χρήση του κόλπου Log-Sum-Exp ακριβώς όπως οι filtering κατανομές. Το αναδρομικό σχήμα backward smoothing συνοψίζεται παρακάτω.

Algorithm 2 Backward Smoothing

Είσοδος: Παρατηρήσεις $y_1, \ldots, y_n$, τρέχουσα εκτίμηση ϑ και filtering κατανομές $\phi_{i|i}(\cdot)$.

1: Υπολογίζουμε τη δισδιάστατη smoothing κατανομή $\phi_{n-1,n|n}(\cdot,\cdot)$.

2: Υπολογίζουμε τις backward μεταβλητές $b_{n-1}(\cdot)$.

3: Για $i = n-1, n-2, \ldots, 2$ υπολογίζουμε:

i: την περιθώρια smoothing κατανομή $\phi_{i|n}(\cdot)$,

ii: τη δισδιάστατη smoothing κατανομή $\phi_{i-1,i|n}(\cdot,\cdot)$,

iii: τις backward μεταβλητές $b_{i-1}(\cdot)$.

4: Υπολογίζουμε την περιθώρια smoothing κατανομή $\phi_{1|n}(\cdot)$.

Έξοδος: Smoothing κατανομές $\phi_{i|n}(\cdot)$ και $\phi_{i-1,i|n}(\cdot,\cdot)$.

Markovian Backward Smoothing

Το Markovian backward smoothing είναι ένα εναλλακτικό αναδρομικό σχήμα που στοχεύει στον υπολογισμό των smoothing κατανομών $\phi_{i|n}(\cdot)$ και $\phi_{i-1,i|n}(\cdot,\cdot)$. Σε αντίθεση με το αναδρομικό σχήμα backward smoothing, μπορεί να οδηγήσει στον υπολογισμό της από κοινού smoothing κατανομής ολόκληρης της κρυφής Μαρκοβιανής αλυσίδας. Αυτό επιτρέπει μεταξύ άλλων τον σχεδιασμό αποτελεσματικών αλγορίθμων MCMC για την εκτίμηση της posterior κατανομής των αγνώστων παραμέτρων του ΚΜΜ. Στη θέση των backward μεταβλητών κάνει χρήση των backward πιθανοτήτων μετάβασης:

$$B_i(k,\ell) = \mathbb{P}_\vartheta\left(X_i = k \mid X_{i+1} = \ell, y_1, \ldots, y_i\right), \quad i = 1, 2, \ldots, n-1.$$

Παρατηρούμε ότι $\sum_{k=1}^{K} B_i(k,\ell) = 1$. Εφαρμόζοντας τον πολλαπλασιαστικό νόμο, υπολογίζουμε ότι:

$$\begin{aligned}
\phi_{i-1,i|n}(k,\ell) &= \mathbb{P}_\vartheta\left(X_{i-1} = k, X_i = \ell \mid y_1, \ldots, y_n\right) \\
&= \mathbb{P}_\vartheta\left(X_i = \ell \mid y_1, \ldots, y_n\right) \underbrace{\mathbb{P}_\vartheta\left(X_{i-1} = k \mid X_i = \ell, y_1, \ldots, y_{i-1}, \cancel{y_i, \ldots, y_n}\right)}_{((Y_i,\ldots,Y_n) \perp\!\!\!\perp X_{i-1} \mid X_i)} \\
&= \phi_{i|n}(\ell) B_{i-1}(k,\ell), \quad i = 2, 3, \ldots, n.
\end{aligned}$$

Βλέπουμε ότι ο υπολογισμός κάθε δισδιάστατης smoothing κατανομής καθορίζεται πλήρως από τον υπολογισμό της αμέσως προηγούμενης περιθώριας smoothing κατανομής και των τρεχουσών backward πιθανοτήτων μετάβασης. Έπειτα, εφαρμόζοντας το θεώρημα ολικής πιθανότητες, παίρνουμε ότι:

$$\begin{aligned}
\phi_{i|n}(k) &= \mathbb{P}_\vartheta\left(X_i = k \mid y_1, \ldots, y_n\right) \\
&= \sum_{\ell=1}^{K} \mathbb{P}_\vartheta\left(X_i = k, X_{i+1} = \ell \mid y_1, \ldots, y_n\right) \\
&= \sum_{\ell=1}^{K} \phi_{i,i+1}(k,\ell), \quad i = 1, 2, \ldots, n-1.
\end{aligned}$$

Βλέπουμε ότι ο υπολογισμός κάθε περιθώριας smoothing κατανομής καθορίζεται πλήρως από τον υπολογισμό της τρέχουσας δισδιάστατης smoothing κατανομής. Οι backward πιθανότητες μετάβασης υπολογίζονται σύμφωνα με το θεώρημα του Bayes:

$$\begin{aligned} B_i(k,\ell) &= \mathbb{P}_\vartheta\left(X_i = k \mid X_{i+1} = \ell, y_1, \dots, y_i\right) \\ &\propto \mathbb{P}_\vartheta\left(X_i = k \mid y_1, \dots, y_i\right) \underbrace{\mathbb{P}_p\left(X_{i+1} = \ell \mid X_i = k, \cancel{y_1, \dots, y_i}\right)}_{((Y_1,\dots,Y_i) \perp\!\!\!\perp X_{i+1}) \mid X_i} \\ &= \phi_{i|i}(k) p_{k,\ell}, \quad i = 1, 2, \dots, n-1. \end{aligned}$$

Βλέπουμε ότι ο υπολογισμός των backward πιθανοτήτων μετάβασης καθορίζεται πλήρως από την αντίστοιχη filtering κατανομή. Συνεπώς, παίρνουμε ένα αναδρομικό σχήμα σύμφωνα με το οποίο υπολογίζουμε εναλλάξ τις backward πιθανότητες μετάβασης και τις smoothing κατανομές της κρυφής Μαρκοβιανής αλυσίδας από το τέλος μέχρι την αρχή.

Τελικά, μπορούμε να αναλύσουμε την από κοινού smoothing κατανομή ολόκληρης της κρυφής Μαρκοβιανής αλυσίδας με χρήση του πολλαπλασιαστικού νόμου:

$$\begin{aligned} \phi_{1,\dots,n|n}(x_1, \dots, x_n) &= \mathbb{P}_\vartheta(X_1 = x_1, \dots, X_n = x_n \mid y_1, \dots, y_n) \\ &= \mathbb{P}_\vartheta\left(X_n = x_n \mid y_1, \dots, y_n\right) \\ &\quad \times \prod_{i=1}^{n-1} \underbrace{\mathbb{P}_\vartheta\left(X_i = x_i \mid X_{i+1} = x_{i+1}, \cancel{X_{i+2} = x_{i+2}, \dots, X_n = x_n}, y_1, \dots, y_i, \cancel{y_{i+1}, \dots, y_n}\right)}_{\substack{((X_{i+2},\dots,X_n)) \perp\!\!\!\perp X_i \mid X_{i+1} \\ ((Y_{i+1},\dots,Y_n) \perp\!\!\!\perp X_i) \mid X_{i+1}}} \\ &= \phi_{n|n}(x_n) \prod_{i=1}^{n-1} B_i\left(x_i, x_{i+1}\right). \end{aligned}$$

Η από κοινού smoothing κατανομή είναι ουσιαστικά η δεσμευμένη posterior ολόκληρης της κρυφής Μαρκοβιανής αλυσίδας. Παρατηρούμε ότι η δεσμευμένη posterior κατανομή της στοχαστικής διαδικασίας $\{X_i\}$ εξακολουθεί να έχει τη Μαρκοβιανή ιδιότητα, αλλά δεν είναι πλέον χρονικά ομογενής. Οι backward πιθανότητες $B_i(\cdot,\cdot)$ ουσιαστικά δίνουν τον πίνακα πιθανοτήτων μετάβασης της δεσμευμένης posterior κατανομής της αντίστροφης Μαρκοβιανής αλυσίδας κατά την $(n-i)$-οστή μετάβασή της. Βλέπουμε εμφανώς ότι αυτός ο πίνακας πιθανοτήτων μετάβασης είναι διαφορετικός σε κάθε μετάβαση της Μαρκοβιανής αλυσίδας.

Οι backward πιθανότητες μετάβασης υπολογίζονται με χρήση του κόλπου Log-Sum-Exp. Ορίζουμε:

$$v_i(k,\ell) = \log \phi_{i|i}(k) + \log p_{k,\ell}, \quad m_i(\ell) = \max_{k \in \{1,\dots,K\}} v_i(k,\ell).$$

Τότε, παίρνουμε ότι:

$$B_i(k,\ell) = \frac{e^{v_i(k,\ell) - m_i(\ell)}}{\sum_{j=1}^K e^{v_i(j,\ell) - m_i(\ell)}}.$$

Το αναδρομικό σχήμα Markovian backward smoothing συνοψίζεται παρακάτω.

