ISABELLE MICHAUD

# Application de l'algorithme EM au modèle des risques concurrents avec causes de panne masquées

Mémoire présenté
à la Faculté des études supérieures de l'Université Laval
dans le cadre du programme de maîtrise en statistique
pour l'obtention du grade de maître ès sciences (M.Sc.)

FACULTÉ DES SCIENCES ET DE GÉNIE
UNIVERSITÉ LAVAL
QUÉBEC

Août 2005

# Résumé

Dans un modèle de durées de vie avec des risques concurrents, les systèmes peuvent tomber en panne dans le temps. Ces pannes sont dues à une cause parmi plusieurs possibles et il arrive parfois que celle-ci soit inconnue. C'est alors qu'on peut faire appel à l'algorithme EM pour calculer les estimateurs du maximum de vraisemblance. Cette technique utilise la fonction de vraisemblance des données complètes pour trouver les estimateurs même si les données observées sont incomplètes.

Pour les systèmes ayant leur cause de panne inconnue, on peut en prendre un échantillon pour une inspection plus approfondie qui dévoilera les vraies causes de panne. Cette étape peut améliorer l'estimation des probabilités de masque et des fonctions de risque spécifiques aux causes de panne. Après avoir expliqué la théorie de l'algorithme EM, le modèle des risques concurrents, ainsi que les travaux réalisés sur le sujet, on étudie l'impact qu'a sur les estimateurs le fait de ne pas envoyer un échantillon des systèmes masqués à un examen approfondi qui permettrait de trouver la vraie cause de panne.

# Avant-propos

Tout d'abord, je tiens à remercier mon directeur de recherche, Thierry Duchesne, pour son dévouement et sa grande disponibilité. Il sait rester quelqu'un de simple et d'accessible. C'est ce qui fait de lui un excellent professeur et une personne exceptionnelle.

Merci aussi à la très belle équipe de professeurs et de professionnels en statistique de l'Université Laval. Un merci tout spécial à Jean-Claude Massé qui m'a permis de travailler avec lui dans les cours de Statistique-Mathématique I et à l'équipe du Service de Consultation Statistique qui m'a donné l'énorme chance de travailler au service d'aide aux étudiants chercheurs. Ce sont des expériences plus qu'enrichissantes qui m'ont fait énormément grandir en tant que statisticienne et en tant que personne.

Un merci plus personnel à mes parents qui m'ont toujours supporté dans mes choix et à Jean-Lou pour son amour et pour la confiance qu'il a toujours eue en moi.

# Table des matières

# Liste des tableaux

# Table des figures

# Chapitre 1

# Introduction

Dans les applications de la statistique, il arrive fréquemment que certaines valeurs soient manquantes dans les jeux de données utilisés par les praticiens. Lorsque cette situation survient, plusieurs méthodes sont mises à la disposition des statisticiens pour remplacer les données manquantes par des valeurs imputées. De plus, lorsque toutes les données sont présentes et que leur distribution est connue, certaines méthodes peuvent être utilisées pour estimer les paramètres de la distribution.

Dans cet ouvrage, une technique qui permet d'estimer les paramètres lorsque des données sont absentes est étudiée ; c'est l'algorithme EM. Ce dernier est un processus itératif qui utilise la distribution des données complètes pour calculer les estimateurs du maximum de vraisemblance lorsque les données observées sont incomplètes. Ce procédé se fait en deux étapes. La première est l'étape E, c'est-à-dire l'étape d'espérance. Elle consiste à prendre l'espérance conditionnelle de la fonction de log-vraisemblance des données complètes sachant les données observées. La deuxième étape est l'étape M, soit la maximisation de la log-vraisemblance obtenue à l'étape E. On trouve alors l'estimateur qui maximise l'équation trouvée sous l'étape E. Ces étapes sont répétées itérativement jusqu'à convergence et, on l'espère, l'obtention de l'estimateur du maximum de vraisemblance.

Dans le texte qui suit, on utilise l'algorithme EM pour résoudre un problème dans le contexte des données de survie. Plus spécifiquement, on s'intéresse aux risques concurrents, c'est-à-dire qu'on est en présence de systèmes qui peuvent tomber en panne de l'une des $J$ causes possibles. Aussi, un certain nombre de systèmes peuvent avoir une cause de panne masquée dans un sous-groupe des $J$ causes, c'est-à-dire qu'on ne connaît pas la cause de panne exacte, mais qu'on a tout de même l'information que la cause est dans un sous-groupe des $J$ causes. Dans ces cas, on se retrouve en présence de données

manquantes pour lesquelles on ne dispose que d'information partielle. De plus, certains des systèmes dont la cause de panne est masquée au départ peuvent être envoyés à une deuxième étape où une investigation plus profonde est effectuée afin de découvrir la cause exacte de panne. Plusieurs exemples de la vie courante sont utilisés dans la littérature pour illustrer le modèle des risques concurrents avec des données masquées. On n'a qu'à penser à l'exemple de la carcinogenèse de Dinse (1986) et à ceux du cancer du poumon et de la fiabilité des disques durs de Flehinger, Reiser et Yashchin (1998, 2002).

En tenant compte du contexte des risques concurrents et des données masquées, la fonction de vraisemblance pour les données complètes est déduite. À l'aide de cette fonction et de l'algorithme EM, on peut estimer plusieurs paramètres. En effet, on peut entre autres estimer les probabilités de masque qui représentent la probabilité que la cause de panne d'un système soit masquée dans le groupe $g$ sachant que le système est tombé en panne de la cause $j$. On peut aussi estimer la fonction de risque spécifique à la cause $j$ définie comme étant la probabilité de tomber en panne de la cause $j$ dans l'instant suivant $t$. Finalement, ces deux probabilités permettent de calculer la probabilité de diagnostic, c'est-à-dire la probabilité d'être tombé en panne de la cause $j$ étant donné que la cause est masquée dans le groupe $g$.

Dans ce mémoire, on s'intéresse également à la situation où aucun système masqué n'est envoyé à la deuxième étape. Dans ce cas, on veut savoir si les estimateurs demeurent précis et si l'algorithme EM va converger. Pour ce faire, on fait des simulations dans un contexte où il y a seulement deux causes de panne possibles et où les fonctions de risque spécifiques aux causes de panne sont constantes par paliers. On remarque que les estimateurs semblent moins précis lorsqu'il n'y a pas de données de deuxième étape. Par contre, plus la taille des échantillons augmente, plus les estimations sont précises. Aussi, lorsque les risques concurrents sont proportionnels, l'algorithme est beaucoup plus lent à converger et les estimations sont plus variables que lorsque les risques sont non-proportionnels. De plus, l'algorithme EM ne converge pas toujours vers le même point lorsqu'on change les valeurs initiales des paramètres à estimer. Il semble que cette situation survienne lorsque les estimateurs du maximum de vraisemblance ne sont pas identifiables. Des conditions d'identifiabilité sont trouvées lorsqu'on est en présence de deux causes de panne et de deux intervalles pour les fonctions de risque spécifiques.

Dans ce travail, on fera d'abord un rappel de la méthode du maximum de vraisemblance. Par la suite, la théorie de l'algorithme EM et le modèle des risques concurrents seront présentés. On poursuivra avec une description des travaux réalisés sur les causes de panne masquées. Finalement, on terminera avec une étude théorique et une étude par simulations de l'effet de l'absence des données de deuxième étape.

# Chapitre 2

# La méthode du maximum de vraisemblance

La méthode du maximum de vraisemblance est l'une des techniques les plus populaires pour estimer des paramètres, puisqu'en général, elle donne des estimateurs convergents et de faible variance. C'est la méthode utilisée dans l'application de l'algorithme EM. Un rappel de la méthode du maximum de vraisemblance pour des données complètement observables est le sujet abordé dans ce chapitre.

Soit $X_1, X_2, ..., X_n$, un échantillon aléatoire indépendant identiquement distribué d'une population ayant comme fonction de masse de probabilité ou de densité $f(x|\boldsymbol{\theta})$, où $\boldsymbol{\theta} = (\theta_1, ..., \theta_d)^T$. Casella et Berger (2002, p.516) donnent les cinq conditions de régularité suivantes concernant $f(x|\boldsymbol{\theta})$ :

1. Le paramètre est identifiable, c'est-à-dire que si $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$, alors $f(x|\boldsymbol{\theta}) \neq f(x|\boldsymbol{\theta}')$.
2. Les densités $f(x|\boldsymbol{\theta})$ ont un support commun et $f(x|\boldsymbol{\theta})$ est dérivable en $\boldsymbol{\theta}$.
3. L'espace paramétrique $\boldsymbol{\Theta}$ contient un ensemble ouvert dont la vraie valeur du paramètre $\boldsymbol{\theta}_0$ est un point intérieur.
4. Soit $\mathcal{X}$, l'espace échantillonnal de $x$. Pour tout $x \in \mathcal{X}$, $f(x|\boldsymbol{\theta})$ est dérivable trois fois par rapport à $\boldsymbol{\theta}$, la troisième dérivée est continue en $\boldsymbol{\theta}$ et $\int f(x|\boldsymbol{\theta})dx$ peut être dérivée trois fois sous le signe de l'intégrale.
5. Pour tout $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, il existe un vecteur $\boldsymbol{c} = (c_1, ..., c_d)^T$ avec $c_1, ..., c_d$ des nombres

positifs et une fonction $M(x)$ tels que pour tout $x \in \mathcal{X}$,

$$\left| \frac{\partial^3}{\partial \boldsymbol{\theta}^3} \ln \Big( f(x|\boldsymbol{\theta}) \Big) \right| \leq M(x), \quad \boldsymbol{\theta}_0 - \boldsymbol{c} < \boldsymbol{\theta} < \boldsymbol{\theta}_0 + \boldsymbol{c},$$

avec $E_{\boldsymbol{\theta}_0}[M(x)] < \infty$.

Sous ces conditions, l'estimateur du maximum de vraisemblance satisfait les théorèmes 2.1 à 2.3 donnés à la fin du chapitre. Aussi, la fonction de vraisemblance pour un tel échantillon est donnée par

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}) = f(\boldsymbol{x}|\boldsymbol{\theta}), \tag{2.1}$$

où $\boldsymbol{x} = (x_1, ..., x_n)^T$.

Le logarithme naturel de (2.1) est la fonction de log-vraisemblance, dénotée

$$l(\boldsymbol{\theta}|\boldsymbol{x}) = \ln \Big( L(\boldsymbol{\theta}|\boldsymbol{x}) \Big). \tag{2.2}$$

La dérivée de la fonction (2.2) est la fonction score

$$\boldsymbol{S}(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{\partial l(\boldsymbol{\theta}|\boldsymbol{x})}{\partial \boldsymbol{\theta}}, \tag{2.3}$$

où $\boldsymbol{S}(\boldsymbol{\theta}|\boldsymbol{x}) = (S_1(\boldsymbol{\theta}|\boldsymbol{x}), ..., S_d(\boldsymbol{\theta}|\boldsymbol{x}))^T$ avec

$$\boldsymbol{S}_i(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{\partial l(\boldsymbol{\theta}|\boldsymbol{x})}{\partial \boldsymbol{\theta}_i}, i = 1, ..., d. \tag{2.4}$$

L'estimateur du maximum de vraisemblance (EMV) $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$ est une valeur qui maximise la fonction de vraisemblance ou, de façon équivalente, la fonction de log-vraisemblance. Sous les conditions de régularité précédentes sur $f(x|\boldsymbol{\theta})$, l'EMV peut être obtenu en solutionnant les équations de score suivantes :

$$\frac{\partial L(\boldsymbol{\theta}|\boldsymbol{x})}{\partial \boldsymbol{\theta}} = \mathbf{0}, \tag{2.5}$$

ou

$$\boldsymbol{S}(\boldsymbol{\theta}|\boldsymbol{x}) = \mathbf{0}. \tag{2.6}$$

De plus, la matrice d'information observée est dénotée par

$$\boldsymbol{I}(\boldsymbol{\theta}|\boldsymbol{x}) = -\frac{\partial \boldsymbol{S}(\boldsymbol{\theta}|\boldsymbol{x})}{\partial \boldsymbol{\theta}^T} = -\frac{\partial^2 l(\boldsymbol{\theta}|\boldsymbol{x})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}, \tag{2.7}$$

où $\boldsymbol{\theta}^T = (\theta_1 \ \ldots \ \theta_d)$ est la transposée du vecteur $\boldsymbol{\theta}$.

La matrice d'information de Fisher est donnée par

$$\mathcal{I}(\boldsymbol{\theta}) = -E_{\boldsymbol{\theta}}\Big[\boldsymbol{I}(\boldsymbol{\theta}|\boldsymbol{X})\Big]. \tag{2.8}$$

De plus, les estimateurs du maximum de vraisemblance ont les propriétés suivantes :

**Théorème 2.1 (Propriété d'invariance des EMV)** *Si $\hat{\boldsymbol{\theta}}$ est l'EMV de $\boldsymbol{\theta}$, alors pour toute fonction $\tau(\boldsymbol{\theta})$, l'EMV de $\tau(\boldsymbol{\theta})$ est $\tau(\hat{\boldsymbol{\theta}})$.*

**Théorème 2.2 (Convergence des EMV)** *Soit $\hat{\boldsymbol{\theta}}$, l'EMV de $\boldsymbol{\theta}$. Sous les conditions de régularité sur $f(\boldsymbol{x}|\boldsymbol{\theta})$, pour tout $\epsilon > 0$ et pour tout $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,*

$$\lim_{n\to\infty} P_{\boldsymbol{\theta}}(|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}| \geq \epsilon) = 0.$$

*C'est-à-dire que $\hat{\boldsymbol{\theta}}$ est un estimateur convergent de $\boldsymbol{\theta}$.*

**Théorème 2.3 (Normalité asymptotique des EMV)** *Soit $\hat{\boldsymbol{\theta}}$, l'EMV de $\boldsymbol{\theta}$. Sous les conditions de régularité sur $f(\boldsymbol{x}|\boldsymbol{\theta})$,*

$$\sqrt{n}[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}] \rightarrow^D N(\mathbf{0}, \mathcal{I}^{-1}(\boldsymbol{\theta})).$$

*C'est-à-dire que $\hat{\boldsymbol{\theta}}$ est un estimateur asymptotiquement normal de $\boldsymbol{\theta}$.*

# Chapitre 3

# La théorie de l'algorithme EM

## 3.1 Introduction

D'après Dempster, Laird et Rubin (1977), l'algorithme EM est une approche générale qui fait un calcul itératif pour trouver des estimateurs du maximum de vraisemblance lorsque les données sont incomplètes. On l'appelle « l'algorithme EM » puisque chaque itération de l'algorithme consiste en une étape d'**E**spérance et une étape de **M**aximisation.

### 3.1.1 Notation

Soit $\boldsymbol{Y}$, le vecteur aléatoire correspondant aux données observées $\boldsymbol{y}$, ayant une fonction de densité dénotée $f(\boldsymbol{y}|\boldsymbol{\theta})$, où $\boldsymbol{\theta} = (\theta_1, ..., \theta_d)^T$ est un vecteur de paramètres inconnus dans l'espace $\boldsymbol{\Theta}$.

Le vecteur des valeurs observées $\boldsymbol{y}$ est incomplet ; c'est-à-dire que certaines de ses données sont manquantes. Si la situation était idéale, toutes les données seraient présentes. Dans ce cas, ce serait le vecteur $\boldsymbol{x}$ qui serait observé. Mais dans les cas qui nous intéressent, c'est $\boldsymbol{y}$ qui est observé et ce dernier a des valeurs manquantes qui sont

contenues dans le vecteur $\boldsymbol{z}$. Donc si on ajoutait le vecteur $\boldsymbol{z}$ au vecteur $\boldsymbol{y}$, toutes les données seraient présentes et ainsi, le vecteur $\boldsymbol{x}$ serait formé. Par exemple, admettons que 3 dés sont lancés et que les résultats de la face gagnante sont notés. Le vecteur $\boldsymbol{x}$ serait formé des 3 valeurs obtenues par les dés, soit par exemple $\boldsymbol{x} = (1, 5, 2)$. Par contre, si le deuxième dé ne pouvait être lu, alors on observerait $\boldsymbol{y} = (1, ., 2)$ et donc $\boldsymbol{z} = (5)$.