Algorithm 3 Markovian Backward Smoothing

Είσοδος: Τρέχουσα εκτίμηση ϑ και filtering κατανομές $\phi_{i|i}(\cdot)$.

1: Για $i = n-1, n-2, \dots, 1$ υπολογίζουμε:

i: τις backward πιθανότητες μετάβασης $B_i(\cdot, \cdot)$,

ii: τη δισδιάστατη smoothing κατανομή $\phi_{i,i+1|n}(\cdot, \cdot)$,

iii: την περιθώρια smoothing κατανομή $\phi_{i|n}(\cdot)$.

Έξοδος: Smoothing κατανομές $\phi_{i|n}(\cdot)$ και $\phi_{i,i+1|n}(\cdot, \cdot)$.

10 Αλγόριθμος Baum-Welch

Ξεκινάμε με κάποια αρχική εκτίμηση $\vartheta^{(0)}$. Στο E-step του αλγορίθμου EM υλοποιούμε τον αλγόριθμο Forward-Backward για $\vartheta = \vartheta^{(0)}$ και τελικά λαμβάνουμε τις smoothing κατανομές $\phi_{i|n}(\cdot)$, $\phi_{i-1,i|n}(\cdot, \cdot)$. Ο αλγόριθμος EM που ενσωματώνει στο E-step τον αλγόριθμο Forward-Backward για τον υπολογισμό των smoothing κατανομών ονομάζεται αλγόριθμος Baum-Welch. Με αυτόν τον τρόπο προσδιορίζουμε πλήρως την ενδιάμεση ποσότητα του EM και προχωρούμε στο M-step.

Έστω ότι $(Y_i \mid X_i = k) \sim \text{Poisson}(\theta_k)$. Στη συγκεκριμένη περίπτωση, υπολογίζουμε ότι:

$$
\begin{aligned}
Q_{\vartheta^{(0)}}(\vartheta) = \sum_{k=1}^{K} \phi_{1|n}(k) \log p_k &+ \sum_{i=2}^{n} \sum_{k=1}^{K} \sum_{\ell=1}^{K} \phi_{i-1,i|n}(k, \ell) \log p_{k,\ell} \\
&+ \sum_{i=1}^{n} \sum_{k=1}^{K} \phi_{i|n}(k) \left[-\theta_k + y_i \log \theta_k - \log(y_i!)\right].
\end{aligned}
$$

Αρχικά, σημειώνουμε ότι η εκτίμηση της αρχικής κατανομής της κρυφής Μαρκοβιανής αλυσίδας βασίζεται αποκλειστικά στη δεσμευμένη posterior κατανομή της X_1. Επομένως, δεν υπάρχει ελπίδα να πάρουμε κάποια συνεπή εκτίμηση για την αρχική κατανομή βασιζόμενοι μόνο σε μία ακολουθία παρατηρήσεων $y_1, \dots, y_n$. Εκτός αν έχουμε κάποια εκ των προτέρων γνώση για την αρχική κατανομή, η πιο συνήθης τακτική είναι να θεωρήσουμε ότι είναι γνωστή και ίση με τη διακριτή ομοιόμορφη στον χώρο καταστάσεων $S = \{1, 2, \dots, K\}$ της κρυφής Μαρκοβιανής αλυσίδας.

Γνωρίζουμε ότι $\sum_{\ell=1}^{K} p_{k,\ell} = 1$. Επομένως, μεγιστοποιούμε την ενδιάμεση ποσότητα του EM ως προς κάθε γραμμή του πίνακα πιθανοτήτων μετάβασης ξεχωριστά. Ορίζουμε τη Λαγκραντζιανή συνάρτηση:

$$
\mathcal{L}(p_k, \lambda_k) = \sum_{i=2}^{n} \sum_{\ell=1}^{K} \phi_{i-1,i|n}(k, \ell) \log p_{k,\ell} - \lambda_k \left(\sum_{\ell=1}^{K} p_{k,\ell} - 1 \right).
$$

Παραγωγίζοντας ως προς $p_{k,\ell}$, υπολογίζουμε ότι:

$$
\frac{\partial \mathcal{L}(p_k, \lambda_k)}{\partial p_{k,\ell}} = \frac{1}{p_{k,\ell}} \sum_{i=2}^{n} \phi_{i-1,i|n}(k, \ell) - \lambda_k \quad \Rightarrow \quad p_{k,\ell}^{(1)} = \frac{1}{\lambda_k} \sum_{i=2}^{n} \phi_{i-1,i|n}(k, \ell) = \frac{1}{\lambda_k} \sum_{i=1}^{n-1} \phi_{i,i+1|n}(k, \ell).
$$

Παρατηρούμε ότι $\sum_{\ell=1}^{K} \phi_{i,i+1|n}(k, \ell) = \phi_{i|n}(k)$. Εφαρμόζοντας τον περιορισμό $\sum_{\ell=1}^{K} p_{k,\ell}^{(1)} = 1$, υπολογίζουμε την τιμή του πολλαπλασιαστή Lagrange λ_k:

$$1 = \sum_{\ell=1}^{K} p_{k,\ell}^{(1)} = \frac{1}{\lambda_k}\sum_{\ell=1}^{K}\sum_{i=1}^{n-1}\phi_{i,i+1|n}(k,\ell) = \frac{1}{\lambda_k}\sum_{i=1}^{n-1}\sum_{\ell=1}^{K}\phi_{i,i+1|n}(k,\ell) = \frac{1}{\lambda_k}\sum_{i=1}^{n-1}\phi_{i|n}(k) \quad \Rightarrow \quad \lambda_k = \sum_{i=1}^{n-1}\phi_{i|n}(k).$$

Επομένως, καταλήγουμε ότι:

$$p_{k,\ell}^{(1)} = \frac{\sum_{i=1}^{n-1}\phi_{i,i+1|n}(k,\ell)}{\sum_{i=1}^{n-1}\phi_{i|n}(k)},$$

δηλαδή η νέα εκτίμηση του $p_{k,\ell}$ είναι ο σταθμισμένος μέσος όρος των δεσμευμένων posterior πιθανοτήτων να μεταβεί η κρυφή Μαρκοβιανή αλυσίδα από την κατάσταση k στην κατάσταση ℓ, όπου κάθε πιθανότητα σταθμίζεται με τη δεσμευμένη posterior πιθανότητα να βρίσκεται η Μαρκοβιανή αλυσίδα στην κατάσταση k. Προφανώς, η ίδια μεγιστοποίηση ισχύει για οποιοδήποτε ΚΜΜ με πεπερασμένο χώρο καταστάσεων.

Έπειτα, παραγωγίζουμε την ενδιάμεση ποσότητα του EM ως προς θ_k:

$$\frac{\partial Q_{\vartheta^{(0)}}(\vartheta)}{\partial \theta_k} = \sum_{i=1}^{n}\phi_{i|n}(k)\left(-1 + \frac{y_i}{\theta_k}\right) \quad \Rightarrow \quad \theta_k^{(1)} = \frac{\sum_{i=1}^{n}\phi_{i|n}(k)y_i}{\sum_{i=1}^{n}\phi_{i|n}(k)},$$

δηλαδή η νέα εκτίμηση του θ_k είναι ο σταθμισμένος μέσος όρος όλων των παρατηρήσεων, όπου κάθε παρατήρηση σταθμίζεται με τη δεσμευμένη posterior πιθανότητα να προέρχεται από την κατάσταση k. Βλέπουμε ότι η εκτίμηση των αγνώστων παραμέτρων των κατανομών $f_{\theta_k}(\cdot)$ είναι ίδια είτε πρόκειται για πεπερασμένη μίξη κατανομών είτε για ΚΜΜ.

```
forward = function(Y, theta, p, P) {
    n = length(Y)
    K = length(theta)
    filter = matrix(0, n, K)
    logfilter = log(p[p > 0]) + dpois(Y[1], theta[p > 0], log = TRUE)
    maximum = max(logfilter)
    unnormalized = exp(logfilter - maximum)
    c = sum(unnormalized)
    filter[1, p > 0] = unnormalized/c
    loglik = maximum + log(c)
    for (i in 2:n) {
        predict = colSums(filter[i - 1, ] * P)
        logfilter = log(predict) + dpois(Y[i], theta, log = TRUE)
        maximum = max(logfilter)
        unnormalized = exp(logfilter - maximum)
        c = sum(unnormalized)
        filter[i, ] = unnormalized/c
        loglik = loglik + maximum + log(c)
    }
    return(list(filter = filter, loglik = loglik))
}
```

```r
backward = function(filter, P) {
    n = dim(filter)[1]
    K = dim(P)[1]
    bivariate = array(0, c(K, K, n - 1))
    marginal = matrix(0, n, K)
    marginal[n, ] = filter[n, ]
    for (i in (n - 1):1) {
        for (k in 1:K) {
            logB = log(filter[i, ]) + log(P[, k])
            unnormalized = exp(logB - max(logB))
            B = unnormalized/sum(unnormalized)
            bivariate[, k, i] = marginal[i + 1, k] * B
        }
        marginal[i, ] = rowSums(bivariate[, , i])
    }
    return(list(marginal = marginal, bivariate = bivariate))
}

EMpoisHMM = function(Y, theta, p, P, tol = 1e-05) {
    steps = 1
    f = forward(Y, theta, p, P)
    loglik = f$loglik
    b = backward(f$filter, P)
    marginal = b$marginal
    bivariate = b$bivariate
    err = Inf
    while (err > tol) {
        steps = steps + 1
        P = apply(bivariate, 1:2, sum)
        P = P/rowSums(P)
        theta = colSums(marginal * Y)/colSums(marginal)
        f = forward(Y, theta, p, P)
        loglik[steps] = f$loglik
        b = backward(f$filter, P)
        marginal = b$marginal
        bivariate = b$bivariate
        err = loglik[steps] - loglik[steps - 1]
    }
    return(list(theta = theta, P = P, marginal = marginal, loglik = loglik))
}

MLE = EMpoisHMM(Y, mean(Y) + sd(Y) * ((1 - K)/2):((K - 1)/2), p, matrix(1, K,
    K)/K)
```

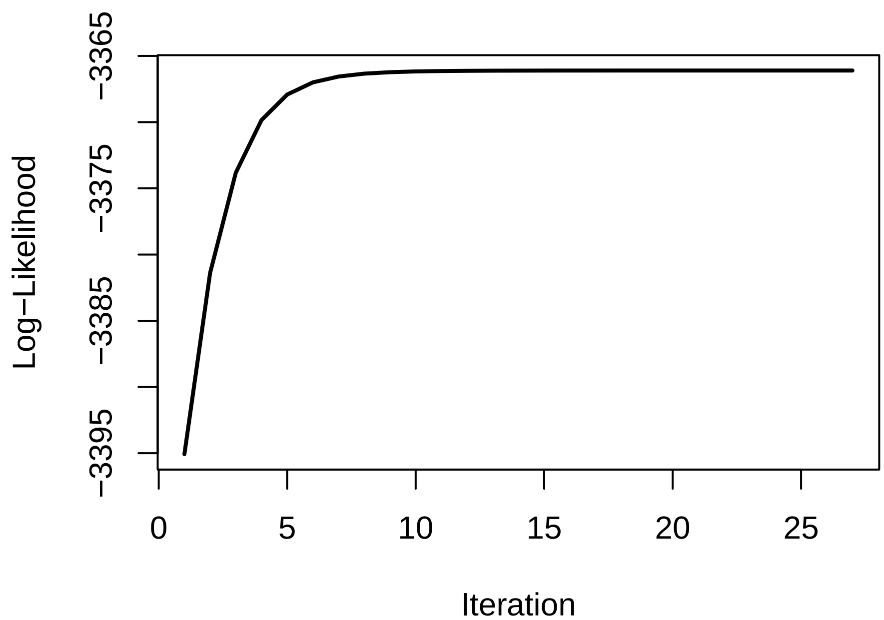
```r
print(MLE$theta)
```

```
## [1]  4.928938 15.160789 24.837596
```

```r
print(MLE$P)
```

```
##           [,1]      [,2]       [,3]
## [1,] 0.5733331 0.2123264 0.21434048
## [2,] 0.2633581 0.6434395 0.09320237
## [3,] 0.2191574 0.1412264 0.63961621
```

```r
plot(MLE$loglik[-1], type = "l", xlab = "Iteration", ylab = "Log-Likelihood",
    lwd = 2)
```



11 Αποκωδικοποίηση Κρυφών Καταστάσεων

Έχοντας εκτιμήσει όλες τις άγνωστες παραμέτρους ενός ΚΜΜ με χρήση του αλγορίθμου Baum-Welch ή κάποιου άλλου αλγορίθμου, μπορούμε να εκτιμήσουμε τις κρυφές καταστάσεις $x_1, \dots, x_n$ από τις οποίες προέχονται οι παρατηρήσεις $y_1, \dots, y_n$. Αυτή η διαδικασία καλείται αποκωδικοποίηση της κρυφής Μαρκοβιανής αλυσίδας. Η αποκωδικοποίηση διακρίνεται σε καθολική και τοπική.

Καθολική Αποκωδικοποίηση

Η καθολική αποκωδικοποίηση στοχεύει στη μεγιστοποίηση της δεσμευμένης posterior κατανομής ολόκληρης της κρυφής Μαρκοβιανής αλυσίδας δεδομένων όλων των παρατηρήσεων, δηλαδή της από κοινού smoothing κατανομής $\phi_{1,\dots,n|n}(\cdot, \dots, \cdot)$. Εφόσον αυτή είναι μία n-διάστατη κατανομή, η μεγιστοποίησή της δεν είναι άμεσα υλοποιήσιμη, αλλά απαιτεί κάποιον αλγόριθμο που υλοποιεί την αρχή του δυναμικού προγραμματισμού.

Ο αλγόριθμος Viterbi βασίζεται στην αρχή του δυναμικού προγραμματισμού και χρησιμοποιείται για την καθολική αποκωδικοποίηση της κρυφής Μαρκοβιανής αλυσίδας. Θέλουμε να ορίσουμε τη συνάρτηση βέλτιστης τιμής και να διατυπώσουμε τις εξισώσεις βελτιστότητας Bellman. Εφαρμόζοντας το θεώρημα του Bayes, παρατηρούμε ότι:

$$
\begin{aligned}
\phi_{1,\dots,i|i}(x_1,\dots,x_i) &= \mathbb{P}_\vartheta\left(X_1=x_1,\dots,X_i=x_i \mid y_1,\dots,y_i\right) \\
&= \frac{\mathbb{P}_\vartheta\left(X_1=x_1,\dots,X_i=x_i\right) f_\theta\left(y_1,\dots,y_i \mid X_1=x_1,\dots,X_i=x_i\right)}{f_\vartheta(y_1,\dots,y_i)} \\
&= \frac{1}{\mathcal{L}\left(\vartheta \mid y_1,\dots,y_i\right)} \cdot a_{x_1} \cdot \prod_{j=2}^{i} p_{x_{j-1},x_j} \cdot \prod_{j=1}^{i} f_{\theta_{x_j}}(y_j) \\
&= \frac{\mathcal{L}\left(\vartheta \mid y_1,\dots,y_{i-1}\right)}{\mathcal{L}\left(\vartheta \mid y_1,\dots,y_i\right)} \cdot \frac{p_{x_{i-1},x_i} f_{\theta_{x_i}}(y_i)}{\mathcal{L}\left(\vartheta \mid y_1,\dots,y_{i-1}\right)} \cdot a_{x_1} \cdot \prod_{j=2}^{i-1} p_{x_{j-1},x_j} \cdot \prod_{j=1}^{i-1} f_{\theta_{x_j}}(y_j) \\
&= \frac{\mathcal{L}\left(\vartheta \mid y_1,\dots,y_{i-1}\right)}{\mathcal{L}\left(\vartheta \mid y_1,\dots,y_i\right)} \cdot \phi_{1,\dots,i-1|i-1}(x_1,\dots,x_{i-1}) \cdot p_{x_{i-1},x_i} f_{\theta_{x_i}}(y_i), \quad i=2,3,\dots,n.
\end{aligned}
$$

Επομένως, ορίζουμε σε λογαριθμική κλίμακα τη συνάρτηση βέλτιστης τιμής:

$$
v_i(k) = \ell\left(\vartheta \mid y_1,\dots,y_i\right) + \max_{(x_1,\dots,x_{i-1})\in S^{i-1}} \log \phi_{1,\dots,i|i}(x_1,\dots,x_{i-1},k), \quad i=1,2,\dots,n.
$$