Au lieu d'observer le vecteur des données complètes $\boldsymbol{x}$ dans $\mathcal{X}$, c'est plutôt le vecteur des données incomplètes $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})$ dans $\mathcal{Y}$ qui est observé. $\mathcal{X}$ et $\mathcal{Y}$ sont des espaces échantillonnals et il y a plusieurs valeurs dans $\mathcal{X}$ pour une valeur dans $\mathcal{Y}$. On définit ainsi $\mathcal{X}(\boldsymbol{y}) = \{\boldsymbol{x} : \boldsymbol{y}(\boldsymbol{x}) = \boldsymbol{y}\}$. Dans l'exemple sur les dés mentionné plus haut, on aurait

$$\mathcal{X}(\boldsymbol{y}) = \{(1, 1, 2), (1, 2, 2), (1, 3, 2), (1, 4, 2), (1, 5, 2), (1, 6, 2)\}, \tag{3.1}$$

soit toutes les valeurs possibles que les dés peuvent prendre lorsque le deuxième dé ne peut être lu et que le premier et le dernier dé ont respectivement une valeur de 1 et de 2.

Même lorsque c'est le vecteur $\boldsymbol{x}$ qui est observé, c'est-à-dire lorsque toutes les données sont présentes, l'algorithme EM peut être utilisé pour faciliter le calcul de l'estimateur du maximum de vraisemblance. En effet, moins il y a de données manquantes, plus l'estimation du maximum de vraisemblance par l'algorithme EM sera simple.

### 3.1.2 Vraisemblance pour les données observées

Soit $f_c(\boldsymbol{x}|\boldsymbol{\theta})$, la fonction de densité du vecteur aléatoire $\boldsymbol{X}$ correspondant au vecteur de données complètes $\boldsymbol{x}$. Alors la fonction de log-vraisemblance qui pourrait être formée pour $\boldsymbol{\theta}$ si $\boldsymbol{x}$ était complètement observable est donnée par

$$l_c(\boldsymbol{\theta}|\boldsymbol{x}) = \ln\Big(L_c(\boldsymbol{\theta}|\boldsymbol{x})\Big) = \ln\Big(f_c(\boldsymbol{x}|\boldsymbol{\theta})\Big). \tag{3.2}$$

Aussi, McLachlan et Krishnan (1997, p.22) donnent la relation suivante pour la vraisemblance sous les données observées $\boldsymbol{y}$ :

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\theta}) = \int_{\mathcal{X}(\boldsymbol{y})} f_c(\boldsymbol{x}|\boldsymbol{\theta})dx. \tag{3.3}$$

C'est-à-dire que le principe pour calculer $L(\boldsymbol{\theta}|\boldsymbol{y})$ est que, pour toutes les valeurs manquantes de $\boldsymbol{y}$, on somme sur toutes leurs valeurs possibles dans $\mathcal{X}(\boldsymbol{y})$. Dans l'illustration

des dés, la fonction de vraisemblance pour les données observées serait donc donnée par

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{x_2=1}^{6} f_c\Big((1, x_2, 2)|\boldsymbol{\theta}\Big).$$

### 3.1.3 Les étapes E et M

L'algorithme EM tente de résoudre le problème suivant : sachant qu'un échantillon de $\boldsymbol{Y}$ est observé, mais que les $\boldsymbol{X}$ correspondants ne sont pas observés entièrement (ou cachés), il faut trouver l'estimateur du maximum de vraisemblance $\hat{\boldsymbol{\theta}}$ qui maximise $l(\boldsymbol{\theta}|\boldsymbol{y}) = \ln(f(\boldsymbol{y}|\boldsymbol{\theta}))$. L'approche de l'algorithme EM envers ce problème est de résoudre la fonction score pour les données incomplètes, $\boldsymbol{S}(\boldsymbol{\theta}|\boldsymbol{y}) = \partial l(\boldsymbol{\theta}|\boldsymbol{y})/\partial\boldsymbol{\theta}$, indirectement en faisant des itérations à l'aide de la fonction de log-vraisemblance pour les données complètes donnée par l'équation (3.2). Mais puisque les éléments de cette dernière équation ne sont pas tous observés, elle est remplacée par son espérance conditionnelle sachant $\boldsymbol{y}$, en utilisant le $\boldsymbol{\theta}$ de l'itération en cours.

Plus spécifiquement, MacLachlan et Krishnan (1997, p.22) présentent les étapes E et M de cet algorithme comme suit :

Soit $\boldsymbol{\theta}^{(0)}$, une valeur initiale de $\boldsymbol{\theta}$ choisie arbitrairement, à laquelle l'algorithme débute. Alors à la première itération, l'étape E (Espérance) de l'algorithme EM se calcule comme suit :

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)}) = E_{\boldsymbol{\theta}^{(0)}}[l_c(\boldsymbol{\theta}|\boldsymbol{X})|\boldsymbol{y}]. \tag{3.4}$$

Ensuite, l'étape M (Maximisation) maximise (3.4) par rapport à $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. En fait, $\boldsymbol{\theta}^{(1)}$ est choisi selon l'inéquation $Q(\boldsymbol{\theta}^{(1)}|\boldsymbol{\theta}^{(0)}) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)})$ pour tout $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

À la deuxième itération, les étapes E et M sont refaites, mais cette fois-ci avec $\boldsymbol{\theta}^{(1)}$ au lieu de $\boldsymbol{\theta}^{(0)}$.

Voici donc les étapes E et M pour l'itération $(b+1)$ :

1. **Étape E** : Calculer $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(b)})$, où

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(b)}) = E_{\boldsymbol{\theta}^{(b)}}[l_c(\boldsymbol{\theta}|\boldsymbol{X})|\boldsymbol{y}]. \tag{3.5}$$

2. **Étape M** : Choisir $\boldsymbol{\theta}^{(b+1)}$ qui est une valeur de $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ qui maximise (3.5), c'est-à-dire qui est telle que $Q(\boldsymbol{\theta}^{(b+1)}|\boldsymbol{\theta}^{(b)}) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(b)})$ pour tout $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Ces deux étapes sont répétées jusqu'à ce que la différence entre la fonction de vraisemblance de l'itération $(b+1)$ et celle de l'itération $(b)$ ne change pratiquement plus (dans les cas où la suite de $L(\boldsymbol{\theta}^{(b)}|\boldsymbol{y})$ converge vers un point stationnaire) :

$$L(\boldsymbol{\theta}^{(b+1)}|\boldsymbol{y}) - L(\boldsymbol{\theta}^{(b)}|\boldsymbol{y}) \leq \varepsilon,$$

où $\varepsilon$ est une valeur arbitraire positive très près de zéro.

Il sera démontré à la section 3.3 que $L(\boldsymbol{\theta}|\boldsymbol{y})$ ne décroît pas après une itération, c'est-à-dire que $L(\boldsymbol{\theta}^{(b+1)}|\boldsymbol{y}) \geq L(\boldsymbol{\theta}^{(b)}|\boldsymbol{y})$, avec l'égalité survenant seulement aux points stationnaires de $L(\boldsymbol{\theta}|\boldsymbol{y})$. La convergence de la suite des fonctions de vraisemblance vers un point stationnaire est donc obtenue lorsque cette suite est bornée par le haut.

## 3.2 L'algorithme EM pour les familles exponentielles

Si la distribution des données complètes est un membre de la famille exponentielle, alors l'algorithme EM est beaucoup plus simple à calculer.

McLachlan et Krishnan (1997, p.26) définissent la famille exponentielle comme la famille des distributions dont la fonction de densité/probabilité est de la forme suivante :

$$f_c(\boldsymbol{x}|\boldsymbol{\theta}) = b(\boldsymbol{x}) \exp\{\boldsymbol{c}^T(\boldsymbol{\theta})\boldsymbol{t}(\boldsymbol{x})\}/a(\boldsymbol{\theta}), \tag{3.6}$$

où $\boldsymbol{t}(\boldsymbol{x})$ est une statistique exhaustive de dimension $k \times 1$ $(k \geq d)$, $\boldsymbol{c}(\boldsymbol{\theta})$ est un vecteur de dimension $k \times 1$ qui est fonction du vecteur de paramètres $\boldsymbol{\theta}$ de dimension $d \times 1$ et $a(\boldsymbol{\theta})$ et $b(\boldsymbol{x})$ sont des fonctions scalaires.

Si $k = d$ et que le Jacobien de $\boldsymbol{c}(\boldsymbol{\theta})$ est de rang complet, alors $f_c(\boldsymbol{x}|\boldsymbol{\theta})$ appartient à la famille exponentielle régulière. Si $f_c(\boldsymbol{x}|\boldsymbol{\theta})$ appartient à la famille exponentielle régulière de forme canonique, alors $\boldsymbol{c}(\boldsymbol{\theta}) = \boldsymbol{\theta}$ et

$$f_c(\boldsymbol{x}|\boldsymbol{\theta}) = b(\boldsymbol{x}) \exp\{\boldsymbol{\theta}^T\boldsymbol{t}(\boldsymbol{x})\}/a(\boldsymbol{\theta}). \tag{3.7}$$

Dans ce cas, la fonction de log-vraisemblance a la forme suivante :

$$l_c(\boldsymbol{\theta}|\boldsymbol{x}) = \ln\Big(b(\boldsymbol{x})\Big) - \ln\Big(a(\boldsymbol{\theta})\Big) + \boldsymbol{\theta}^T\boldsymbol{t}(\boldsymbol{x}). \tag{3.8}$$

On voit bien que maximiser (3.8) par rapport à $\boldsymbol{\theta}$ revient au même que de maximiser $l_c(\boldsymbol{\theta}|\boldsymbol{t}(\boldsymbol{x})) = -\ln(a(\boldsymbol{\theta})) + \boldsymbol{\theta}^T\boldsymbol{t}(\boldsymbol{x})$ qui ne dépend de $\boldsymbol{x}$ seulement que par $\boldsymbol{t}(\boldsymbol{x})$.

On a aussi que

$$\begin{aligned} E_{\boldsymbol{\theta}}\Big[l_c\big(\boldsymbol{\theta}|\boldsymbol{X}\big)\Big] &= E_{\boldsymbol{\theta}}\Big[l_c\big(\boldsymbol{\theta}|\boldsymbol{t}(\boldsymbol{X})\big)\Big] \\ &= -\ln\big(a(\boldsymbol{\theta})\big) + \boldsymbol{\theta}^T E_{\boldsymbol{\theta}}\Big[\boldsymbol{t}(\boldsymbol{X})\Big] \\ &= l_c\big(\boldsymbol{\theta}|E_{\boldsymbol{\theta}}[\boldsymbol{t}(\boldsymbol{X})]\big). \end{aligned} \tag{3.9}$$

De plus, si on intègre par rapport à $\boldsymbol{x}$ des deux côtés de l'équation (3.7), on obtient que

$$\begin{aligned} \int_{-\infty}^{\infty} f_c(\boldsymbol{x}|\boldsymbol{\theta})d\boldsymbol{x} &= \int_{-\infty}^{\infty}\Big[b(\boldsymbol{x})\exp\Big\{\boldsymbol{\theta}^T\boldsymbol{t}(\boldsymbol{x})\Big\}/a(\boldsymbol{\theta})\Big]d\boldsymbol{x} \\ \Leftrightarrow \\ a(\boldsymbol{\theta}) &= \int_{-\infty}^{\infty} b(\boldsymbol{x})\exp\Big\{\boldsymbol{\theta}^T\boldsymbol{t}(\boldsymbol{x})\Big\}d\boldsymbol{x} \end{aligned} \tag{3.10}$$

En dérivant des deux côtés de (3.10) par rapport à $\boldsymbol{\theta}$, on a que

$$\begin{aligned} \frac{d}{d\boldsymbol{\theta}}a(\boldsymbol{\theta}) &= \frac{d}{d\boldsymbol{\theta}}\int_{-\infty}^{\infty} b(\boldsymbol{x})\exp\Big\{\boldsymbol{\theta}^T\boldsymbol{t}(\boldsymbol{x})\Big\}d\boldsymbol{x} \\ \Leftrightarrow \\ a(\boldsymbol{\theta})\frac{d}{d\boldsymbol{\theta}}\ln\Big(a(\boldsymbol{\theta})\Big) &= \int_{-\infty}^{\infty}\frac{d}{d\boldsymbol{\theta}}b(\boldsymbol{x})\exp\Big\{\boldsymbol{\theta}^T\boldsymbol{t}(\boldsymbol{x})\Big\}d\boldsymbol{x} \\ \Leftrightarrow \\ a(\boldsymbol{\theta})\frac{d}{d\boldsymbol{\theta}}\ln\Big(a(\boldsymbol{\theta})\Big) &= \int_{-\infty}^{\infty} b(\boldsymbol{x})\boldsymbol{t}(\boldsymbol{x})\exp\Big\{\boldsymbol{\theta}^T\boldsymbol{t}(\boldsymbol{x})\Big\}d\boldsymbol{x} \\ \Leftrightarrow \\ a(\boldsymbol{\theta})\frac{d}{d\boldsymbol{\theta}}\ln\Big(a(\boldsymbol{\theta})\Big) &= \int_{-\infty}^{\infty}\boldsymbol{t}(\boldsymbol{x})a(\boldsymbol{\theta})f_c(\boldsymbol{x}|\boldsymbol{\theta})d\boldsymbol{x} \qquad \text{(de (3.7))} \\ \Leftrightarrow \\ \frac{d}{d\boldsymbol{\theta}}\ln\Big(a(\boldsymbol{\theta})\Big) &= \int_{-\infty}^{\infty}\boldsymbol{t}(\boldsymbol{x})f_c(\boldsymbol{x}|\boldsymbol{\theta})d\boldsymbol{x} \\ \Leftrightarrow \\ \frac{d}{d\boldsymbol{\theta}}\ln\Big(a(\boldsymbol{\theta})\Big) &= E_{\boldsymbol{\theta}}[\boldsymbol{t}(\boldsymbol{x})], \end{aligned} \tag{3.11}$$

qui est l'espérance de la statistique exhaustive $\boldsymbol{t}(\boldsymbol{X})$. Les équations (3.9) et (3.11) sont utilisées dans les étapes E et M de l'algorithme EM pour les familles exponentielles régulières canoniques comme il sera mentionné ultérieurement.

Une autre propriété intéressante concernant les familles exponentielles régulières canoniques est que la matrice d'information espérée de $\boldsymbol{\theta}$ est égale à la matrice de

variance-covariance de $\boldsymbol{t}(\boldsymbol{X})$. McLachlan et Krishnan (1997, p.27) donnent une façon plus formelle d'écrire cette propriété :

$$\mathcal{I}_c(\boldsymbol{\theta}) = \text{cov}_{\boldsymbol{\theta}}[\boldsymbol{t}(\boldsymbol{X})]. \tag{3.12}$$

D'après Dempster, Laird et Rubin (1977) et McLachlan et Krishnan (1997, p.27), voici les étapes E et M de l'algorithme EM pour les familles exponentielles régulières de forme canonique :

1. **Étape E** : Calculer $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(b)})$ à partir de l'équation (3.9).

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(b)}) &= E_{\boldsymbol{\theta}^{(b)}}\Big[l_c(\boldsymbol{\theta}|\boldsymbol{X})|\boldsymbol{y}\Big] \\ &= -\ln\Big(a(\boldsymbol{\theta})\Big) + \boldsymbol{\theta}^T E_{\boldsymbol{\theta}^{(b)}}\Big[\boldsymbol{t}(\boldsymbol{X})|\boldsymbol{y}\Big]. \end{aligned} \tag{3.13}$$

   (Les termes qui n'impliquent pas $\boldsymbol{\theta}$ ont été ignorés.)