Αυτή αντιπροσωπεύει τη μέγιστη δεσμευμένη πιθανότητα μίας ακολουθίας κρυφών καταστάσεων η οποία καταλήγει την περίοδο i στην κατάσταση k δεδομένων όλων των παρατηρήσεων μέχρι την περίοδο i. Αρχικά, παρατηρούμε ότι:

$$
\begin{aligned}
v_1(k) &= \ell(\vartheta \mid y_1) + \log \phi_{1|1}(k) \\
&= \log\left[f_\vartheta(y_1)\mathbb{P}_\vartheta\left(X_1=k \mid y_1\right)\right] \\
&= \log\left[\mathbb{P}_p\left(X_1=k\right) f_\theta\left(y_1 \mid X_1=x_1\right)\right] \\
&= \log p_k + \log f_{\theta_k}(y_1).
\end{aligned}
$$

Αντικαθιστώντας, παίρνουμε τις εξισώσεις βελτιστότητας:

$$
\begin{aligned}
v_i(k) = &\cancel{\ell\left(\vartheta \mid y_1,\dots,y_i\right)} + \ell\left(\vartheta \mid y_1,\dots,y_{i-1}\right) - \cancel{\ell\left(\vartheta \mid y_1,\dots,y_i\right)} \\
&\qquad + \max_{(x_1,\dots,x_{i-1})\in S^{i-1}} \left[\log \phi_{1,\dots,i-1|i-1}(x_1,\dots,x_{i-1}) + \log p_{x_{i-1},k}\right] + \log f_{\theta_k}(y_i) \\
= &\log f_{\theta_k}(y_i) + \ell\left(\vartheta \mid y_1,\dots,y_{i-1}\right) \\
&\qquad + \max_{\ell\in\{1,\dots,K\}} \left[\max_{(x_1,\dots,x_{i-2})\in S^{i-2}} \log \phi_{1,\dots,i-1|i-1}(x_1,\dots,x_{i-2},\ell) + \log p_{\ell,k}\right] \\
= &\log f_{\theta_k}(y_i) + \max_{\ell\in\{1,\dots,K\}} \left[v_{i-1}(\ell) + \log p_{\ell,k}\right], \quad i=2,3,\dots,n.
\end{aligned}
$$

Επιπλέον, ορίζουμε:

$$
m_i(k) = \operatorname*{argmax}_{\ell\in\{1,\dots,K\}} \left[v_{i-1}(\ell) + \log p_{\ell,k}\right], \quad i=2,3,\dots,n.
$$

Επομένως, πρώτα υπολογίζουμε αναδρομικά όλες τις συναρτήσεις βέλτιστης τιμής από την αρχή μέχρι το τέλος της Μαρκοβιανής αλυσίδας. Στη συνέχεια, υπολογίζουμε την τελική κρυφή κατάσταση που μεγιστοποιεί την τελευταία συνάρτηση βέλτιστης τιμής, δηλαδή $x_n^* = \operatorname{argmax}_{k\in\{1,\dots,K\}} v_n(k)$. Αυτή είναι η τελική κατάσταση που μεγιστοποιεί την από κοινού δεσμευμένη posterior πιθανότητα ολόκληρης της ακολουθίας κρυφών καταστάσεων. Έπειτα, θέλουμε να μεγιστοποιήσουμε τη δεσμευμένη posterior πιθανότητα μίας ακολουθίας κρυφών καταστάσεων η οποία καταλήγει στην κατάσταση x_n^*. Η προτελευταία κατάσταση αυτής της κρυφής Μαρκοβιανής αλυσίδας δίνεται ως $x_{n-1}^* = m_n\left(x_n^*\right)$. Οι υπόλοιπες κρυφές καταστάσεις εκτιμούνται αναδρομικά με τον ίδιο τρόπο. Η βέλτιστη ακολουθία κρυφών καταστάσεων $x_1^*, \ldots, x_n^*$ καλείται μονοπάτι Viterbi. Τα βήματα του αλγορίθμου Viterbi συνοψίζονται παρακάτω.

Algorithm 4 Viterbi

Είσοδος: Παρατηρήσεις $y_1, \ldots, y_n$ και εκτίμηση ϑ.

1: Υπολογίζουμε τη συνάρτηση βέλτιστης τιμής $v_1(\cdot)$.

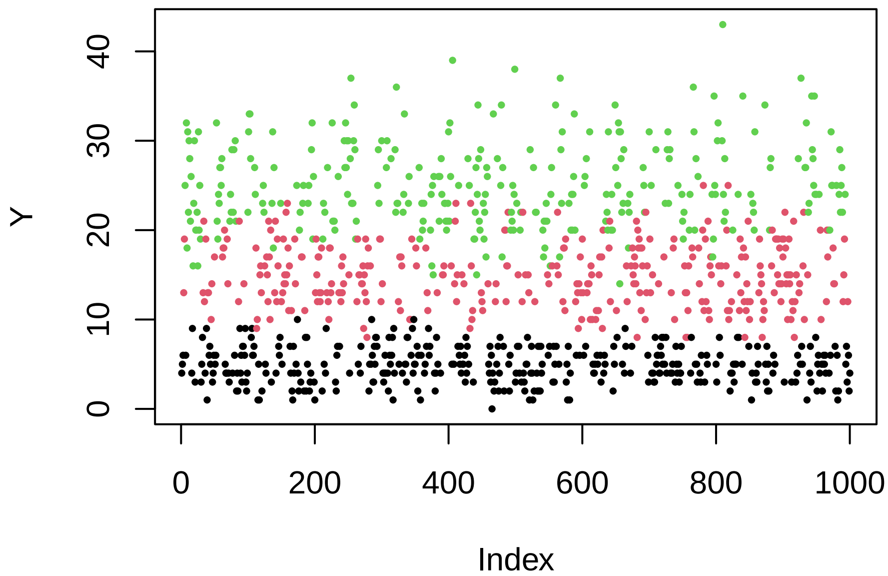
2: Για $i = 2, 3, \ldots, n$ υπολογίζουμε αναδρομικά τις υπόλοιπες συναρτήσεις βέλτιστης τιμής $v_i(\cdot)$.

3: Υπολογίζουμε τη βέλτιστη τελική κατάσταση $x_n^* = \operatorname{argmax}_{k \in \{1,\ldots,K\}} v_n(k)$.

4: Για $i = n-1, n-2, \ldots, 1$ υπολογίζουμε τις υπόλοιπες βέλτιστες καταστάσεις $x_i^* = m_{i+1}\left(x_{i+1}^*\right)$.

Έξοδος: Βέλτιστη ακολουθία κρυφών καταστάσεων $x_1^*, \ldots, x_n^*$.

```r
Viterbi = function(Y, theta, p, P) {
    n = length(Y)
    K = length(theta)
    v = matrix(0, n, K)
    m = matrix(0, n - 1, K)
    v[1, ] = log(p) + dpois(Y[1], theta, log = TRUE)
    for (i in 2:n) {
        for (k in 1:K) {
            temp = v[i - 1, ] + log(P[, k])
            m[i - 1, k] = which.max(temp)
            v[i, k] = dpois(Y[i], theta[k], log = TRUE) + max(temp)
        }
    }
    X = numeric(n)
    X[n] = which.max(v[n, ])
    for (i in (n - 1):1) {
        X[i] = m[i, X[i + 1]]
    }
    return(X)
}

XMAPjoint = Viterbi(Y, MLE$theta, p, MLE$P)
plot(Y, col = XMAPjoint, pch = 16, cex = 0.5)
```




```
table(X, XMAPjoint)
```

```
##    XMAPjoint
## X     1   2   3
##   1 347  10   0
##   2  17 301  21
##   3   0  21 283
```

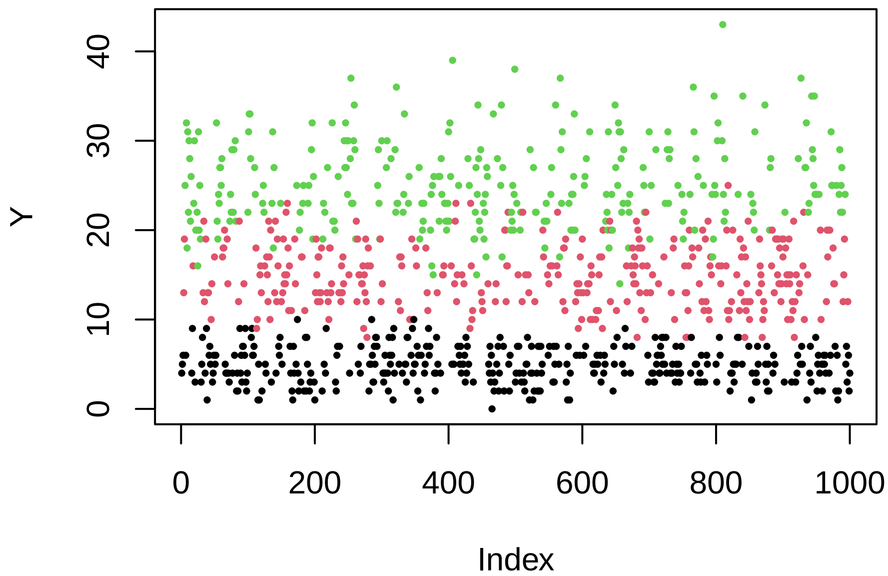
Τοπική Αποκωδικοποίηση

Η τοπική αποκωδικοποίηση στοχεύει στην ξεχωριστή μεγιστοποίηση των επί μέρους περιθωρίων δεσμευμένων posterior κάθε κρυφής κατάστασης, δηλαδή των περιθωρίων smoothing κατανομών $\phi_{i|n}(\cdot)$. Παρότι η καθολική αποκωδικοποίηση φαίνεται πιο ισχυρό αποτέλεσμα από την τοπική αποκωδικοποίηση, καμία από τις δύο αποκωδικοποιήσεις δε συνεπάγεται την άλλη. Για παράδειγμα, η τοπική αποκωδικοποίηση μπορεί να είναι προτιμότερη αν στόχος μας είναι η ορθότερη ταξινόμηση συγκεκριμένων παρατηρήσεων.

Έστω ότι έχουμε κάποια εκτίμηση ϑ^* για το ϑ. Εφαρμόζουμε τον αλγόριθμο Forward-Backward για $\vartheta = \vartheta^*$ και παίρνουμε τις περιθώριες smoothing κατανομές $\phi_{i|n}(\cdot)$ για $i = 1, 2, \dots, n$. Τότε, υπολογίζουμε ότι:

$$x_i^* = \underset{k \in \{1,\dots,K\}}{\operatorname{argmax}} \ \phi_{i|n}(k), \quad i = 1, 2, \dots, n.$$

```
XMAPmarginal = apply(MLE$marginal, 1, which.max)
plot(Y, col = XMAPmarginal, pch = 16, cex = 0.5)
```




```r
table(X, XMAPmarginal)
```

```
##    XMAPmarginal
## X     1   2   3
##   1 347  10   0
##   2  17 305  17
##   3   0  21 283
```

12 Αλγόριθμος Viterbi Training

Ο αλγόριθμος Viterbi training είναι μία εναλλακτική του αλγορίθμου Baum-Welch για την εκτίμηση των αγνώστων παραμέτρων του ΚΜΜ. Ξεκινάμε με κάποια αρχική εκτίμηση $\vartheta^{(0)}$. Στη θέση του αλγορίθμου Forward-Backward στο E-step του αλγορίθμου Baum-Welch εφαρμόζουμε τον αλγόριθμο Viterbi για $\vartheta = \vartheta^{(0)}$ και παίρνουμε το μονοπάτι Viterbi $x_1^{(1)}, \ldots, x_n^{(1)}$. Στη θέση του M-step του αλγορίθμου Baum-Welch μεγιστοποιούμε την πλήρη πιθανοφάνεια των δεδομένων για $x = x^{(1)}$. Επαναλαμβάνουμε αυτά τα δύο βήματα μέχρι ο αλγόριθμος Viterbi training να συγκλίνει σε κάποιο μονοπάτι Viterbi x^*.

Έστω ότι $(Y_i \mid X_i = k) \sim \text{Poisson}(\theta_k)$. Τότε, η πλήρης λογαριθμο-πιθανοφάνεια γράφεται ως:

$$\ell\left(\vartheta \mid y, x\right) = \sum_{k=1}^{K} \mathbb{1}\left\{x_1 = k\right\} \log p_k + \sum_{i=2}^{n} \sum_{k=1}^{K} \sum_{\ell=1}^{K} \mathbb{1}\left\{x_{i-1} = k, x_i = \ell\right\} \log p_{k,\ell}$$
$$+ \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{1}\left\{x_i = k\right\} \left[-\theta_k + y_i \log \theta_k - \log\left(y_i!\right)\right].$$

Ορίζουμε:

$$M_{k,\ell} = \sum_{i=2}^{n} \mathbb{1}\left\{x_{i-1} = k, x_i = \ell\right\} = \sum_{i=1}^{n-1} \mathbb{1}\left\{x_i = k, x_{i+1} = \ell\right\},$$

$$M_k = \sum_{i=1}^{n-1} \mathbb{1}\left\{x_i = k\right\}, \quad N_k = \sum_{i=1}^{n} \mathbb{1}\left\{x_i = k\right\}, \quad S_k = \sum_{i=1}^{n} \mathbb{1}\left\{x_i = k\right\} y_i.$$

Πρώτα μεγιστοποιούμε την πλήρη πιθανοφάνεια ως προς κάθε γραμμή του πίνακα πιθανοτήτων μετάβασης ξεχωριστά. Ορίζουμε τη Λαγκραντζιανή συνάρτηση:

$$\mathcal{L}(p_k, \lambda_k) = \sum_{\ell=1}^{K} M_{k,\ell} \log p_{k,\ell} - \lambda_k \left(\sum_{\ell=1}^{K} p_{k,\ell} - 1 \right).$$

Παραγωγίζοντας ως προς $p_{k,\ell}$, υπολογίζουμε ότι:

$$\frac{\partial \mathcal{L}(p_k, \lambda_k)}{\partial p_{k,\ell}} = \frac{1}{p_{k,\ell}} M_{k,\ell} - \lambda_k \quad \Rightarrow \quad p_{k,\ell}^{(1)} = \frac{1}{\lambda_k} M_{k,\ell}.$$

Παρατηρούμε ότι $\sum_{\ell=1}^{K} M_{k,\ell} = M_k$. Εφαρμόζοντας τον περιορισμό $\sum_{\ell=1}^{K} p_{k,\ell}^{(1)} = 1$, υπολογίζουμε την τιμή του πολλαπλασιαστή Lagrange λ_k:

$$1 = \sum_{\ell=1}^{K} p_{k,\ell}^{(1)} = \frac{1}{\lambda_k} \sum_{\ell=1}^{K} M_{k,\ell} = \frac{1}{\lambda_k} M_k \quad \Rightarrow \quad \lambda_k = M_k.$$

Επομένως, καταλήγουμε ότι:

$$p_{k,\ell}^{(1)} = \frac{M_{k,\ell}}{M_k},$$

δηλαδή η νέα εκτίμηση του $p_{k,\ell}$ είναι το ποσοστό των μεταβάσεων από την κατάσταση k που οδηγούν στην κατάσταση ℓ σύμφωνα με το τρέχον μονοπάτι Viterbi. Προφανώς, η ίδια μεγιστοποίηση ισχύει για οποιοδήποτε ΚΜΜ.

Έπειτα, παραγωγίζουμε την πλήρη λογαριθμο-πιθανοφάνεια ως προς θ_k:

$$\frac{\partial \ell(\vartheta \mid y, x)}{\partial \theta_k} = \sum_{i=1}^{n} \mathbb{1}\left\{x_i = k\right\} \left(-1 + \frac{y_i}{\theta_k} \right) \quad \Rightarrow \quad \theta_k^{(1)} = \frac{S_k}{N_k},$$

δηλαδή η νέα εκτίμηση του θ_k είναι ο μέσος όρος των παρατηρήσεων που προέρχονται από την κατάσταση k σύμφωνα με το τρέχον μονοπάτι Viterbi.