2. **Étape M** : Dériver (3.13) par rapport à $\boldsymbol{\theta}$ et poser cette équation égale à zéro. On trouve alors que $\boldsymbol{t}^{(b)} = E_{\boldsymbol{\theta}^{(b)}}[\boldsymbol{t}(\boldsymbol{X})|\boldsymbol{y}] = \frac{\partial}{\partial\boldsymbol{\theta}}[\ln(a(\boldsymbol{\theta}))]$. En utilisant l'équation (3.11), on doit donc résoudre l'équation suivante pour déterminer $\boldsymbol{\theta}^{(b+1)}$ :

$$E_{\boldsymbol{\theta}}\Big(\boldsymbol{t}(\boldsymbol{X})\Big) = \boldsymbol{t}^{(b)}. \tag{3.14}$$

   Si l'équation (3.14) peut être résolue algébriquement pour $\boldsymbol{\theta}^{(b+1)} \in \boldsymbol{\Theta}$, alors la solution est unique. Sinon, le $\boldsymbol{\theta}^{(b+1)}$ maximum est sur la frontière de $\boldsymbol{\Theta}$.

## 3.3 Monotonocité de l'algorithme EM

Dans cette section, il est démontré que la fonction de vraisemblance pour les données incomplètes, $L(\boldsymbol{\theta}|\boldsymbol{y})$, ne décroît pas après une itération de l'algorithme EM, c'est-à-dire que

$$L(\boldsymbol{\theta}^{(b+1)}|\boldsymbol{y}) \geq L(\boldsymbol{\theta}^{(b)}|\boldsymbol{y}),\ b = 0, 1, 2, ... \tag{3.15}$$

McLachlan et Krishnan (1997, p.83) donnent la fonction de densité conditionnelle de $\boldsymbol{X}$ sachant $\boldsymbol{Y} = \boldsymbol{y}$ suivante :

$$k(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) = \frac{f_c(\boldsymbol{x}|\boldsymbol{\theta})}{f(\boldsymbol{y}|\boldsymbol{\theta})}. \tag{3.16}$$

Alors la fonction de log-vraisemblance pour les données incomplètes peut être écrite comme suit :

$$\begin{aligned} l(\boldsymbol{\theta}|\boldsymbol{y}) = \ln\Big(L(\boldsymbol{\theta}|\boldsymbol{y})\Big) &= \ln\Big(f(\boldsymbol{y}|\boldsymbol{\theta})\Big) \\ &= \ln\Big(f_c(\boldsymbol{x}|\boldsymbol{\theta})/k(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})\Big) \qquad \text{(de 3.16)} \\ &= \ln\Big(f_c(\boldsymbol{x}|\boldsymbol{\theta})\Big) - \ln\Big(k(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})\Big) \\ &= l_c\Big(\boldsymbol{\theta}|\boldsymbol{x}\Big) - \ln\Big(k(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})\Big). \end{aligned} \tag{3.17}$$

L'espérance de l'équation (3.17) est prise par rapport à la distribution conditionnelle de $\boldsymbol{X}$ sachant $\boldsymbol{Y} = \boldsymbol{y}$, en utilisant la valeur de $\boldsymbol{\theta}^{(b)}$ pour $\boldsymbol{\theta}$. L'équation devient alors :

$$E_{\boldsymbol{\theta}}\Big[l(\boldsymbol{\theta}|\boldsymbol{y})|\boldsymbol{y}\Big] = l(\boldsymbol{\theta}|\boldsymbol{y}) = E_{\boldsymbol{\theta}^{(b)}}\Big[l_c(\boldsymbol{\theta}|\boldsymbol{X})|\boldsymbol{y}\Big] - E_{\boldsymbol{\theta}^{(b)}}\Big[\ln\Big(k(\boldsymbol{X}|\boldsymbol{y},\boldsymbol{\theta})\Big)|\boldsymbol{y}\Big]. \tag{3.18}$$

De l'équation (3.5), on a que

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(b)}) = E_{\boldsymbol{\theta}^{(b)}}\Big[l_c(\boldsymbol{\theta}|\boldsymbol{X})|\boldsymbol{y}\Big]$$

et en posant

$$H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(b)}) = E_{\boldsymbol{\theta}^{(b)}}\Big[\ln\Big(k(\boldsymbol{X}|\boldsymbol{y},\boldsymbol{\theta})\Big)|\boldsymbol{y}\Big],$$

alors l'équation (3.18) peut se réécrire de la façon suivante :

$$l(\boldsymbol{\theta}|\boldsymbol{y}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(b)}) - H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(b)}). \tag{3.19}$$

En utilisant l'équation (3.19), on trouve que

$$\begin{aligned} l(\boldsymbol{\theta}^{(b+1)}|\boldsymbol{y}) - l(\boldsymbol{\theta}^{(b)}|\boldsymbol{y}) &= \Big[Q(\boldsymbol{\theta}^{(b+1)}|\boldsymbol{\theta}^{(b)}) - H(\boldsymbol{\theta}^{(b+1)}|\boldsymbol{\theta}^{(b)})\Big] \\ &\quad -\Big[Q(\boldsymbol{\theta}^{(b)}|\boldsymbol{\theta}^{(b)}) - H(\boldsymbol{\theta}^{(b)}|\boldsymbol{\theta}^{(b)})\Big] \\ &= \Big[Q(\boldsymbol{\theta}^{(b+1)}|\boldsymbol{\theta}^{(b)}) - Q(\boldsymbol{\theta}^{(b)}|\boldsymbol{\theta}^{(b)})\Big] \\ &\quad -\Big[H(\boldsymbol{\theta}^{(b+1)}|\boldsymbol{\theta}^{(b)}) - H(\boldsymbol{\theta}^{(b)}|\boldsymbol{\theta}^{(b)})\Big]. \end{aligned} \tag{3.20}$$

La première soustraction du côté droit de l'égalité de l'équation (3.20) est positive puisque $\boldsymbol{\theta}^{(b+1)}$ est choisi tel que

$$Q(\boldsymbol{\theta}^{(b+1)}|\boldsymbol{\theta}^{(b)}) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(b)}) \text{ pour tout } \boldsymbol{\theta} \in \boldsymbol{\Theta}. \tag{3.21}$$

Ainsi, l'inégalité (3.15) est vraie si on a que

$$H(\boldsymbol{\theta}^{(b+1)}|\boldsymbol{\theta}^{(b)}) - H(\boldsymbol{\theta}^{(b)}|\boldsymbol{\theta}^{(b)}) \leq 0. \tag{3.22}$$

Pour tout $\boldsymbol{\theta}$,

$$\begin{aligned} H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(b)}) - H(\boldsymbol{\theta}^{(b)}|\boldsymbol{\theta}^{(b)}) &= E_{\boldsymbol{\theta}^{(b)}}\Big[\ln\Big(k(\boldsymbol{X}|\boldsymbol{y},\boldsymbol{\theta})\Big)|\boldsymbol{y}\Big] \\ &\quad - E_{\boldsymbol{\theta}^{(b)}}\Big[\ln\Big(k(\boldsymbol{X}|\boldsymbol{y},\boldsymbol{\theta}^{(b)})\Big)|\boldsymbol{y}\Big] \\ &= E_{\boldsymbol{\theta}^{(b)}}\Big[\ln\Big(k(\boldsymbol{X}|\boldsymbol{y},\boldsymbol{\theta})/k(\boldsymbol{X}|\boldsymbol{y},\boldsymbol{\theta}^{(b)})\Big)|\boldsymbol{y}\Big] \\ &\leq \ln\left(E_{\boldsymbol{\theta}^{(b)}}\Big[k(\boldsymbol{X}|\boldsymbol{y},\boldsymbol{\theta})/k(\boldsymbol{X}|\boldsymbol{y},\boldsymbol{\theta}^{(b)})\Big]|\boldsymbol{y}\right) \qquad (3.23) \\ &= \ln\left(\int_{\mathcal{X}(\boldsymbol{y})} \frac{k(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})}{k(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}^{(b)})} k(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}^{(b)})d\boldsymbol{x}\right) \\ &= \ln\left(\int_{\mathcal{X}(\boldsymbol{y})} k(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})d\boldsymbol{x}\right) \\ &= \ln(1) \\ &= 0. \qquad (3.24) \end{aligned}$$

L'inégalité en (3.23) est une conséquence directe de l'inégalité de Jensen et de la concavité de la fonction logarithmique.

On a donc montré l'inégalité (3.22) et par le fait même l'inégalité (3.15). Ainsi, $L(\boldsymbol{\theta}|\boldsymbol{y})$ ne décroît pas après une itération de l'algorithme EM. Cette fonction de vraisemblance sera strictement croissante si l'inégalité (3.21) est stricte. Ainsi, pour une suite bornée de valeurs de vraisemblances $\{L(\boldsymbol{\theta}|\boldsymbol{y})\}$, $L(\boldsymbol{\theta}|\boldsymbol{y})$ converge de façon monotone vers une certaine valeur $L^\star$, où $L^\star = L(\boldsymbol{\theta}^\star|\boldsymbol{y})$ pour un point $\boldsymbol{\theta}^\star$ pour lequel $\partial L(\boldsymbol{\theta}|\boldsymbol{y})/\partial\boldsymbol{\theta} = \mathbf{0}$ ou $\partial l(\boldsymbol{\theta}|\boldsymbol{y})/\partial\boldsymbol{\theta} = \mathbf{0}$. En d'autres mots, la fonction de vraisemblance augmente de façon monotone croissante à chaque itération de l'algorithme et atteint son sommet au point $L^\star$.

## 3.4 Convergence d'une suite EM vers une valeur stationnaire

Comme il a été vu à la section précédente, pour une suite bornée $\{L(\boldsymbol{\theta}|\boldsymbol{y})\}$, $L(\boldsymbol{\theta}|\boldsymbol{y})$ est monotone croissante et converge vers une certaine valeur $L^\star$. McLachlan et Krishnan (1997, p.85) mentionnent que, dans presque toutes les applications, $L^\star$ est une valeur stationnaire.

La plupart du temps, $L^\star$ est un maximum local. Par contre, si la suite EM $\{\boldsymbol{\theta}^{(b)}\}$ est prise dans un point stationnaire $\boldsymbol{\theta}^\star$ qui n'est ni un maximum local ni un maximum global (par exemple, un point de selle), une petite perturbation aléatoire de $\boldsymbol{\theta}$ en-dehors du point de selle $\boldsymbol{\theta}^\star$ éloignera l'algorithme EM de ce point de selle.

Si $L(\boldsymbol{\theta}|\boldsymbol{y})$ a plusieurs points stationnaires, la convergence de la suite EM vers un maximum local, un maximum global ou un point de selle dépendra de la valeur initiale $\boldsymbol{\theta}^{(0)}$. Cependant, si la fonction de vraisemblance est unimodale dans $\boldsymbol{\Theta}$, alors la suite EM convergera nécessairement vers l'unique estimateur du maximum de vraisemblance, peu importe le point de départ $\boldsymbol{\theta}^{(0)}$ utilisé.

Finalement, le théorème 3.1 de Wu (1983) est un résultat important sur la convergence d'une suite EM vers un point stationnaire et s'énonce comme suit :

**Théorème 3.1** *Supposons que $Q(\boldsymbol{\theta}|\boldsymbol{\theta}')$ est continue en $\boldsymbol{\theta}$ et en $\boldsymbol{\theta}'$. Alors tous les points limites de la suite $\{\boldsymbol{\theta}^{(b)}\}$ de l'algorithme EM sont des points stationnaires de $L(\boldsymbol{\theta}|\boldsymbol{y})$ et $L(\boldsymbol{\theta}^{(b)}|\boldsymbol{y})$ converge de façon monotone vers $L^*$ pour n'importe quel point stationnaire $\boldsymbol{\theta}^*$.*

## 3.5 Taux de convergence de l'algorithme EM

McLachlan et Krishnan (1997, p.105) expliquent que l'algorithme EM définit implicitement une fonction $\boldsymbol{\theta} \mapsto \boldsymbol{M}(\boldsymbol{\theta})$ d'un espace d'un paramètre $\boldsymbol{\theta}$, $\boldsymbol{\Theta}$, vers lui-même telle que chaque itération est définie par

$$\boldsymbol{\theta}^{(b+1)} = \boldsymbol{M}(\boldsymbol{\theta}^{(b)}), \quad b = 0, 1, ... \tag{3.25}$$

Si $\boldsymbol{\theta}^{(b)}$ converge vers un point $\boldsymbol{\theta}^*$ et $\boldsymbol{M}(\boldsymbol{\theta})$ est continue, alors $\boldsymbol{\theta}^*$ doit satisfaire

$$\boldsymbol{\theta}^* = \boldsymbol{M}(\boldsymbol{\theta}^*). \tag{3.26}$$

Dans le voisinage de $\boldsymbol{\theta}^*$, par un développement en série de Taylor de $\boldsymbol{\theta}^{(b+1)} = \boldsymbol{M}(\boldsymbol{\theta}^{(b)})$ autour du point $\boldsymbol{\theta}^{(b)} = \boldsymbol{\theta}^*$, on obtient de Meng et Rubin (1991) :

$$\boldsymbol{\theta}^{(b+1)} - \boldsymbol{\theta}^* \approx \boldsymbol{D}\boldsymbol{M}(\boldsymbol{\theta}^*)(\boldsymbol{\theta}^{(b)} - \boldsymbol{\theta}^*), \tag{3.27}$$

où $\boldsymbol{DM}(\boldsymbol{\theta}^*)$ est le Jacobien de dimension $d \times d$ pour $\boldsymbol{M}(\boldsymbol{\theta}) = (M_1(\boldsymbol{\theta}), ..., M_d(\boldsymbol{\theta}))$ évaluée à $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ dont le $(i,j)$ème élément $(DM(\boldsymbol{\theta}^*))_{ij}$ est donné par

$$r_{ij} = (DM(\boldsymbol{\theta}^*))_{ij} = \left(\frac{\partial M_j(\boldsymbol{\theta})}{\partial \theta_i}\right)\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}. \tag{3.28}$$

Ainsi, dans le voisinage de $\boldsymbol{\theta}^*$, l'algorithme EM est essentiellement une itération linéaire avec la matrice de taux $\boldsymbol{DM}(\boldsymbol{\theta}^*)$, puisque $\boldsymbol{DM}(\boldsymbol{\theta}^*)$ est différente de zéro. On dit alors que la matrice $\boldsymbol{DM}(\boldsymbol{\theta}^*)$ est la matrice du taux de convergence de l'algorithme EM.

McLachlan et Krishnan (1997, p.105) définissent une mesure du taux de convergence observé qui peut être déterminée par le taux de convergence global comme suit :

$$r = \lim_{b\to\infty} ||\boldsymbol{\theta}^{(b+1)} - \boldsymbol{\theta}^*||/||\boldsymbol{\theta}^{(b)} - \boldsymbol{\theta}^*||, \tag{3.29}$$

où $||.||$ est n'importe quelle norme sur l'espace $\mathcal{R}^d$. Sous certaines conditions de régularité,

$$r \equiv \text{ la plus grande valeur propre de } \boldsymbol{DM}^T(\boldsymbol{\theta}^*). \tag{3.30}$$

Ainsi, une grande valeur de $r$ implique que l'algorithme converge lentement.

## 3.6 Forces et faiblesses de l'algorithme EM

Dans cette section, l'objectif est de donner une idée du potentiel de l'algorithme EM comme outil utile pour l'estimation de paramètres dans des problèmes d'estimation statistique. Quelques critiques de cet algorithme seront aussi mentionnées. De plus, il sera comparé à ses compétiteurs, comme les méthodes de Newton-Raphson et du score de Fisher, car ceux-ci utilisent aussi des algorithmes itératifs pour trouver les EMV. McLachlan et Krishnan (1997, p.32) présentent quelques avantages de l'algorithme EM comparé à ces méthodes :

- L'algorithme EM est stable numériquement et la vraisemblance croît à chaque itération (sauf à un point fixe de l'algorithme).
- L'algorithme EM converge globalement sous certaines conditions. En effet, en partant d'un point arbitraire $\boldsymbol{\theta}^{(0)}$ dans l'espace du paramètre, la convergence se

fait presque toujours à un maximum local. Il peut arriver que ce ne soit pas le cas, mais cela arrive très rarement ; soit que le choix de $\boldsymbol{\theta}^{(0)}$ ait été très malchanceux ou encore qu'il y ait une pathologie locale dans la fonction de log-vraisemblance.