Σε αντίθεση με τον αλγόριθμο Baum-Welch, ο οποίος σταθμίζει όλα τα δυνατά μονοπάτια της κρυφής Μαρκοβιανής αλυσίδας σύμφωνα με τη δεσμευμένη posterior πιθανότητά τους προκειμένου να ανανεώσει τις εκτιμήσεις των αγνώστων παραμέτρων του ΚΜΜ, ο αλγόριθμος Viterbi training λαμβάνει υπόψιν του μόνο το μονοπάτι με τη συνολικά μέγιστη δεσμευμένη posterior πιθανότητα σε κάθε επανάληψη. Ως αποτέλεσμα, ο αλγόριθμος Viterbi training έχει μικρότερο υπολογιστικό κόστος σε σύγκριση με τον αλγόριθμο Baum-Welch, χωρίς όμως να προσφέρει καμία εγγύηση ότι η τελική εκτίμηση του ϑ θα βελτιστοποιεί πράγματι την παρατηρούμενη πιθανοφάνεια των δεδομένων.

```
VTpoisHMM = function(Y, theta, p, P) {
    K = length(theta)
    steps = 1
    loglik = forward(Y, theta, p, P)$loglik
    Xprev = numeric(n)
    X = Viterbi(Y, theta, p, P)
    while (sum(X != Xprev) > 0) {
```

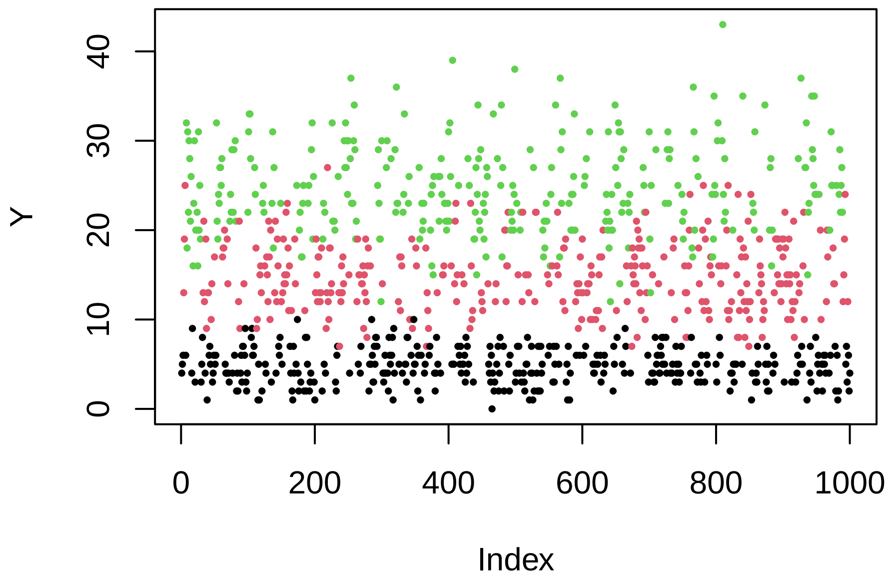
```
        steps = steps + 1
        x = factor(X, levels = 1:K)
        M = matrix(aggregate(numeric(n - 1), data.frame(x[-n], x[-1]), length,
            drop = FALSE)[, 3], K)
        M[is.na(M)] = 0
        P = M/rowSums(M)
        N = table(x)
        S = aggregate(Y ~ x, FUN = sum, drop = FALSE)[, 2]
        S[is.na(S)] = 0
        theta = S/N
        loglik[steps] = forward(Y, theta, p, P)$loglik
        Xprev = X
        X = Viterbi(Y, theta, p, P)
    }
    return(list(theta = theta, P = P, loglik = loglik))
}

VT = VTpoisHMM(Y, mean(Y) + sd(Y) * ((1 - K)/2):((K - 1)/2), p, matrix(1, K,
    K)/K)
print(VT$theta)
```

```
## x
##         1         2         3
##  4.750708 14.899110 24.745161
```

```
print(VT$P)
```

```
##           [,1]      [,2]       [,3]
## [1,] 0.5823864 0.1875000 0.23011364
## [2,] 0.2581602 0.7091988 0.03264095
## [3,] 0.1935484 0.1032258 0.70322581
```

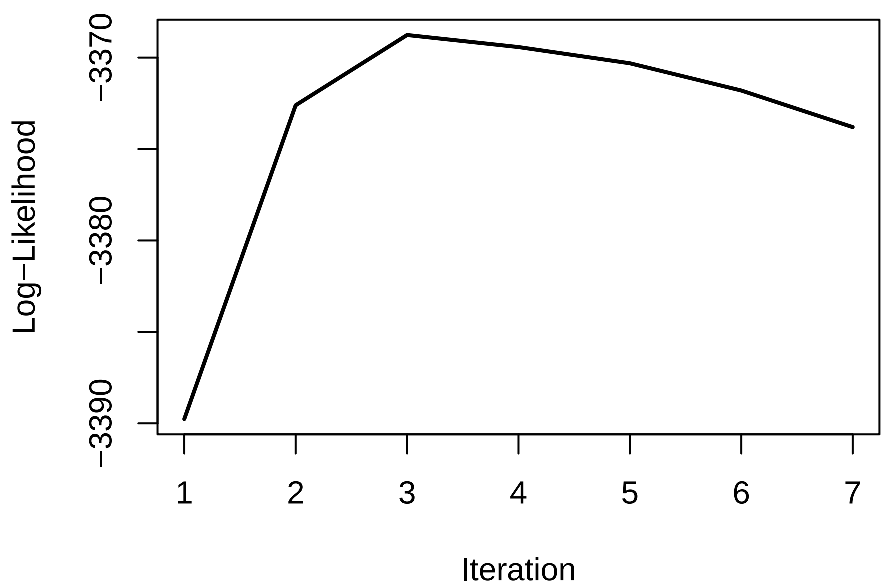
```
XMAPjoint = Viterbi(Y, VT$theta, p, VT$P)
plot(Y, col = XMAPjoint, pch = 16, cex = 0.5)
```




```r
table(X, XMAPjoint)
```

```
##    XMAPjoint
## X     1   2   3
##   1 336  21   0
##   2  17 298  24
##   3   0  18 286
```

```r
plot(VT$loglik[-1], type = "l", xlab = "Iteration", ylab = "Log-Likelihood",
    lwd = 2)
```



13 Εφαρμογή του Δειγματολήπτη Gibbs σε ΚΜΜ

Θεωρούμε ότι οι γραμμές του πίνακα πιθανοτήτων μετάβασης είναι a priori ανεξάρτητες με δεσμευμένα συζυγείς Dirichlet(α) prior και συναρτήσεις πυκνότητας πιθανότητας:

$$f(p_k) = \frac{\Gamma(K\alpha)}{[\Gamma(\alpha)]^K} \prod_{\ell=1}^{K} p_{k,\ell}^{\alpha-1} \propto \prod_{\ell=1}^{K} p_{k,\ell}^{\alpha-1}.$$

Τότε, η δεσμευμένη posterior κατανομή του πίνακα πιθανοτήτων μετάβασης P δίνεται από τη σχέση:

$$\begin{aligned} f(P \mid x) &\propto \prod_{k=1}^{K} f(p_k) \prod_{i=2}^{n} \mathbb{P}(X_i = x_i \mid X_{i-1} = x_{i-1}, P) \\ &\propto \prod_{k=1}^{K} \prod_{\ell=1}^{K} \left[p_{k,\ell}^{\alpha-1} \prod_{i=2}^{n} p_{k,\ell}^{\mathbb{1}_{\{x_{i-1}=k, x_i=\ell\}}} \right] = \prod_{k=1}^{K} \prod_{\ell=1}^{K} p_{k,\ell}^{M_{k,\ell}+\alpha-1}, \end{aligned}$$

δηλαδή οι γραμμές $p_k \mid x \sim \text{Dirichlet}\big(M_{k,1}+\alpha, M_{k,2}+\alpha, \dots, M_{k,K}+\alpha\big)$ είναι a posteriori ανεξάρτητες, όπου $M_{k,\ell} = \sum_{i=2}^{n} \mathbb{1}_{\{x_{i-1}=k, x_i=\ell\}}$ για $k, \ell = 1, 2, \dots, K$. Ως ειδική περίπτωση μπορούμε να θεωρήσουμε την prior του Jeffreys για τις γραμμές του πίνακα πιθανοτήτων μετάβασης, η οποία προκύπτει για $\alpha = 0.5$.

Έπειτα, θεωρούμε τη δεσμευμένα συζυγή Gamma(β, λ) prior για την παράμετρο θ_k με συνάρτηση πυκνότητας πιθανότητας:

$$f(\theta_k) = \frac{\lambda^\beta}{\Gamma(\beta)} \theta_k^{\beta-1} e^{-\lambda\theta_k} \propto \theta_k^{\beta-1} e^{-\lambda\theta_k}.$$

Τότε, η δεσμευμένη posterior κατανομή της παραμέτρου θ_k δίνεται από τη σχέση:

$$f(\theta_k \mid x, y) \propto f(\theta_k) \prod_{i=1}^{n} f(y_i \mid x_i, \theta_k) \propto \theta_k^{\beta-1} e^{-\lambda\theta_k} \prod_{i=1}^{n} \left(e^{-\theta_k} \frac{\theta_k^{y_i}}{y_i!} \right)^{\mathbb{1}_{\{x_i=k\}}} \propto \theta_k^{S_k+\beta-1} e^{-(N_k+\lambda)\theta_k},$$

δηλαδή $\theta_k \mid x, y \sim \text{Gamma}(S_k+\beta, N_k+\lambda)$, όπου $S_k = \sum_{i=1}^{n} y_i \mathbb{1}_{\{x_i=k\}}$ για $k = 1, 2, \dots, K$. Ως ειδική περίπτωση μπορούμε να θεωρήσουμε τη μη-ορθή prior του Jeffreys για την παράμετρο θ_k, η οποία δίνεται από τη σχέση $f(\theta_k) \propto \theta_k^{-0.5}$ και προκύπτει για $\beta = 0.5$, $\lambda = 0$.