- L'algorithme EM est facilement mis en application parce qu'il s'appuie sur le calcul des données complètes. En effet, l'étape E ne prend que l'espérance sur la distribution conditionnelle des données complètes à chaque itération, tandis que l'étape M n'exige, pour sa part, que l'estimation du maximum de vraisemblance des données complètes à chaque itération, qui est souvent sous une forme simple.
- L'algorithme EM est souvent facile à programmer, puisque ni l'évaluation de sa vraisemblance des données observées ni celle de ses dérivées ne sont nécessaires.
- L'algorithme EM demande peu d'espace de stockage et peut généralement être utilisé sur un petit ordinateur. Par exemple, il n'a pas besoin d'emmagasiner la matrice d'information ni son inverse.
- Souvent, le problème des données complètes est un problème « classique », donc l'étape M peut souvent être résolue en utilisant des logiciels statistiques.
- Le travail analytique nécessaire est plus simple que celui des autres méthodes puisque seulement l'espérance conditionnelle de la log-vraisemblance pour les données complètes a besoin d'être maximisée. Il y a une certaine quantité de travail analytique à faire pour exécuter l'étape E, mais dans la plupart des applications, cette étape n'est pas compliquée.
- Le coût par itération étant généralement bas, un plus grand nombre d'itérations que les autres méthodes peut donc être exécuté par l'algorithme EM pour un coût donné.
- En observant la croissance monotone de la vraisemblance à chaque itération, il est facile de contrôler sa convergence et les erreurs de programmation.
- L'algorithme EM peut être utilisé pour fournir des valeurs estimées des données manquantes.

Voici maintenant quelques critiques de l'algorithme EM :

- L'algorithme EM n'a pas de procédure incluse qui pourrait produire la matrice de variance-covariance des paramètres estimés.
- L'algorithme EM peut converger lentement même pour les problèmes qui semblent inoffensifs. Il peut converger lentement aussi lorsqu'il y a beaucoup d'information manquante.
- Il n'est pas certain que l'algorithme EM convergera à un maximum global ou local lorsqu'il y a plusieurs maxima.
- Dans certains problèmes, l'étape E peut être analytiquement impossible à trouver.

## 3.7 Statistiques de score et matrices d'information

Dans cette section, on décrit les statistiques de score et les matrices d'information pour les données observées et pour les données complètes, ainsi que le principe de l'information manquante.

Les statistiques de score pour les données observées et pour les données complètes sont respectivement données par

$$\boldsymbol{S}(\boldsymbol{y}|\boldsymbol{\theta}) = \partial l(\boldsymbol{\theta}|\boldsymbol{y})/\partial\boldsymbol{\theta} \tag{3.31}$$

et

$$\boldsymbol{S}_c(\boldsymbol{x}|\boldsymbol{\theta}) = \partial l_c(\boldsymbol{\theta}|\boldsymbol{x})/\partial\boldsymbol{\theta}. \tag{3.32}$$

McLachlan et Krishnan (1997, p.100) montrent que la statistique de score pour les données observées peut être exprimée en fonction de celle pour les données complètes. Cette relation s'exprime de la façon suivante :

$$\boldsymbol{S}(\boldsymbol{y}|\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[\boldsymbol{S}_c(\boldsymbol{X}|\boldsymbol{\theta})|\boldsymbol{y}]. \tag{3.33}$$

Voici la démonstration de l'expression (3.33) :

$$\begin{aligned}
\boldsymbol{S}(\boldsymbol{y}|\boldsymbol{\theta}) &= \partial l(\boldsymbol{\theta}|\boldsymbol{y})/\partial\boldsymbol{\theta} && \text{(de (3.31))}\\
&= \partial \ln\Big(f(\boldsymbol{y}|\boldsymbol{\theta})\Big)/\partial\boldsymbol{\theta}\\
&= \frac{\partial f(\boldsymbol{y}|\boldsymbol{\theta})/\partial\boldsymbol{\theta}}{f(\boldsymbol{y}|\boldsymbol{\theta})}\\
&= \left[\int_{\mathcal{X}(\boldsymbol{y})} \frac{\partial f_c(\boldsymbol{x}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} d\boldsymbol{x}\right]/f(\boldsymbol{y}|\boldsymbol{\theta}) && \text{(de (3.3))}\\
&= \left[\int_{\mathcal{X}(\boldsymbol{y})} \frac{\partial f_c(\boldsymbol{x}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \frac{f_c(\boldsymbol{x}|\boldsymbol{\theta})}{f_c(\boldsymbol{x}|\boldsymbol{\theta})} d\boldsymbol{x}\right]/f(\boldsymbol{y}|\boldsymbol{\theta})\\
&= \int_{\mathcal{X}(\boldsymbol{y})} \frac{\partial \ln\Big(f_c(\boldsymbol{x}|\boldsymbol{\theta})\Big)}{\partial\boldsymbol{\theta}} \frac{f_c(\boldsymbol{x}|\boldsymbol{\theta})}{f(\boldsymbol{y}|\boldsymbol{\theta})} d\boldsymbol{x}\\
&= \int_{\mathcal{X}(\boldsymbol{y})} \frac{\partial \ln\Big(L_c(\boldsymbol{\theta}|\boldsymbol{x})\Big)}{\partial\boldsymbol{\theta}} k(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) d\boldsymbol{x} && \text{(de (3.16))}\\
&= E_{\boldsymbol{\theta}}\left[\frac{\partial l_c(\boldsymbol{\theta}|\boldsymbol{X})}{\partial\boldsymbol{\theta}}\Big|\boldsymbol{y}\right]\\
&= E_{\boldsymbol{\theta}}[\boldsymbol{S}_c(\boldsymbol{X}|\boldsymbol{\theta})|\boldsymbol{y}].
\end{aligned}$$

Ensuite, les matrices d'information observées, basées respectivement sur les données disponibles ($\boldsymbol{y}$) et complètes ($\boldsymbol{x}$) s'écrivent comme suit :

$$\boldsymbol{I}(\boldsymbol{\theta}|\boldsymbol{y}) = -\partial^2 \, l(\boldsymbol{\theta}|\boldsymbol{y}) \,/\, \partial\boldsymbol{\theta}\, \partial\boldsymbol{\theta}^T \tag{3.34}$$

et

$$\boldsymbol{I}_c(\boldsymbol{\theta}|\boldsymbol{x}) = -\partial^2 \, l_c(\boldsymbol{\theta}|\boldsymbol{x}) \,/\, \partial\boldsymbol{\theta}\, \partial\boldsymbol{\theta}^T. \tag{3.35}$$

Les matrices d'information de Fisher pour les données disponibles et complètes respectivement sont données par

$$\begin{aligned} \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) &= E_{\boldsymbol{\theta}}[\boldsymbol{S}(\boldsymbol{Y}|\boldsymbol{\theta})\, \boldsymbol{S}^T(\boldsymbol{Y}|\boldsymbol{\theta})] \\ &= E_{\boldsymbol{\theta}}[\boldsymbol{I}(\boldsymbol{\theta}|\boldsymbol{Y})] \end{aligned} \tag{3.36}$$

et

$$\begin{aligned} \boldsymbol{\mathcal{I}}_c(\boldsymbol{\theta}) &= E_{\boldsymbol{\theta}}[\boldsymbol{S}_c(\boldsymbol{X}|\boldsymbol{\theta})\, \boldsymbol{S}_c^T(\boldsymbol{X}|\boldsymbol{\theta})] \\ &= E_{\boldsymbol{\theta}}[\boldsymbol{I}_c(\boldsymbol{\theta}|\boldsymbol{X})]. \end{aligned} \tag{3.37}$$

Finalement, à partir de l'équation (3.17), on prend le négatif de la dérivée seconde par rapport à $\boldsymbol{\theta}$ des deux côtés de l'égalité. On obtient alors

$$\boldsymbol{I}(\boldsymbol{\theta}|\boldsymbol{y}) = \boldsymbol{I}_c(\boldsymbol{\theta}|\boldsymbol{x}) + \partial^2 \, \ln\Big(k(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})\Big)/\partial\boldsymbol{\theta}\, \partial\, \boldsymbol{\theta}^T. \tag{3.38}$$

En prenant l'espérance des deux côtés de (3.38) sur la distribution conditionnelle de $\boldsymbol{z}$ étant donné $\boldsymbol{y}$, McLachlan et Krishnan (1997, p.100) obtiennent le principe de l'information manquante résumé dans le théorème 3.2.

**Théorème 3.2 (Le principe de l'information manquante)** *Soit*

1. $\boldsymbol{\mathcal{I}}_c(\boldsymbol{\theta}|\boldsymbol{y}) = E_{\boldsymbol{\theta}}[\boldsymbol{I}_c(\boldsymbol{\theta}|\boldsymbol{X})|\boldsymbol{y}]$, *l'espérance conditionnelle de la matrice d'information complète sachant* $\boldsymbol{y}$,
2. $\boldsymbol{\mathcal{I}}_m(\boldsymbol{\theta}|\boldsymbol{y}) = -\,E_{\boldsymbol{\theta}}[\partial^2 \, \ln(k(\boldsymbol{X}|\boldsymbol{y},\boldsymbol{\theta})) \,/\, \partial\boldsymbol{\theta}\, \partial\boldsymbol{\theta}^T|\boldsymbol{y}]$, *la matrice d'information espérée pour* $\boldsymbol{\theta}$ *basée sur* $\boldsymbol{x}$ *conditionnant sur* $\boldsymbol{y}$. *Elle peut être vue comme « l'information manquante » puisqu'elle est une conséquence de l'observation du vecteur* $\boldsymbol{y}$ *seulement.*

*Alors,*

$$E_{\boldsymbol{\theta}}[\boldsymbol{I}(\boldsymbol{\theta}|\boldsymbol{y})|\boldsymbol{y}] = \boldsymbol{I}(\boldsymbol{\theta}|\boldsymbol{y}) = \boldsymbol{\mathcal{I}}_c(\boldsymbol{\theta}|\boldsymbol{y}) - \boldsymbol{\mathcal{I}}_m(\boldsymbol{\theta}|\boldsymbol{y}), \tag{3.39}$$

*c'est-à-dire que l'information observée peut être vue comme l'information complète moins l'information manquante.*

En prenant l'espérance des deux côtés de (3.39) sur la distribution de $\boldsymbol{Y}$, on obtient une équation analogue à (3.39) pour l'information espérée des données observées :

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = \boldsymbol{\mathcal{I}}_c(\boldsymbol{\theta}) - E_{\boldsymbol{\theta}}[\boldsymbol{\mathcal{I}}_m(\boldsymbol{\theta}|\boldsymbol{Y})]. \tag{3.40}$$

## 3.8 Matrice de variance-covariance des paramètres estimés

Tel que mentionné dans la section 3.6, une faiblesse de l'algoritme EM est de ne pas inclure une procédure calculant la matrice de variance-covariance des paramètres estimés. Dans ce chapitre, on montre comment cette matrice peut être calculée à l'aide des méthodes proposées par Louis (1982) et par Meng et Rubin (1991).

### 3.8.1 Méthode de Louis

La première méthode, celle développée par Louis (1982), nécessite le calcul de la statistique de score des données complètes, $\boldsymbol{S}_c(\boldsymbol{x}|\boldsymbol{\theta})$, de l'équation (3.32) et de la matrice $\boldsymbol{\mathcal{I}}_c(\boldsymbol{\theta}|\boldsymbol{y}) = E_{\boldsymbol{\theta}}[\boldsymbol{I}_c(\boldsymbol{\theta}|\boldsymbol{X})|\boldsymbol{y}]$ du théorème 3.2. Cette méthode peut être insérée assez facilement dans les itérations de l'algorithme EM. En fait, elle utilise les données complètes pour calculer la matrice d'information observée des paramètres estimés, $\boldsymbol{I}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$. La matrice de variance-covariance est simplement l'inverse de cette matrice. Notons qu'en pratique, on prend un point de convergence de l'algorithme EM comme valeur de $\hat{\boldsymbol{\theta}}$.

D'après Louis (1982), la matrice d'information manquante, $\boldsymbol{\mathcal{I}}_m(\boldsymbol{\theta}|\boldsymbol{y})$, qui a été vue à l'équation (3.39), peut être exprimée sous la forme suivante :

$$\begin{aligned}
\boldsymbol{\mathcal{I}}_m(\boldsymbol{\theta}|\boldsymbol{y}) &= \mathrm{cov}_{\boldsymbol{\theta}}[\boldsymbol{S}_c(\boldsymbol{X}|\boldsymbol{\theta})|\boldsymbol{y}] && (3.41)\\
&= E_{\boldsymbol{\theta}}[\boldsymbol{S}_c(\boldsymbol{X}|\boldsymbol{\theta})\boldsymbol{S}_c^T(\boldsymbol{X}|\boldsymbol{\theta})|\boldsymbol{y}] - E_{\boldsymbol{\theta}}[\boldsymbol{S}_c(\boldsymbol{X}|\boldsymbol{\theta})|\boldsymbol{y}]\, E_{\boldsymbol{\theta}}^T[\boldsymbol{S}_c(\boldsymbol{X}|\boldsymbol{\theta})|\boldsymbol{y}]\\
&= E_{\boldsymbol{\theta}}[\boldsymbol{S}_c(\boldsymbol{X}|\boldsymbol{\theta})\boldsymbol{S}_c^T(\boldsymbol{X}|\boldsymbol{\theta})|\boldsymbol{y}] - \boldsymbol{S}(\boldsymbol{y}|\boldsymbol{\theta})\, \boldsymbol{S}^T(\boldsymbol{y}|\boldsymbol{\theta}), && (3.42)
\end{aligned}$$

où la dernière égalité s'obtient de l'équation (3.33).

En substituant (3.42) dans l'équation (3.39), on trouve que

$$\begin{aligned}
\boldsymbol{I}(\boldsymbol{\theta}|\boldsymbol{y}) &= \boldsymbol{\mathcal{I}}_c(\boldsymbol{\theta}|\boldsymbol{y}) - \boldsymbol{\mathcal{I}}_m(\boldsymbol{\theta}|\boldsymbol{y})\\
&= \boldsymbol{\mathcal{I}}_c(\boldsymbol{\theta}|\boldsymbol{y}) - E_{\boldsymbol{\theta}}[\boldsymbol{S}_c(\boldsymbol{X}|\boldsymbol{\theta})\boldsymbol{S}_c^T(\boldsymbol{X}|\boldsymbol{\theta})|\boldsymbol{y}] + \boldsymbol{S}(\boldsymbol{y}|\boldsymbol{\theta})\, \boldsymbol{S}^T(\boldsymbol{y}|\boldsymbol{\theta}). && (3.43)
\end{aligned}$$

Ainsi, la matrice d'information observée des paramètres estimés, $\boldsymbol{I}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$, peut être calculée grâce à l'équation (3.43) comme suit :

$$\begin{aligned} \boldsymbol{I}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) &= \boldsymbol{\mathcal{I}}_c(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) - \boldsymbol{\mathcal{I}}_m(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) && (3.44) \\ &= \boldsymbol{\mathcal{I}}_c(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) - \left\{ E_{\boldsymbol{\theta}}\Big[\boldsymbol{S}_c(\boldsymbol{X}|\boldsymbol{\theta})\boldsymbol{S}_c^T(\boldsymbol{X}|\boldsymbol{\theta})|\boldsymbol{y}\Big] \right\}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. && (3.45) \end{aligned}$$

Le dernier terme à la droite de l'égalité de l'équation (3.43) a disparu dans l'équation (3.45) puisque $\hat{\boldsymbol{\theta}}$ doit satisfaire $\boldsymbol{S}(\boldsymbol{y}|\boldsymbol{\theta}) = \boldsymbol{0}$.

Lorsqu'on est dans le cas des familles exponentielles régulières, le calcul est beaucoup plus simple. En effet, dans ce cas, la matrice $\boldsymbol{I}_c(\boldsymbol{\theta}|\boldsymbol{x})$ n'est pas une fonction des données. Ainsi, en utilisant l'équation (3.12), McLachlan et Krishnan (1997, p.113) présentent le résultat suivant :

$$\begin{aligned} \boldsymbol{\mathcal{I}}_c(\boldsymbol{\theta}|\boldsymbol{y}) &= \boldsymbol{\mathcal{I}}_c(\boldsymbol{\theta}) \\ &= \text{cov}_{\boldsymbol{\theta}}[\boldsymbol{t}(\boldsymbol{X})]. && (3.46) \end{aligned}$$

On sait aussi de l'équation (3.37) que

$$\boldsymbol{\mathcal{I}}_c(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[\boldsymbol{S}_c(\boldsymbol{X}|\boldsymbol{\theta})\, \boldsymbol{S}_c^T(\boldsymbol{X}|\boldsymbol{\theta})].$$

Ainsi,

$$\text{cov}_{\boldsymbol{\theta}}\Big(\boldsymbol{t}(\boldsymbol{X})\Big) = E_{\boldsymbol{\theta}}[\boldsymbol{S}_c(\boldsymbol{X}|\boldsymbol{\theta})\, \boldsymbol{S}_c^T(\boldsymbol{X}|\boldsymbol{\theta})]. \quad (3.47)$$

De plus, de la définition de la covariance,

$$\begin{aligned} \text{cov}_{\boldsymbol{\theta}}[\boldsymbol{S}_c(\boldsymbol{X}|\boldsymbol{\theta})] &= E_{\boldsymbol{\theta}}[\boldsymbol{S}_c(\boldsymbol{X}|\boldsymbol{\theta})\, \boldsymbol{S}_c^T(\boldsymbol{X}|\boldsymbol{\theta})] - E_{\boldsymbol{\theta}}[\boldsymbol{S}_c(\boldsymbol{X}|\boldsymbol{\theta})]\, E_{\boldsymbol{\theta}}^T[\boldsymbol{S}_c(\boldsymbol{X}|\boldsymbol{\theta})] \\ &= E_{\boldsymbol{\theta}}[\boldsymbol{S}_c(\boldsymbol{X}|\boldsymbol{\theta})\, \boldsymbol{S}_c^T(\boldsymbol{X}|\boldsymbol{\theta})], && (3.48) \end{aligned}$$

puisque $E_{\boldsymbol{\theta}}[\boldsymbol{S}_c(\boldsymbol{X}|\boldsymbol{\theta})] = 0$.

Les équations (3.47) et (3.48) impliquent que $\text{cov}_{\boldsymbol{\theta}}[\boldsymbol{S}_c(\boldsymbol{X}|\boldsymbol{\theta})] = \text{cov}_{\boldsymbol{\theta}}[\boldsymbol{t}(\boldsymbol{X})]$. De (3.41), on a finalement que

$$\begin{aligned} \boldsymbol{\mathcal{I}}_m(\boldsymbol{\theta}|\boldsymbol{y}) &= \text{cov}_{\boldsymbol{\theta}}[\boldsymbol{S}_c(\boldsymbol{X}|\boldsymbol{\theta})|\boldsymbol{y}] \\ &= \text{cov}_{\boldsymbol{\theta}}[\boldsymbol{t}(\boldsymbol{X})|\boldsymbol{y}]. && (3.49) \end{aligned}$$

Ainsi, en remplaçant (3.46) et (3.49) dans l'équation (3.44), McLachlan et Krishnan (1997, p.114) obtiennent le résultat suivant pour la famille exponentielle régulière :

$$\begin{aligned} \boldsymbol{I}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) &= \boldsymbol{\mathcal{I}}_c(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) - \boldsymbol{\mathcal{I}}_m(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) \\ &= \Big[\text{cov}_{\boldsymbol{\theta}}[\boldsymbol{t}(\boldsymbol{X})] - \text{cov}_{\boldsymbol{\theta}}[\boldsymbol{t}(\boldsymbol{X})|\boldsymbol{y}]\Big]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. && (3.50) \end{aligned}$$

À partir de la matrice $\boldsymbol{I}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$ obtenue soit par l'équation (3.45) ou par l'équation (3.50), on peut facilement obtenir la matrice de variance-covariance des paramètres estimés ; il suffit simplement de calculer l'inverse de cette matrice et le tour est joué ! Par contre, la méthode de Louis (1982) nécessite souvent de l'analyse algébrique difficile à résoudre pour le calcul du dernier terme à la droite de l'égalité de l'équation (3.45). On privilégera donc la méthode suivante.

### 3.8.2 Algorithme SEM

La deuxième méthode pour calculer la matrice de variance-covariance des paramètres estimés est celle de Meng et Rubin (1991). Cette technique s'appelle l'algorithme SEM (« Supplemented EM algorithm »), c'est-à-dire qu'on ajoute un supplément aux estimateurs du maximum de vraisemblance trouvés avec l'algorithme EM. Ce supplément est la matrice de variance-covariance asymptotique qui peut être obtenue en utilisant seulement le code pour la matrice de variance-covariance asymptotique pour les données complètes, le code pour l'algorithme EM et du code standard pour effectuer des opérations sur des matrices. Aucune vraisemblance ou log-vraisemblance n'a besoin d'être évaluée dans cette technique. Aussi, lorsqu'il y a une grande augmentation de la variance à cause d'une grande quantité d'information manquante, la séquence d'évaluations requises par l'algorithme SEM semble être très stable.

Dans sa discussion sur l'article de Dempster, Laird et Rubin (1977), Smith (1977) a étudié la possibilité de calculer la variance asymptotique pour l'estimateur du maximum de vraisemblance dans le cas d'un seul paramètre en utilisant le taux de convergence de l'algorithme EM. Smith (1977) a trouvé l'expression suivante :

$$V = V_c/(1-r), \tag{3.51}$$

où $V$ est la variance asymptotique des données observées (c'est ce qu'on cherche), $V_c$ est la variance asymptotique des données complètes et $r$ est le taux de convergence de l'algorithme EM défini à l'équation (3.29).

L'équation (3.51) peut être réécrite sous la forme suivante :

$$V = V_c + \Delta V = V_c\left(1 + \frac{r}{1-r}\right), \tag{3.52}$$

où $\Delta V = [r/(1-r)]V_c$ est l'augmentation dans la variance due aux données manquantes. L'équation (3.52) montre donc ce qu'on pourrait penser de façon intuitive ; la vitesse de convergence de l'algorithme EM est inversement proportionnelle à la quantité d'information manquante. En effet, plus il y a d'information manquante, plus $r$ est grand,

ce qui signifie une convergence lente, et donc plus l'augmentation dans la variance due aux données manquantes est grande.

L'algorithme SEM développé par Meng et Rubin (1991) donne une formulation générale de cette simple procédure pour pouvoir l'appliquer au cas multivarié, c'est-à-dire lorsqu'on a $d > 1$ paramètres.

Notons d'abord que la matrice de variance-covariance asymptotique des paramètres estimés, $\boldsymbol{V}$, est en fait l'inverse de la matrice d'information observée,

$$\boldsymbol{V} = \boldsymbol{I}^{-1}(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}). \tag{3.53}$$

Ainsi, de façon analogue à l'équation (3.52), Meng et Rubin (1991) ont montré qu'en multivarié on a que

$$\boldsymbol{V} \;=\; \boldsymbol{\mathcal{I}}_c^{-1}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) + \boldsymbol{\Delta V}, \tag{3.54}$$

où $\boldsymbol{\Delta V} = \boldsymbol{\mathcal{I}}_c^{-1}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})\boldsymbol{DM}(\hat{\boldsymbol{\theta}})(\boldsymbol{I}_d - \boldsymbol{DM}(\hat{\boldsymbol{\theta}}))^{-1}$ est l'augmentation dans la variance due à l'information manquante et $\boldsymbol{I}_d$ est la matrice identité de dimension $d \times d$.

Pour arriver au résultat de l'équation (3.54), on utilise l'équation (3.39) pour $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ de la façon suivante :

$$\begin{aligned} \boldsymbol{I}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) &= \boldsymbol{\mathcal{I}}_c(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) - \boldsymbol{\mathcal{I}}_m(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) \\ &= \Big[\boldsymbol{I}_d - \boldsymbol{\mathcal{I}}_m(\hat{\boldsymbol{\theta}}|\boldsymbol{y})\boldsymbol{\mathcal{I}}_c^{-1}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})\Big]\boldsymbol{\mathcal{I}}_c(\hat{\boldsymbol{\theta}}|\boldsymbol{y}). \end{aligned} \tag{3.55}$$

Dempster et coll. (1977) ont montré que si $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(b)})$ est maximisé à l'étape M en posant sa première dérivée égale à zéro ($\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(b)})/\partial\boldsymbol{\theta} = \boldsymbol{0}$), alors

$$\boldsymbol{DM}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\mathcal{I}}_m(\hat{\boldsymbol{\theta}}|\boldsymbol{y})\boldsymbol{\mathcal{I}}_c^{-1}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}). \tag{3.56}$$

En remplaçant l'équation (3.56) dans (3.55), on obtient

$$\boldsymbol{I}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) \;=\; \Big[\boldsymbol{I}_d - \boldsymbol{DM}(\hat{\boldsymbol{\theta}})\Big]\boldsymbol{\mathcal{I}}_c(\hat{\boldsymbol{\theta}}|\boldsymbol{y}). \tag{3.57}$$

En inversant l'équation (3.57), on a finalement que

$$\begin{aligned} \boldsymbol{V} &= \boldsymbol{I}^{-1}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) \\ &= \boldsymbol{\mathcal{I}}_c^{-1}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})\Big[\boldsymbol{I}_d - \boldsymbol{DM}(\hat{\boldsymbol{\theta}})\Big]^{-1} \end{aligned}$$

$$
\begin{aligned}
&= \boldsymbol{\mathcal{I}}_c^{-1}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})\Big[(\boldsymbol{I}_d - \boldsymbol{DM}(\hat{\boldsymbol{\theta}})) + \boldsymbol{DM}(\hat{\boldsymbol{\theta}})\Big]\Big[\boldsymbol{I}_d - \boldsymbol{DM}(\hat{\boldsymbol{\theta}})\Big]^{-1} \\
&= \boldsymbol{\mathcal{I}}_c^{-1}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})(\boldsymbol{I}_d - \boldsymbol{DM}(\hat{\boldsymbol{\theta}}))(\boldsymbol{I}_d - \boldsymbol{DM}(\hat{\boldsymbol{\theta}}))^{-1} + \boldsymbol{\mathcal{I}}_c^{-1}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})\boldsymbol{DM}(\hat{\boldsymbol{\theta}})(\boldsymbol{I}_d - \boldsymbol{DM}(\hat{\boldsymbol{\theta}}))^{-1} \\
&= \boldsymbol{\mathcal{I}}_c^{-1}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) + \boldsymbol{\mathcal{I}}_c^{-1}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})\boldsymbol{DM}(\hat{\boldsymbol{\theta}})\Big[\boldsymbol{I}_d - \boldsymbol{DM}(\hat{\boldsymbol{\theta}})\Big]^{-1}, \qquad (3.58)
\end{aligned}
$$

où l'équation (3.58) est la même que l'équation (3.54).

On peut maintenant noter que l'algorithme SEM est composé de 3 parties :

1. L'évaluation de $\boldsymbol{\mathcal{I}}_c^{-1}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$
2. L'évaluation de $\boldsymbol{DM}(\hat{\boldsymbol{\theta}})$
3. L'évaluation de $\boldsymbol{V}$ à partir de (3.58).

On développera chacune de ces parties dans les lignes suivantes.

**1.** *L'évaluation de* $\boldsymbol{\mathcal{I}}_c^{-1}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$.

Dans plusieurs applications de l'algorithme EM, la fonction de densité des données complètes provient de la famille exponentielle. Dans ce cas, il est facile de calculer la matrice d'information observée pour les données complètes, $\boldsymbol{\mathcal{I}}_c(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$. En effet, si on a une fonction de densité de la forme de l'équation (3.6), alors on a de (3.35) que

$$
\begin{aligned}
\boldsymbol{I}_c(\boldsymbol{\theta}|\boldsymbol{x}) &= -\partial^2\, l_c(\boldsymbol{\theta}|\boldsymbol{x}) \,/\, \partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T \\
&= \frac{-\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\Big[\ln\Big(b(\boldsymbol{x})\Big) - \ln\Big(a(\boldsymbol{\theta})\Big) + \boldsymbol{c}^T(\boldsymbol{\theta})\boldsymbol{t}(\boldsymbol{x})\Big] \\
&= \frac{-\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\Big[-\ln\Big(a(\boldsymbol{\theta})\Big) + \boldsymbol{c}^T(\boldsymbol{\theta})\boldsymbol{t}(\boldsymbol{x})\Big] \\
&= \boldsymbol{I}_c\Big(\boldsymbol{\theta}|\boldsymbol{t}(\boldsymbol{x})\Big). \qquad (3.59)
\end{aligned}
$$

Donc $\boldsymbol{I}_c(\boldsymbol{\theta}|\boldsymbol{x})$ est une fonction linéaire de $\boldsymbol{t}(\boldsymbol{x})$. On a vu au théorème 3.2 que $\boldsymbol{\mathcal{I}}_c(\boldsymbol{\theta}|\boldsymbol{y}) = E_{\boldsymbol{\theta}}[\boldsymbol{I}_c(\boldsymbol{\theta}|\boldsymbol{X})|\boldsymbol{y}]$ est l'espérance conditionnelle de la matrice d'information complète sachant $\boldsymbol{y}$. Ainsi, dans le cas des familles exponentielles, on a de Meng et Rubin (1991) et de l'équation (3.59) que

$$
\begin{aligned}
\boldsymbol{\mathcal{I}}_c(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) &= E_{\hat{\boldsymbol{\theta}}}[\boldsymbol{I}_c(\hat{\boldsymbol{\theta}}|\boldsymbol{X})|\boldsymbol{y}] \\
&= E_{\hat{\boldsymbol{\theta}}}[\boldsymbol{I}_c(\hat{\boldsymbol{\theta}}|\boldsymbol{t}(\boldsymbol{X}))|\boldsymbol{y}] \\
&= \boldsymbol{I}_c(\hat{\boldsymbol{\theta}}|E_{\hat{\boldsymbol{\theta}}}[\boldsymbol{t}(\boldsymbol{X})|\boldsymbol{Y}]), \qquad (3.60)
\end{aligned}
$$

où $E_{\hat{\boldsymbol{\theta}}}[\boldsymbol{t}(\boldsymbol{X})|\boldsymbol{Y}]$ est obtenue à l'étape E de la dernière itération de l'algorithme EM pour la famille exponentielle.

Pour la famille exponentielle régulière (lorsque $k = d$ et que le Jacobien de $\boldsymbol{c}(\boldsymbol{\theta})$ est de rang complet), on a que $\boldsymbol{\mathcal{I}}_c(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) = \boldsymbol{\mathcal{I}}_c(\hat{\boldsymbol{\theta}})$.

Meng et Rubin (1991) mentionnent que si on n'est pas dans le cas des familles exponentielles, la fonction de log-vraisemblance des données complètes, $l_c(\boldsymbol{\theta}|\boldsymbol{x})$, n'est plus une fonction linéaire des statistiques exhaustives. La solution pour régler ce problème, lorsque l'échantillon est assez grand, est de faire le développement en série de Taylor pour linéariser $l_c(\boldsymbol{\theta}|\boldsymbol{x})$ en terme des statistiques exhaustives pour de grands échantillons. Une fois cette linéarisation formulée pour l'étape E, on peut calculer $\boldsymbol{\mathcal{I}}_c(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$ en utilisant la même méthode que pour les familles exponentielles puisque $\boldsymbol{I}_c(\boldsymbol{\theta}|\boldsymbol{x})$ est aussi une fonction linéaire de ces statistiques exhaustives pour de grands échantillons.

Après avoir obtenu $\boldsymbol{\mathcal{I}}_c(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$, alors il reste simplement à inverser cette fonction pour finalement avoir la matrice désirée, $\boldsymbol{\mathcal{I}}_c^{-1}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$.