Τέλος, γνωρίζουμε ότι η δεσμευμένη posterior κατανομή της κρυφής Μαρκοβιανής αλυσίδας δίνεται από τη σχέση:

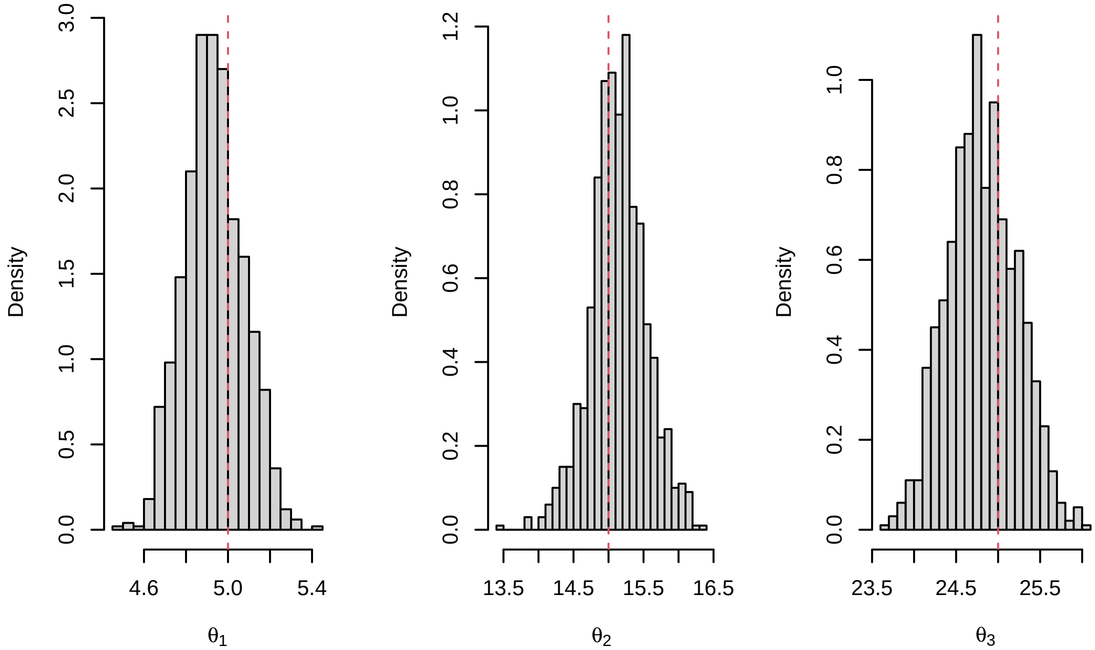
$$\mathbb{P}(X = x \mid y, \vartheta) = \phi_{n|n}(x_n) \prod_{i=1}^{n-1} B_i(x_i, x_{i+1}).$$

Επομένως, πρώτα υλοποιούμε τον αλγόριθμο Forward Filtering για να υπολογίσουμε τις filtering κατανομές $\phi_{i|i}(\cdot)$ για $i = 1, 2, \dots, n$. Έπειτα, προσομοιώνουμε την τελική κατάσταση X_n της κρυφής Μαρκοβιανής αλυσίδας από την τελευταία filtering κατανομή $\phi_{n|n}(\cdot)$. Τέλος, προσομοιώνουμε την κατάσταση X_i από την backward κατανομή μετάβασης $B_i(\cdot, x_{i+1}) = \phi_{i|i}(\cdot) p_{\cdot, x_{i+1}}$ για $i = n-1, n-2, \dots, 1$.

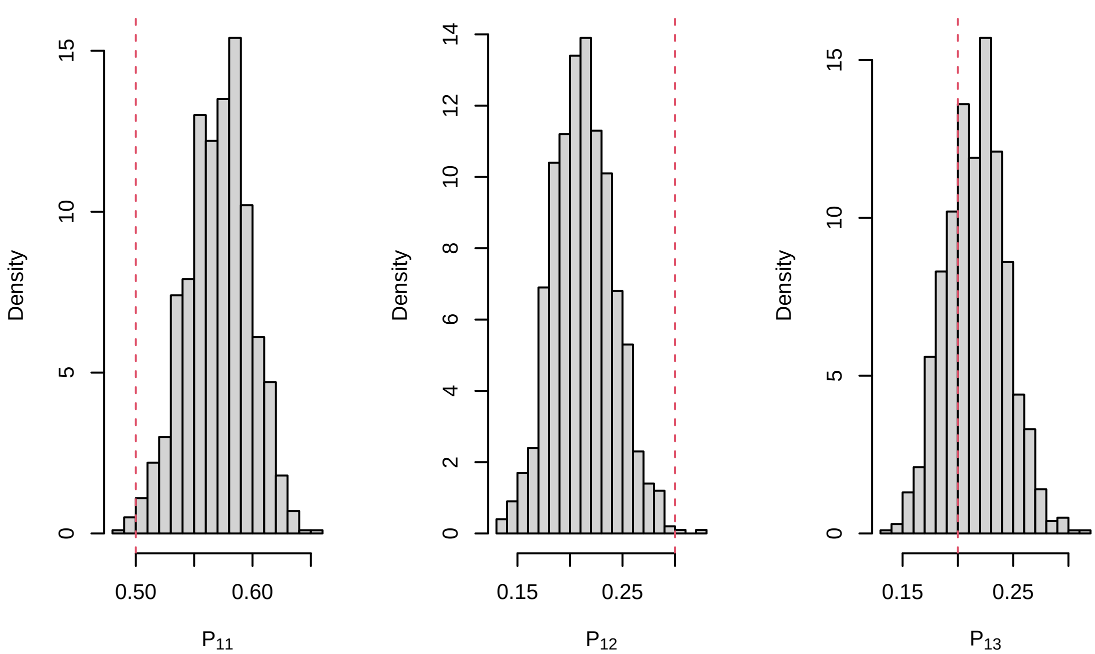
```
backward = function(filter, P) {
    n = dim(filter)[1]
    K = dim(P)[1]
    X = numeric(n)
    X[n] = sample(K, 1, prob = filter[n, ])
```

```
    for (i in (n - 1):1) {
        logB = log(filter[i, ]) + log(P[, X[i + 1]])
        B = exp(logB - max(logB))
        X[i] = sample(K, 1, prob = B)
    }
    return(X)
}

MCMCpoisHMM = function(Y, theta0, p, P0, alpha, beta, lambda, niter, nburn) {
    n = length(Y)
    K = length(theta)
    theta = matrix(0, niter, K)
    P = array(0, c(K, K, niter))
    X = matrix(0, niter, n)
    theta[1, ] = theta0
    P[, , 1] = P0
    X[1, ] = backward(forward(Y, theta[1, ], p, P[, , 1])$filter, P[, , 1])
    for (j in 2:niter) {
        x = factor(X[j - 1, ], levels = 1:K)
        M = aggregate(numeric(n - 1), data.frame(x[-n], x[-1]), length, drop = FALSE)[,
            3]
        P[, , j] = rgamma(K^2, M + alpha)
        P[, , j] = P[, , j]/rowSums(P[, , j])
        N = table(x)
        S = aggregate(Y ~ x, FUN = sum, drop = FALSE)[, 2]
        S[is.na(S)] = 0
        theta[j, ] = rgamma(K, S + beta, N + lambda)
        X[j, ] = backward(forward(Y, theta[j, ], p, P[, , j])$filter, P[, ,
            j])
        I = order(theta[j, ])
        theta[j, ] = theta[j, I]
        P[, , j] = P[I, I, j]
        X[j, ] = I[X[j, ]]
    }
    return(list(theta = theta[-(1:nburn), ], P = P[, , -(1:nburn)], X = X[-(1:nburn),
        ]))
}

posterior = MCMCpoisHMM(Y, mean(Y) + sd(Y) * ((1 - K)/2):((K - 1)/2), p, matrix(1,
    K, K)/K, 0.5, 0.5, 0, 2000, 1000)
par(mfrow = c(1, 3))
hist(posterior$theta[, 1], "FD", freq = FALSE, main = NA, xlab = expression(theta[1]))
abline(v = theta[1], col = 2, lty = 2)
```