**2.** *L'évaluation de* $\boldsymbol{DM}(\hat{\boldsymbol{\theta}})$.

Soit $r_{ij}$, le $(i,j)$ème élément de la matrice $\boldsymbol{DM}(\hat{\boldsymbol{\theta}})$, et $b$ $(b = 1, 2, ...)$, la $b$ième itération de l'algorithme EM. On définit $\boldsymbol{\theta}^{(b)}(i)$ de la façon suivante :

$$\boldsymbol{\theta}^{(b)}(i) = (\hat{\theta}_1, ..., \hat{\theta}_{i-1}, \theta_i^{(b)}, \hat{\theta}_{i+1}, ..., \hat{\theta}_d). \tag{3.61}$$

On remarque que seulement la $i$ème composante de $\boldsymbol{\theta}^{(b)}(i)$ peut changer. Les autres composantes sont fixées à leur valeur d'EMV. De l'équation (3.28) et d'après Meng et Rubin (1991),

$$\begin{aligned} r_{ij} = (DM(\hat{\boldsymbol{\theta}}))_{ij} &= \left(\frac{\partial M_j(\hat{\boldsymbol{\theta}})}{\partial \theta_i}\right) \\ &= \lim_{\theta_i \to \hat{\theta}_i} \frac{M_j(\hat{\theta}_1, ..., \hat{\theta}_{i-1}, \theta_i^{(b)}, \hat{\theta}_{i+1}, ..., \hat{\theta}_d) - M_j(\hat{\boldsymbol{\theta}})}{\theta_i - \hat{\theta}_i} \\ &= \lim_{b \to \infty} \frac{M_j(\boldsymbol{\theta}^{(b)}(i)) - \hat{\theta}_j}{\theta_i^{(b)} - \hat{\theta}_i} \\ &\equiv \lim_{b \to \infty} r_{ij}^{(b)}. \end{aligned} \tag{3.62}$$

Étant donné que $\boldsymbol{M}(\boldsymbol{\theta})$ est implicitement définie par la sortie des étapes E et M de l'algorithme EM, toutes les valeurs de l'équation (3.62) peuvent être obtenues dans le code de cet algorithme. Les étapes suivantes sont alors suggérées, entre autres par Meng et Rubin (1991), McLachlan et Krishnan (1997, p.130) et Little et Rubin (2002, p.192), pour calculer $r_{ij}^{(b)}$ $(b = 1, 2, ...)$ :

ENTRÉE : $\hat{\boldsymbol{\theta}}$ et $\boldsymbol{\theta}^{(b)}$.

1. Rouler l'algorithme EM pour obtenir $\boldsymbol{\theta}^{(b+1)}$.

   *Répéter les étapes 2 et 3 pour* $i = 1, ..., d$.
2. Calculer $\boldsymbol{\theta}^{(b)}(i)$ à partir de l'équation (3.61) et le prendre comme l'estimation courante de $\boldsymbol{\theta}$. Rouler une itération de l'algorithme EM pour obtenir $\tilde{\boldsymbol{\theta}}^{(b+1)}(i)$.
3. Calculer

$$r_{ij}^{(b)} = \frac{\tilde{\theta}_j^{(b+1)}(i) - \hat{\theta}_j}{\theta_i^{(b)} - \hat{\theta}_i}, \; j = 1, ..., d. \tag{3.63}$$

SORTIE : $\boldsymbol{\theta}^{(b+1)}$ et $\{r_{ij}^{(b)}, i, j = 1, ..., d\}$.

On obtient $r_{ij}$ lorsqu'une suite $r_{ij}^{(b^*)}, r_{ij}^{(b^*+1)}, ...$ est stable pour un $b^*$ donné. Cet algorithme peut être fait pour plusieurs valeurs de $b^*$ pour des éléments différents de $r_{ij}$. Finalement, la matrice $\boldsymbol{DM}(\hat{\boldsymbol{\theta}})$ est obtenue en remplaçant son $(i, j)$ème élément par $r_{ij}$, $i, j = 1, ..., d$.

Notons que cette méthode ne fonctionne pas si un dénominateur de l'équation (3.63) vaut zéro. En effet, lorsque certaines composantes de $\boldsymbol{\theta}$ n'ont pas d'information manquante, l'algorithme EM convergera en une étape pour ces $\theta_i$, $i = 1, ..., d_1$, où $d_1 \leq d$, ce qui entraînera une valeur de 0 au dénominateur de l'équation (3.63) pour les $\theta_i$ en question. Ainsi, la méthode décrite précédemment doit être modifiée comme suit lorsqu'on est dans cette situation.

Supposons que les $d_1$ premières composantes de $\boldsymbol{\theta}$ n'ont pas d'information manquante, Meng et Rubin (1991) ont montré que la matrice $\boldsymbol{DM}(\hat{\boldsymbol{\theta}})$ peut alors s'écrire de la façon suivante :

$$\boldsymbol{DM}(\hat{\boldsymbol{\theta}}) = \begin{pmatrix} \boldsymbol{0}_{d_1 \times d_1} & \boldsymbol{A}_{d_1 \times d_2} \\ \boldsymbol{0}_{d_2 \times d_1} & \boldsymbol{DM}^*(\hat{\boldsymbol{\theta}})_{d_2 \times d_2} \end{pmatrix}, \tag{3.64}$$

où $d_1 + d_2 = d$.

L'algorithme expliqué plus haut pour calculer la matrice $\boldsymbol{DM}(\hat{\boldsymbol{\theta}})$ peut maintenant servir pour calculer la matrice $\boldsymbol{DM}^*(\hat{\boldsymbol{\theta}})$ (pour $i = d_1 + 1, ..., d$) . L'identité suivante, démontrée par Meng et Rubin (1991), montre qu'il suffit de connaître $\boldsymbol{DM}^*(\hat{\boldsymbol{\theta}})$ (et $\boldsymbol{\mathcal{I}}_c^{-1}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$) pour obtenir $\boldsymbol{V}$. Soit

$$\boldsymbol{\mathcal{I}}_c^{-1}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) = \begin{pmatrix} (\boldsymbol{G_1})_{d_1\times d_1} & (\boldsymbol{G_2})_{d_1\times d_2} \\ (\boldsymbol{G_2}^T)_{d_2\times d_1} & (\boldsymbol{G_3})_{d_2\times d_2} \end{pmatrix}. \tag{3.65}$$

Alors on a que

$$\begin{aligned} \boldsymbol{V} &= \boldsymbol{\mathcal{I}}_c^{-1}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) + \begin{pmatrix} \boldsymbol{0}_{d_1\times d_1} & \boldsymbol{0}_{d_1\times d_2} \\ \boldsymbol{0}_{d_2\times d_1} & \boldsymbol{\Delta V}^*_{d_2\times d_2} \end{pmatrix} \\ &= \begin{pmatrix} (\boldsymbol{G_1})_{d_1\times d_1} & (\boldsymbol{G_2})_{d_1\times d_2} \\ (\boldsymbol{G_2}^T)_{d_2\times d_1} & (\boldsymbol{G_3} + \boldsymbol{\Delta V}^*)_{d_2\times d_2} \end{pmatrix}, \end{aligned} \tag{3.66}$$

où $\boldsymbol{\Delta V}^* = (\boldsymbol{G_3} - \boldsymbol{G_2^T}\boldsymbol{G_1}^{-1}\boldsymbol{G_2})\boldsymbol{DM}^*(\hat{\boldsymbol{\theta}})(\boldsymbol{I}_{d_2} - \boldsymbol{DM}^*(\hat{\boldsymbol{\theta}}))^{-1}$ et $\boldsymbol{I}_{d_2}$ est la matrice identité de dimension $d_2 \times d_2$. L'équation (3.66) est un cas spécial de l'équation (3.54).

**3.** *L'évaluation de* $\boldsymbol{V}$.

Il suffit de remplacer $\boldsymbol{\mathcal{I}}_c^{-1}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$ et $\boldsymbol{DM}(\hat{\boldsymbol{\theta}})$ obtenus des étapes 1 et 2 de l'algorithme SEM dans l'équation (3.58) pour obtenir la matrice de variance-covariance des paramètres estimés, $\boldsymbol{V}$. Ceci termine l'explication des 3 parties de l'algorithme SEM.

L'algorithme SEM permet de calculer $\boldsymbol{V}$ de façon stable numériquement pour les raisons suivantes :

1. Lorsqu'il y a peu d'information manquante, alors $\boldsymbol{\Delta V}$ est petit par rapport à $\boldsymbol{\mathcal{I}}_c^{-1}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$. Dans ce cas, même si le calcul de $\boldsymbol{DM}(\hat{\boldsymbol{\theta}})$ est sujet à des imprécisions numériques à cause de la convergence rapide de l'algorithme EM, cela a peu d'effet sur le calcul de $\boldsymbol{V} = \boldsymbol{\mathcal{I}}_c^{-1}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) + \boldsymbol{\Delta V}$.
2. Lorsqu'il y a beaucoup d'information manquante, alors $\boldsymbol{\Delta V}$ a un poids élevé dans le calcul de $\boldsymbol{V}$. Dans ce cas, la convergence assez lente de l'algorithme EM assure une précision numérique de $\boldsymbol{DM}(\hat{\boldsymbol{\theta}})$, et donc le calcul de $\boldsymbol{V}$ se fait encore une fois de façon assez exacte.

De plus, Little et Rubin (2002 , p.195) expliquent que l'algorithme SEM produit des diagnostics pour les erreurs de programmation et pour les erreurs numériques. Par exemple, la matrice $\boldsymbol{\Delta V}$ est symétrique analytiquement, mais il se peut qu'elle ne le

soit pas numériquement à cause d'erreurs de programmation ou à cause d'imprécisions numériques dans le calcul de $\hat{\boldsymbol{\theta}}$ ou de $\boldsymbol{DM}(\hat{\boldsymbol{\theta}})$. Lorsqu'on est en présence d'asymétrie, il est suggéré par Meng et Rubin (1991) d'essayer d'augmenter la précision des étapes E et M en utilisant un critère d'arrêt plus sévère. Si l'asymétrie persiste, c'est signe qu'il y a des erreurs de programmation soit dans le code de l'algorithme EM ou de l'algorithme SEM.

Aussi, même si la matrice $\boldsymbol{V}$ est symétrique, il se peut qu'elle ne soit pas définie non négative, ce qui suggère cette fois soit une erreur de programmation, une erreur numérique ou une convergence de l'algorithme EM vers un point de selle.

# Chapitre 4

# Le modèle des risques concurrents

En durées de vie, on s'intéresse habituellement au temps de décès d'un individu ou au temps de panne d'un système. Dans certaines expériences, il peut y avoir plusieurs causes possibles de décès ou de panne ; c'est ce qu'on appelle les risques concurrents. Dans le reste de ce chapitre, on parlera de temps et de cause de panne dans le but d'alléger le texte.

Soit $J$ ($J \geq 2$) causes de panne possibles. Si $X_j,\ j = 1, ..., J$, est la variable aléatoire du temps de panne dû à la $j$ème cause, alors pour chaque système on observe

$$T = \min(X_1, ..., X_J),$$

c'est-à-dire le temps auquel le système tombe en panne d'une des $J$ causes. De plus, on dénote par $\delta$ la cause de panne, c'est-à-dire que $\delta = j$, si $T = X_j$. Ceci implique que $X_1, ..., X_J$ sont des variables aléatoires latentes, c'est-à-dire qu'elles ne sont pas observables. En effet, ce sont seulement les variables $T$ et $\delta$ qui sont observées.

Klein et Moeschberger (2003, p.50) présentent le paramètre de base des risques concurrents qui est la fonction de risque spécifique à la cause de décès $j$, définie par

$$\begin{aligned} \lambda_j(t) &= \lim_{h \downarrow 0} \frac{P[t \leq T \leq t+h, \delta = j | T \geq t]}{h},\ j = 1, ..., J \\ &= \lim_{h \downarrow 0} \frac{P[t \leq X_j \leq t+h, \delta = j | X_j \geq t, j = 1, ..., J]}{h},\ j = 1, ..., J. \end{aligned} \quad (4.1)$$

Cette fonction donne donc le taux auquel les systèmes à risque tombent en panne

de la cause $j$. Autrement dit, $\lambda_j(t)$ donne la « probabilité » de tomber en panne de la cause $j$ dans l'instant suivant $t$ sachant que le système est toujours fonctionnel à $t$.

La fonction de risque total, c'est-à-dire la fonction de risque pour toutes les causes de panne confondues, est simplement donnée par

$$\lambda(t) \quad = \quad \sum_{j=1}^{J} \lambda_j(t). \tag{4.2}$$

$\lambda(t)$ représente le taux auquel les systèmes qui ne sont pas encore tombés en panne d'aucune des $J$ causes au temps $t$ tombent en panne de n'importe quelle cause.

Ainsi, la fonction de survie totale, c'est-à-dire la probabilité de ne pas être encore tombé en panne d'aucune des $J$ causes au temps $t$, peut s'écrire de la façon suivante :

$$S(t) = P(T > t) = \exp\left\{ -\int_0^t \lambda(u)du \right\}. \tag{4.3}$$

La fonction de densité totale est obtenue de l'équation (4.3) et est définie par la « probabilité » de tomber en panne au temps $t$ d'une des $J$ causes, c'est-à-dire que

$$f(t) = -\frac{\partial}{\partial t} S(t) = \lambda(t) \exp\left\{ -\int_0^t \lambda(u)du \right\}. \tag{4.4}$$

Notons que $\lambda$, $S$ et $f$ sont les fonctions de risque, de survie et de densité de la variable aléatoire $T$, respectivement.

La fonction de risque spécifique, $\lambda_j(t)$, peut être calculée à partir de la fonction de survie jointe des $J$ risques concurrents. En effet, soit $S(t_1, ..., t_J) = P(X_1 > t_1, ..., X_J > t_J)$, la fonction de survie jointe. Alors la fonction de risque spécifique est donnée par

$$\lambda_j(t) \quad = \quad \frac{-\partial S(t_1, ..., t_J)/\partial t_j \mid_{t_1=...=t_J=t}}{S(t, ..., t)}, \; j = 1, ..., J. \tag{4.5}$$

On peut aussi définir la fonction de densité spécifique à la cause de panne $j$ comme suit :

$$f_j(t) = \left. \frac{-\partial S(t_1, ..., t_J)}{\partial t_j} \right|_{t_1=...=t_J=t} = \lambda_j(t) \exp\left\{ -\int_0^t \lambda(u)du \right\}, \; j = 1, ..., J. \tag{4.6}$$

L'équation (4.6) représente la « probabilité » de tomber en panne au temps $t$ de la cause $j$.

Considérons maintenant la fonction suivante :

$$S_j(t) = \exp\left\{ -\int_0^t \lambda_j(u)du \right\},\ j = 1, ..., J. \tag{4.7}$$

On a donc que

$$\begin{aligned} S(t) &= \exp\left\{ -\int_0^t \lambda(u)du \right\} && \text{(de (4.3))} \\ &= \exp\left\{ -\int_0^t \Big[\lambda_1(u) + \cdots + \lambda_J(u)\Big]du \right\} && \text{(de (4.2))} \\ &= \exp\left\{ -\int_0^t \lambda_1(u)du \right\} \cdots \exp\left\{ -\int_0^t \lambda_J(u)du \right\} \\ &= S_1(t) \cdots S_J(t) && \text{(de (4.7))} \\ &= \prod_{j=1}^{J} S_j(t). && (4.8) \end{aligned}$$

Si les risques concurrents $X_1, ..., X_J$ ne sont pas indépendants, alors $S_j(t)$, $j = 1, ..., J$ ne peut pas s'interpréter comme une probabilité ou une fonction de survie. Par contre, si les risques concurrents $X_1, ..., X_J$ sont indépendants, alors $S_j(t)$ est la fonction de survie marginale de $X_j$, c'est-à-dire qu'elle peut être vue comme la probabilité de survivre au moins jusqu'au temps $t$ dans un environnement où la seule cause de panne possible est $j$.

On peut aussi calculer la fonction d'incidence cumulée due à la cause $j$. Celle-ci représente la proportion de systèmes fonctionnels au temps 0 qui tomberont en panne de la cause $j$ avant le temps $t$. Cette fonction est présentée par Klein et Moeschberger (2003, p.52) comme suit :

$$F_j(t) = P\Big(T \leq t, \delta = j\Big) = \int_0^t \lambda_j(u)S(u)du. \tag{4.9}$$

$F_j(t)$ n'est pas une vraie fonction de répartition. En effet, lorsque $t \to \infty$, cette fonction tend vers la probabilité de tomber en panne de la cause $j$ ($F_j(\infty) = P(\delta = j)$). Par contre, cette fonction est non-décroissante avec $F_j(0) = 0$ et $F_j(\infty) \leq 1$.

Admettons maintenant qu'on a $n$ systèmes pour chacun desquels on observe les données suivantes : $[t_i,\ \delta_i,\ j_i],\ i = 1, ..., n$, où

$t_i$ : Le temps de panne ou de censure observé pour le système $i$
$\delta_i$ : L'indicatrice qui vaut 1 si le système $i$ a une panne, 0 sinon.
$j_i$ : Cause de panne du système $i$.

Alors d'après Kalbfleisch et Prentice (2002, p.254), la fonction de vraisemblance pour le modèle des risques concurrents peut s'écrire de la façon suivante :

$$
\begin{aligned}
L &= \prod_{i=1}^{n} \Big( \{f_{j_i}(t_i)\}^{\delta_i} \{S(t_i)\}^{1-\delta_i} \Big) && (4.10)\\
&= \prod_{i=1}^{n} \Big( \{\lambda_{j_i}(t_i) S(t_i)\}^{\delta_i} \{S(t_i)\}^{1-\delta_i} \Big) && \text{(de (4.6))}\\
&= \prod_{i=1}^{n} \Big( \{\lambda_{j_i}(t_i)\}^{\delta_i} S(t_i) \Big) \\
&= \prod_{i=1}^{n} \left( \{\lambda_{j_i}(t_i)\}^{\delta_i} \prod_{j=1}^{J} S_j(t_i) \right\} \Bigg) && \text{(de (4.8))}\\
&= \prod_{i=1}^{n} \left( \{\lambda_{j_i}(t_i)\}^{\delta_i} \prod_{j=1}^{J} \exp \Big\{ - \int_0^{t_i} \lambda_j(u) du \Big\} \right) && \text{(de (4.7)).} \quad (4.11)
\end{aligned}
$$

L'équation (4.10) représente le produit pour $i = 1, ..., n$ de la contribution du système $i$. Chaque contribution se traduit par un système qui tombe en panne de la cause $j_i$ au temps $t_i$ ou encore par la censure du système au temps $t_i$. Pour observer un temps de censure pour le système $i$ au temps $t_i$, il faudrait par exemple que ce système ne soit toujours pas tombé en panne à la fin de l'étude. Dans ce cas, on dit que le système $i$ est censuré à droite à $t_i = \tau$, c'est-à-dire que son temps de panne est plus grand que le temps de la fin de l'étude, $\tau$.

# Chapitre 5

# Travaux réalisés sur les causes de panne masquées

Au chapitre 4, on a vu que dans le contexte des risques concurrents, les systèmes tombent en panne d'une cause parmi $J$ causes possibles. Il arrive souvent, en pratique, que la cause de panne de certains systèmes soit masquée, c'est-à-dire que la cause ne peut pas être identifiée, mais qu'elle se situe dans un sous-groupe des $J$ causes. Par exemple, supposons qu'on a 3 causes de panne possibles et qu'un système tombe en panne d'une cause qui est masquée dans le groupe $\{1, 2\}$. Alors on sait que la cause de panne du système en question est soit 1 ou 2, mais on ne connaît pas la cause exacte. Cependant, certains des systèmes masqués peuvent être analysés plus profondément à une deuxième étape où leur unique cause de panne est alors dévoilée.

La méthode du maximum de vraisemblance décrite au chapitre 2 est une technique qui permet de faire l'estimation de certains paramètres d'une distribution donnée. Par contre, lorsqu'on étudie des données de survie avec des risques concurrents et que certaines causes de panne sont masquées, l'estimation des paramètres par cette méthode peut devenir un peu plus compliquée.

Dans ce chapitre, on donne un bref aperçu de certains travaux qui ont été publiés sur l'utilisation de la méthode du maximum de vraisemblance pour estimer des paramètres lorsqu'il y a des risques concurrents et que certaines causes de panne (ou de décès) sont masquées. Plusieurs modèles de survie pour trouver les estimateurs dans ce contexte ont été développés depuis quelques années. On présente ici quelques-uns de ces modèles.

Le premier à être abordé est le modèle nonparamétrique élaboré par Dinse en 1986. Ensuite, un modèle avec des risques concurrents proportionnels est développé par Flehinger, Reiser et Yashchin (1998). En 2002, ces mêmes auteurs font une modélisation paramétrique des données de survie. Enfin, en 2004, un modèle basé sur des risques concurrents constants par paliers est construit par Craiu et Duchesne. Les trois derniers modèles utilisent le concept des données de deuxième étape. À cette étape, les causes de panne de certains des systèmes dont la cause était masquée au départ sont identifiées. Aussi, les modèles proposés par Dinse (1986) et Reiser et al. (1998 et 2002) supposent l'indépendance des risques concurrents, tandis que le modèle de Craiu et Duchesne (2004) suppose plutôt la dépendance. L'hypothèse des risques indépendants est souvent critiquée comme en fait foi l'article de Prentice, Kalbfleisch, Peterson, Flournoy, Farewell et Breslow (1978).

## 5.1 Modèle de survie nonparamétrique

Dans son article de 1986, Dinse a comme objectif principal d'estimer la prévalence d'une maladie donnée et la mortalité due à cette maladie. Ces estimateurs avaient été trouvés de façon nonparamétrique grâce à la méthode du maximum de vraisemblance dans des études antérieures lorsque la cause exacte de décès était toujours connue. Dans l'ouvrage de Dinse (1986), ces méthodes sont généralisées pour pouvoir estimer les paramètres même lorsque les causes de décès sont inconnues pour certains sujets. On utilise donc un modèle nonparamétrique de survie avec deux risques concurrents. Le premier risque est la maladie d'intérêt et le deuxième est n'importe quelle autre cause qui a pu engendrer le décès.

Tout d'abord, définissons quelques termes importants. La *prévalence* d'une maladie est un indicateur du développement de la maladie à un temps $t$ précis. C'est la proportion de sujets en vie au temps $t$ étant porteurs de la maladie en question. L'indice de *mortalité* d'une maladie est, pour sa part, un indicateur des effets d'une maladie sur la longévité. On peut l'estimer par le taux de risque spécifique de décès causé par la maladie au temps $t$.

Dinse (1986) utilise les données d'une expérience sur 58 souris femelles de laboratoire qui sont suivies jusqu'à ce qu'elles décèdent. Suite à leur décès, les souris sont examinées pour savoir si elles sont porteuses du NRVD (une maladie vasculaire extrarénale). Le tableau 5.1 présente ces données.

TAB. 5.1 – Temps de décès (en jours) et état de la maladie (NRVD) au décès pour 58 souris femelles. Les variables $Y$, $D$ et $U$ sont définies dans le texte.

| État de la maladie (NRVD) au décès | Y | D | U | Temps (T) de décès en jours |
|---|---|---|---|---|
| Absent | 0 | 0 | 0 | 231, 444, 468, 473, 527, 550, 593, 600, 610, 650, 655, 660, 715, 720, 752, 785, 832, 838, 859, 891, 896, 904, 931, 952, 998 |
| Secondaire : maladie présente, mais pas la cause de décès | 1 | 0 | 0 | 559, 595, 596, 603, 765, 783, 794, 811, 856, 870, 883, 897, 975, 978, 991, 1005, 1023, 1026, 1053 |
| Inconnu : maladie présente, mais on ne sait pas si c'est la cause du décès | 1 | · | 1 | 593, 735, 816, 848, 850, 1046 |
| Mortel : maladie présente et cause de décès | 1 | 1 | 0 | 500, 591, 713, 751, 778, 784, 786, 796 |

Soit $T$, le temps de décès. Voici la définition des variables nécessaires à la compréhension du tableau 5.1 :

$Y(t)$ : Variable indicatrice qui vaut 1 si la maladie était présente au temps $t$.
$Y$ : Variable indicatrice qui vaut 1 si la maladie était présente lors du décès (=Y(T)).
$D$ : Variable indicatrice qui vaut 1 si la maladie est la cause du décès.
$U$ : Variable indicatrice qui vaut 1 si on ne sait pas si le décès est dû à la maladie.

Soient $t_1 < t_2 < ... < t_L$, les $L$ temps de décès distincts observés. Alors on définit $a_l, b_l, c_l$ et $d_l$ comme étant le nombre de souris décédées au temps $t_l$ avec un état de maladie « absent », « secondaire », « inconnu » et « mortel », respectivement ($l = 1, ..., L$).

Comme il a été mentionné plus haut, le but de cet article est d'obtenir des estimateurs du maximum de vraisemblance des fonctions de prévalence et de mortalité de façon nonparamétrique. Voici quelques définitions utiles dans l'écriture de la fonction de vraisemblance.

La fonction de prévalence de la maladie est donnée par

$$h(t) = P[Y(t) = 1|T > t]. \tag{5.1}$$

Cette fonction représente la proportion d'animaux ayant la maladie parmi tous les

animaux vivants au temps $t$. Par contre, lorsque les animaux sont vivants, on ne sait pas s'ils sont atteints de la maladie en question. En effet, la présence ou non de la maladie est détectée lors de l'autopsie effectuée au décès de l'animal. On doit donc utiliser une autre probabilité, équivalente à la première, pour représenter la prévalence. Cette fonction est la suivante :

$$p(t) = P[Y = 1|D = 0, T = t]. \tag{5.2}$$

Cette dernière est la proportion d'animaux ayant la maladie parmi tous ceux qui meurent d'une autre cause au temps $t$. On doit prendre seulement les animaux qui meurent d'une autre cause ($D = 0$) au temps $t$ pour que l'équation (5.2) soit équivalente à (5.1). En effet, dans ces deux équations, ce sont seulement les animaux malades et vivants au temps $t$ qui sont considérés.

Il faut noter que la solution nonparamétrique du maximum de vraisemblance distribue la masse seulement aux temps de décès observés. Ainsi, la variable aléatoire $T$ sera traitée de façon discrète. La fonction de mortalité due à la maladie est alors la fonction de risque spécifique discrète suivante :

$$\lambda_l^{(1)} = P[T = t_l, D = 1|T \geq t_l], \tag{5.3}$$

où l'exposant (1) indique que le décès est dû à la cause 1 (maladie). C'est la fonction de risque de décès de la maladie au temps $t_l$ ($l = 1, ..., L$).

On définit aussi la fonction

$$q(t) = P[D = 1|Y = 1, T = t] \tag{5.4}$$

comme étant la proportion d'animaux qui décèdent de la maladie au temps $t$ parmi tous les animaux qui décèdent au temps $t$ avec la maladie présente.

La fonction

$$g(t) = P[Y = 1|T = t] \tag{5.5}$$

désigne la proportion d'animaux ayant la maladie parmi tous ceux décédant au temps $t$ et

$$\lambda_l = P[T = t_l|T \geq t_l] \tag{5.6}$$

définit la fonction de risque discrète de décès de n'importe quelle cause au temps $t_l$ ($l = 1, ..., L$). Remarquons que les fonctions $p(t)$ et $\lambda_l^{(1)}$ définies aux équations (5.2) et (5.3) peuvent s'écrire comme des fonctions de $g(t)$, $q(t)$ et $\lambda_l$ de la façon suivante :

$$p(t) = g(t)\,[1 - q(t)]\,/\,[1 - g(t)\,q(t)], \tag{5.7}$$

$$\lambda_l^{(1)} = \lambda_l\,g(t_l)\,q(t_l). \tag{5.8}$$

TAB. 5.2 – Contribution de chaque animal à la vraisemblance en conditionnant sur le décès au temps $t_l$ pour quatre observations possibles.

| État de la maladie (NRVD) au décès | Observation | Contribution à la vraisemblance sachant que l'animal décède au temps $t_l$ | Nombre d'animaux |
|---|---|---|---|
| Absent | $\{T = t_l, Y = 0, D = 0, U = 0\}$ | $1 - g(t_l)$ | $a_l$ |
| Secondaire | $\{T = t_l, Y = 1, D = 0, U = 0\}$ | $g(t_l)\left(1 - q(t_l)\right)\left(1 - \xi(t_l)\right)$ | $b_l$ |
| Inconnu | $\{T = t_l, Y = 1, D = ., U = 1\}$ | $g(t_l)\left\{\left(1-q(t_l)\right)\xi(t_l)+q(t_l)\,\psi(t_l)\right\}$ | $c_l$ |
| Mortel | $\{T = t_l, Y = 1, D = 1, U = 0\}$ | $g(t_l)\,q(t_l)\left(1 - \psi(t_l)\right)$ | $d_l$ |

Si certains animaux ont une cause de décès inconnue, la fonction de vraisemblance dépend alors de la distribution de la variable $U$ sachant $(D, Y, T)$. Les fonctions suivantes définissent les deux possibilités lorsque la cause de décès est inconnue :

$$\xi(t) = P[U = 1|D = 0, Y = 1, T = t], \tag{5.9}$$

$$\psi(t) = P[U = 1|D = 1, Y = 1, T = t]. \tag{5.10}$$

Ces fonctions représentent la proportion d'animaux ayant une cause de décès inconnue parmi les animaux qui décèdent au temps $t$ d'une autre cause que la maladie lorsque la maladie est présente (état « secondaire ») et de la maladie (état « mortel »), respectivement. Ultérieurement, $\xi(t)$ et $\psi(t)$ seront vues comme des taux d'incertitude et seront traitées comme des fonctions nuisibles.

Grâce aux définitions décrites ci-haut, il est maintenant possible de faire l'estimation nonparamétrique du maximum de vraisemblance. Tout d'abord, on s'intéresse au calcul de la fonction de vraisemblance pour les données observées. Sachant qu'un animal est décédé au temps $t_l$, sa contribution conditionnelle à la vraisemblance est donnée au tableau 5.2. Ces derniers peuvent décéder avec un état de maladie « absent », « secondaire », « inconnu » ou « mortel ». Chaque contribution inconditionnelle de la vraisemblance est un produit du terme approprié du tableau 5.2 et de la probabilité discrète de décéder au temps $t_l$. Cette dernière est donnée par $\lambda_l(1 - \lambda_{l-1})...(1 - \lambda_1)$. La fonction de vraisemblance peut donc s'écrire comme suit :

$$
\begin{aligned}
L(\lambda, g, q, \xi, \psi) = \prod_{l=1}^{L} \Bigg[ & \Big(\big(1 - g(t_l)\big)\, \lambda_l\, (1-\lambda_{l-1})...(1-\lambda_1)\Big)^{a_l} \\
& \times \Big(g(t_l)\, \big(1 - q(t_l)\big)\, \big(1 - \xi(t_l)\big)\, \lambda_l\, (1-\lambda_{l-1})...(1-\lambda_1)\Big)^{b_l} \\
& \times \Big(g(t_l)\Big\{\big(1 - q(t_l)\big)\xi(t_l) + q(t_l)\psi(t_l)\Big\}\lambda_l(1-\lambda_{l-1})...(1-\lambda_1)\Big)^{c_l} \\
& \times \Big(g(t_l)\, q(t_l)\, \big(1 - \psi(t_l)\big)\, \lambda_l\, (1-\lambda_{l-1})...(1-\lambda_1)\Big)^{d_l} \Bigg].
\end{aligned}
$$

La log-vraisemblance s'écrit alors

$$
\begin{aligned}
l(\lambda, g, q, \xi, \psi) = \sum_{l=1}^{L} \Bigg[ & a_l\, \ln\Big(1 - g(t_l)\Big) + b_l\, \ln\Big(g(t_l)\, \big(1 - q(t_l)\big)\, \big(1 - \xi(t_l)\big)\Big) \\
& + c_l\, \ln\Big(g(t_l)\, \Big\{\big(1 - q(t_l)\big)\, \xi(t_l)\, +\, q(t_l)\, \psi(t_l)\Big\}\Big) \\
& + d_l\, \ln\Big(g(t_l)\, q(t_l)\, \big(1 - \psi(t_l)\big)\Big) \\
& + (a_l + b_l + c_l + d_l)\, \ln\Big(\, \lambda_l\, (1-\lambda_{l-1})...(1-\lambda_1)\, \Big) \Bigg].
\end{aligned}
$$

On peut regrouper ensemble tous les termes qui sont fonction de $\lambda$, ensuite tous les termes qui sont fonction de $g$ et finalement tous les termes qui sont fonction de $q, \xi$ et $\psi$. On additionne donc trois log-vraisemblances : $l(\lambda, g, q, \xi, \psi) = l_1(\lambda) + l_2(g) + l_3(q, \xi, \psi)$, où

$$
\begin{aligned}
l_1(\lambda) &= \sum_{l=1}^{L} (a_l + b_l + c_l + d_l)\, \ln\Big(\lambda_l\, (1-\lambda_{l-1})...(1-\lambda_1)\Big) \\
&= \sum_{l=1}^{L} \Big[(a_l + b_l + c_l + d_l)\, \ln(\lambda_l) + (n_l - a_l - b_l - c_l - d_l)\, \ln(1-\lambda_l)\Big], \\
l_2(g) &= \sum_{l=1}^{L} \Big[(b_l + c_l + d_l)\, \ln\Big(g(t_l)\Big) + a_l\, \ln\Big(1 - g(t_l)\Big)\Big], \\
l_3(q, \xi, \psi) &= \sum_{l=1}^{L} \Big[d_l\, \ln\Big(q(t_l)\Big) + b_l\, \ln\Big(1 - q(t_l)\Big) + b_l\, \ln\Big(1 - \xi(t_l)\Big) \\
&\qquad + d_l\, \ln\Big(1 - \psi(t_l)\Big)\, +\, c_l\, \ln\Big(\big(1 - q(t_l)\big)\xi(t_l) + q(t_l)\psi(t_l)\Big)\Big],
\end{aligned}
$$

où $n_l = \sum_{m=l}^{L}(a_m + b_m + c_m + d_m)$ est le nombre d'animaux à risque de décéder de n'importe quelle cause au temps $t_l$.