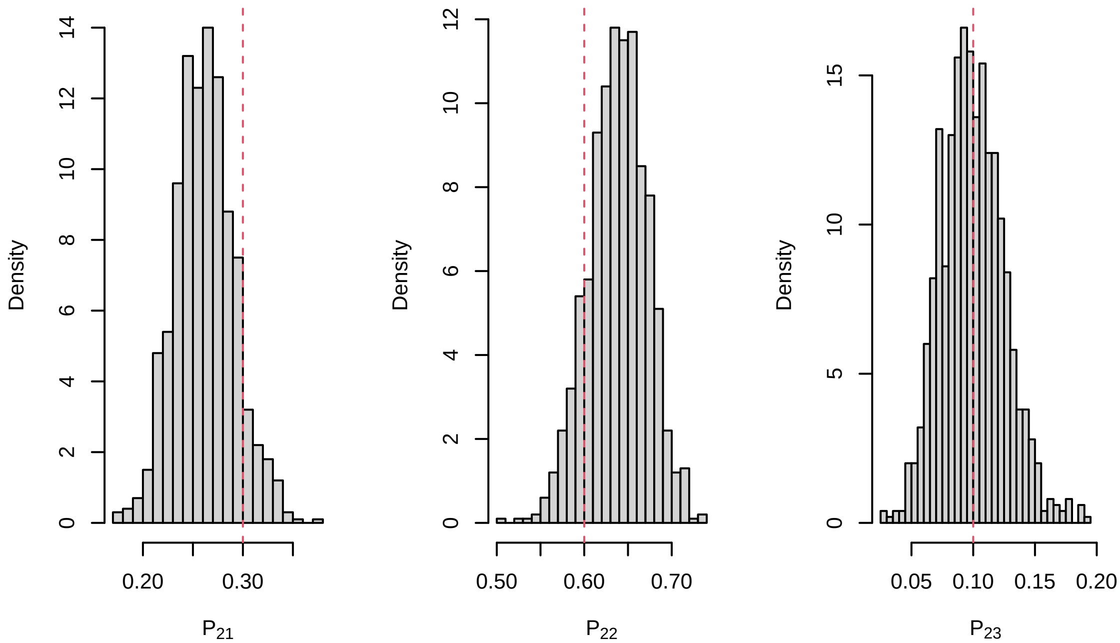
```
hist(posterior$theta[, 2], "FD", freq = FALSE, main = NA, xlab = expression(theta[2]))
abline(v = theta[2], col = 2, lty = 2)
hist(posterior$theta[, 3], "FD", freq = FALSE, main = NA, xlab = expression(theta[3]))
abline(v = theta[3], col = 2, lty = 2)
```



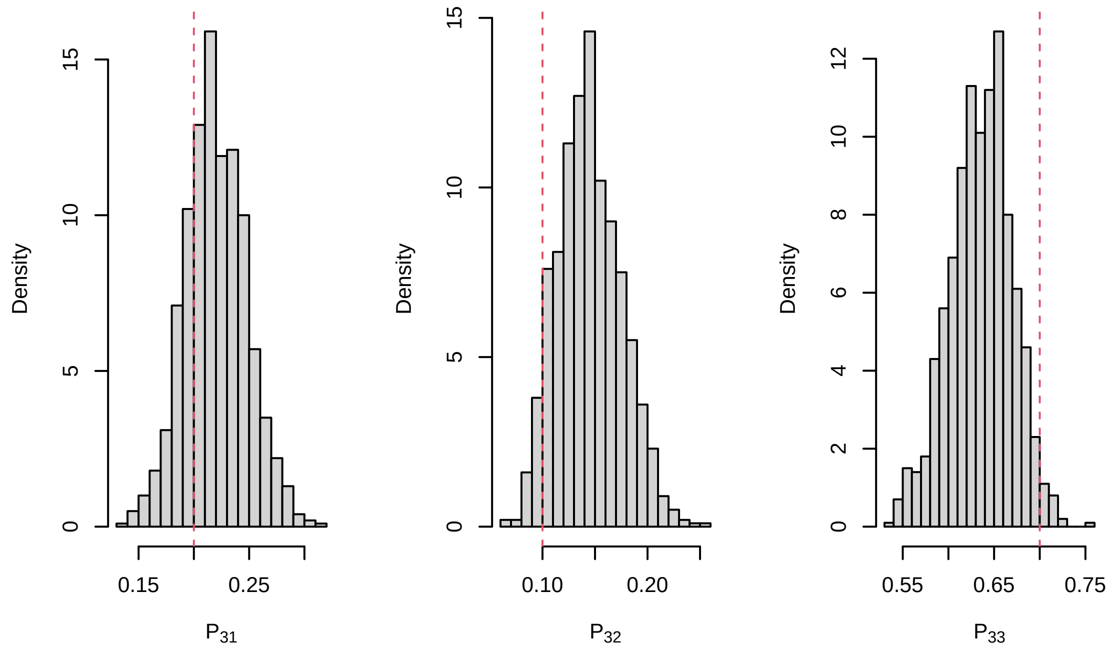

```
hist(posterior$P[1, 1, ], "FD", freq = FALSE, main = NA, xlab = expression(P[11]))
abline(v = P[1, 1], col = 2, lty = 2)
hist(posterior$P[1, 2, ], "FD", freq = FALSE, main = NA, xlab = expression(P[12]))
abline(v = P[1, 2], col = 2, lty = 2)
hist(posterior$P[1, 3, ], "FD", freq = FALSE, main = NA, xlab = expression(P[13]))
abline(v = P[1, 3], col = 2, lty = 2)
```



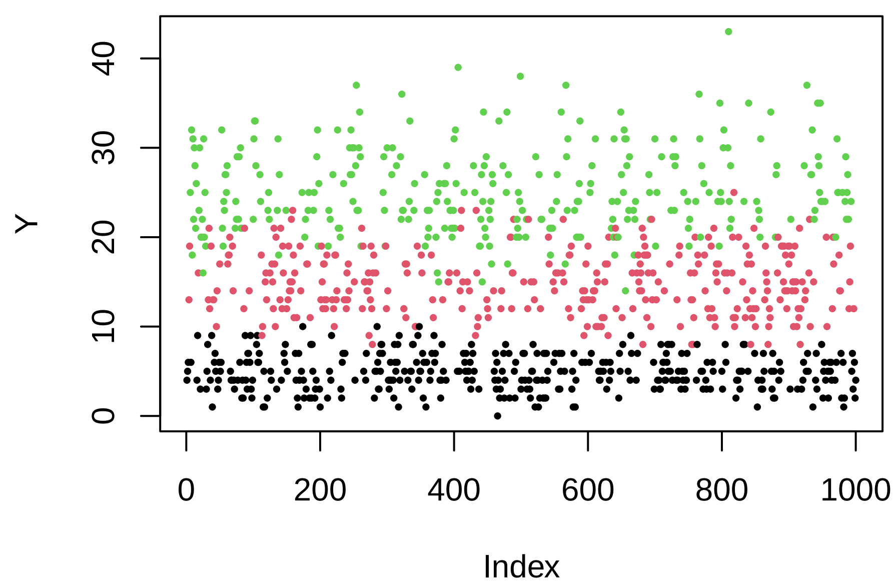
```r
hist(posterior$P[2, 1, ], "FD", freq = FALSE, main = NA, xlab = expression(P[21]))
abline(v = P[2, 1], col = 2, lty = 2)
hist(posterior$P[2, 2, ], "FD", freq = FALSE, main = NA, xlab = expression(P[22]))
abline(v = P[2, 2], col = 2, lty = 2)
hist(posterior$P[2, 3, ], "FD", freq = FALSE, main = NA, xlab = expression(P[23]))
abline(v = P[2, 3], col = 2, lty = 2)
```




```r
hist(posterior$P[3, 1, ], "FD", freq = FALSE, main = NA, xlab = expression(P[31]))
abline(v = P[3, 1], col = 2, lty = 2)
hist(posterior$P[3, 2, ], "FD", freq = FALSE, main = NA, xlab = expression(P[32]))
abline(v = P[3, 2], col = 2, lty = 2)
hist(posterior$P[3, 3, ], "FD", freq = FALSE, main = NA, xlab = expression(P[33]))
abline(v = P[3, 3], col = 2, lty = 2)
```




```
XMAP = apply(posterior$X, 2, function(x) {
    which.max(table(factor(x, levels = 1:K)))
})
plot(Y, col = XMAP, pch = 16, cex = 0.5)
```




```
table(X, XMAP)
```

```
##    XMAP
## X     1   2   3
##   1 347  10   0
##   2  17 305  17
##   3   0  21 283
```