Puisqu'il n'y a que $l_1(\lambda)$ qui dépende de $\lambda$, on peut donc trouver l'estimateur du maximum de vraisemblance de $\lambda_l$ en prenant la dérivée de $l_1(\lambda)$ par rapport à $\lambda_l$ et en l'égalant à 0. Il suffit alors d'isoler $\lambda_l$ et on trouve facilement que

$$\hat{\lambda}_l = (a_l + b_l + c_l + d_l)/n_l, \quad l = 1, ..., L,$$

soit le nombre de souris décédées à $t_l$ sur le nombre de souris à risque de décéder de n'importe quelle cause à ce temps.

Les EMV de $p(t), q(t)$ et $g(t)$ ne sont pas uniquement définis aux temps où il n'y a pas de décès. En fait, à moins qu'il y ait beaucoup de décès aux temps $t_l$ ($l = 1, ..., L$), les estimateurs sont instables même à ces temps. Une façon de stabiliser les estimateurs de $p(t)$, $q(t)$ et $g(t)$ sans faire d'hypothèse sur les distributions serait de supposer que celles-ci sont constantes sur des intervalles de temps. Pour le reste de cet article, on suppose que $p(t), q(t)$ et $g(t)$ sont constantes sur les intervalles appartenant à l'ensemble $\{I_k : k = 1, ..., K\}$, où les $I_k$ sont $K$ intervalles mutuellement exclusifs couvrant l'étendue totale des temps de décès observés. De plus, certaines contraintes sur les taux d'incertitude, $\xi(t)$ et $\psi(t)$, doivent être formulées afin de pouvoir identifier tous les estimateurs. En effet, sans ces contraintes, le problème est surparamétrisé et $q(t)$ n'est pas identifiable.

Comme première contrainte, on peut supposer que les taux d'incertitude sont égaux ($\xi(t) = \psi(t)$). En d'autres mots, on émet l'hypothèse que les décès causés par la maladie ont autant de chances de ne pas être identifiés que ceux qui ne sont pas causés par la maladie, même si celle-ci était présente au décès, et ce, peu importe le temps de décès. En ajoutant la contrainte $\xi(t) = \psi(t)$, la fonction de log-vraisemblance s'exprime plus simplement. En effet, $l_3(q, \xi, \psi)$ peut maintenant se séparer en une somme de deux log-vraisemblances ($l_4(q)$ et $l_5(\psi)$), où

$$\begin{aligned} l_4(q) &= \sum_{l=1}^{L} \left[ d_l \ln\Big(q(t_l)\Big) + b_l \ln\Big(1 - q(t_l)\Big) \right], \\ l_5(\psi) &= \sum_{l=1}^{L} \left[ c_l \ln\Big(\psi(t_l)\Big) + (b_l + d_l) \ln\Big(1 - \psi(t_l)\Big) \right]. \end{aligned}$$

Pour tout $t \in I_k$, les EMV restreints pour $g(t)$ et $q(t)$ s'obtiennent facilement à partir des équations de log-vraisemblance $l_2(g)$ et $l_4(q)$, respectivement. On a donc

$$\begin{aligned} \hat{g}(t) &= \frac{\sum_{\{m:t_m \in I_k\}} \Big(b_m + c_m + d_m\Big)}{\sum_{\{m:t_m \in I_k\}} \Big(a_m + b_m + c_m + d_m\Big)}, \\ \hat{q}(t) &= \frac{\sum_{\{m:t_m \in I_k\}} d_m}{\sum_{\{m:t_m \in I_k\}} \Big(b_m + d_m\Big)}, \end{aligned}$$

où $\hat{g}(t)$ peut s'interpréter comme étant le nombre de souris ayant la maladie décédées dans l'intervalle $k$, sur le nombre de souris décédées dans l'intervalle $k$, pour tout $t \in I_k$. Quant à $\hat{q}(t)$, c'est le nombre de souris décédées dans l'intervalle $k$ dont on sait que la cause de décès est la maladie, sur le nombre de souris décédées dans l'intervalle $k$ porteuses de la maladie et dont on connaît la cause de décès, pour tout $t \in I_k$. Pour trouver les estimateurs $\hat{p}(t)$ et $\hat{\lambda}_l^{(1)}$, il suffit de substituer $\hat{\lambda}_l$, $\hat{g}(t)$ et $\hat{q}(t)$ dans les équations (5.7) et (5.8).

L'hypothèse que $\xi(t) = \psi(t)$ simplifie énormément l'analyse, mais elle n'est pas toujours réaliste. Alors comme deuxième contrainte possible, on peut plutôt supposer que les taux d'incertitude ne dépendent pas du temps ($\xi(t) = \xi$ et $\psi(t) = \psi$). Cette hypothèse implique que la cause de décès d'un jeune animal a la même chance d'être inconnue que celle d'un vieil animal. Les EMV de $\lambda_l$ et de $g(t)$ sont les mêmes qu'avant, mais les EMV de $q_k, \xi$ et $\psi$ sont maintenant trouvés en maximisant $l_3(q, \xi, \psi)$, où $q_k$ est la valeur de $q(t)$ pour tout $t \in I_k$. Ici, l'algorithme EM expliqué au chapitre 3 est utilisé afin de trouver les EMV de $q_k, \xi$ et $\psi$. On procède de la façon suivante :

Soit $e_l$ et $f_l$ le nombre de souris décédées avec un état de maladie « secondaire » et « mortel » , respectivement, parmi les $c_l$ animaux décédant au temps $t_l$ d'un état de maladie « inconnu » ($l = 1, ..., L$). On définit aussi

$$\begin{aligned} w(t) &= P(D = 1 | U = 1, Y = 1, T = t) \\ &= \frac{\psi(t)\, q(t)}{\psi(t)\, q(t) + \xi(t)\big(1 - q(t)\big)} \end{aligned} \tag{5.11}$$

comme étant la proportion de souris qui décèdent de la maladie parmi tous les décès au temps $t$ avec la maladie présente, mais avec la cause de décès inconnue. Soit $q^{(0)}(t_l)$, $\xi^{(0)}(t_l)$ et $\psi^{(0)}(t_l)$, des valeurs initiales arbitraires. À l'itération $(b+1)$, l'étape E de l'algorithme EM pose d'abord $e_l^{(b)}$ et $f_l^{(b)}$ égales à leurs valeurs espérées, en conditionnant sur les données observées et les EMV courants des fonctions qui entrent dans l'algorithme. Ainsi, on pose $f_l^{(b)} = c_l w^{(b)}(t)$ et $e_l^{(b)} = c_l - f_l^{(b)}$, où $w^{(b)}(t)$ est obtenu en substituant les valeurs de $q^{(b)}(t_l)$, $\xi^{(b)}(t_l)$ et $\psi^{(b)}(t_l)$ dans l'équation (5.11). L'étape M estime alors $q^{(b+1)}(t_l)$, $\xi^{(b+1)}(t_l)$ et $\psi^{(b+1)}(t_l)$ par les valeurs qui maximisent la vraisemblance $l_3(q, \xi, \psi)$, où $e_l^{(b)}$ et $f_l^{(b)}$ contribuent de la même façon que $b_l$ et $d_l$, respectivement. On alterne alors les étapes E et M à chaque itération de l'algorithme EM jusqu'à ce que la différence entre les valeurs des EMV de l'itération en cours et de celle qui précède soit négligeable, comme on l'a vu à la section 3.1.

De plus, pour tout $t \in I_k$, l'EMV pour $p(t)$ est obtenu en substituant $\hat{g}(t)$ et $\hat{q}_k$ dans l'équation (5.7). Finalement, l'EMV de $\lambda_l^{(1)}$ est encore donné par l'équation (5.8) en remplaçant les paramètres en place par leur EMV.

D'autres hypothèses sur les taux d'incidence sont possibles, mais on n'en élaborera pas les détails dans ce mémoire. Finalement, les estimateurs pour l'exemple des souris sont calculés en faisant l'hypothèse que $p(t)$, $q(t)$ et $g(t)$ sont constantes sur des intervalles de 6 mois (l'étude a duré 3 ans en tout). On reviendra sur une partie des résultats de cette étude au chapitre 6.

## 5.2 Modèle de risques concurrents proportionnels

Dans l'article de Flehinger, Reiser et Yashchin de 1998, on adopte une approche basée sur l'analyse de survie avec des risques concurrents proportionnels pour trouver des estimateurs du maximum de vraisemblance lorsque des causes de panne sont masquées. Le but de l'article est de faire des inférences statistiques sur :

1. les fonctions de survie marginales de chacun des risques concurrents.
2. les fonctions d'incidence cumulée.
3. les probabilités de diagnostic : la probabilité, pour un groupe masqué donné, que chacun des risques concurrents faisant partie de ce groupe masqué soit responsable de la panne.

Ils montrent aussi comment ils peuvent améliorer la puissance statistique de leur modèle lorsque des données de deuxième étape sont utilisées.

Pour construire le modèle, on considère une situation pour laquelle des systèmes peuvent tomber en panne à cause de certains risques concurrents proportionnels. Les données de survie proviennent d'un échantillon de systèmes pour lesquels on observe des temps de panne ou de censure à droite. La panne d'un système est causée par seulement un risque concurrent, mais on ne peut pas toujours déterminer lequel. De plus, un échantillon des systèmes dont la cause de panne est masquée à la première étape est envoyé à la deuxième étape. Cette façon de faire avait été explorée par Flehinger, Reiser et Yashchin (1996) afin de ne pas être obligé d'émettre d'hypothèses trop fortes sur les données. À cette deuxième étape, la vraie cause de panne des systèmes de l'échantillon est trouvée grâce à une inspection plus approfondie de ceux-ci. La taille de l'échantillon envoyé en deuxième étape dépend des coûts encourus et du temps qui doit être alloué à l'inspection détaillée des systèmes.

On énonce d'abord quelques éléments de notation et des définitions afin de construire la fonction de vraisemblance pour les données de *première étape*. Chaque groupe masqué est dénoté par $g$ et le fait que le $j$ème risque concurrent soit contenu dans le groupe $g$ s'écrit $j \subset g$ ou $g \supset j$. Voici la notation utilisée pour définir les éléments qui feront partie d'un éventuel jeu de données :

$n$ : Nombre de systèmes observés.
$J$ : Nombre de causes de panne possibles.
$n_j$ : Nombre de pannes dues à la cause $j$.
$t_i^{(j)}$ : $i$ème temps de panne dû au risque $j$ $(i = 1, ..., n_j)$.
$n_g$ : Nombre de pannes masquées dans le groupe $g$.
$t_i^{(g)}$ : $i$ème temps de panne du groupe masqué $g$ $(i = 1, ..., n_g)$.
$n_c$ : Nombre de temps de censure.
$t_i^{(c)}$ : $i$ème temps de censure $(i = 1, ..., n_c)$.
$t_i$ : $i$ème temps de panne $(i = 1, ..., n - n_c)$,

où $n - n_c = \sum_{j=1}^{J} n_j + \sum_g n_g$ est le nombre de pannes au total.

Voici maintenant les paramètres du modèle :

$f(t)$ : Fonction de densité totale.
$S(t)$ : Fonction de survie totale.
$\lambda_j(t)$ : Fonction de risque spécifique à la cause de décès $j$.
$f_j(t)$ : Fonction de densité marginale associée à la fonction de risque spécifique, $\lambda_j(t)$.
$S_j(t)$ : Fonction de survie marginale associée à la fonction de risque spécifique, $\lambda_j(t)$.
$P_j$ : Probabilité d'identification. C'est la probabilité qu'une panne d'un système soit correctement identifiée comme étant causée par la cause $j$ à la première étape.
$P_{g|j}$ : Probabilité de masque. C'est la probabilité que la cause de panne soit dans le groupe $g$ à la première étape sachant que la panne est due à la cause $j$.

On suppose que $P_j$ et $P_{g|j}$ ne sont pas dépendantes du temps. La fonction de risque spécifique est définie par l'équation (4.1), tandis que les fonctions de densité et de survie totales sont données par les équations (4.4) et (4.3), respectivement. Dans l'article de Flehinger, Reiser et Yashchin (1998), on se place sous l'hypothèse des risques concurrents indépendants. Sous cette hypothèse, $f_j(t)$ et $S_j(t)$ sont vues comme des fonctions marginales du risque $j$. $S_j(t)$ est définie par l'équation (4.7) et

$f_j(t) = \frac{-\partial}{\partial t} S_j(t) = \lambda_j(t) \exp\{-\int_0^t \lambda_j(u)du\}$. De plus, la fonction de survie totale peut s'écrire comme le produit des $S_j(t)$, comme on l'a vu à l'équation (4.8).

Sous ces hypothèses et en utilisant la notation introduite ci-haut et les développements du chapitre 4, la fonction de vraisemblance pour l'étape 1 peut être enfin déterminée. Celle-ci doit tenir compte des temps de censure, des temps de panne dont la cause est identifiée à la première étape et des temps de panne placés dans des groupes masqués. Alors la fonction de vraisemblance s'écrit comme suit :

$$L = A \times B, \tag{5.12}$$

où $A$ dénote la contribution des systèmes censurés et de ceux dont on connaît la cause de panne à la première étape (c'est la contribution usuelle lorsqu'il n'y a pas de données masquées si on enlève $P_j$) :

$$A = \left\{ \prod_{i=1}^{n_c} S(t_i^{(c)}) \prod_{j=1}^{J} \prod_{i=1}^{n_j} \left[ P_j f_j(t_i^{(j)}) \prod_{l \neq j} S_l(t_i^{(j)}) \right] \right\}, \tag{5.13}$$

et $B$ dénote la contribution des systèmes dont la cause de panne est masquée dans un groupe masquant $g$ :

$$B = \left\{ \prod_{g} \prod_{i=1}^{n_g} \left[ \sum_{j \subset g} P_{g|j} f_j(t_i^{(g)}) \prod_{l \neq j} S_l(t_i^{(g)}) \right] \right\}. \tag{5.14}$$

L'hypothèse des risques concurrents proportionnels implique que

$$\begin{aligned} S_j(t) &= \{S(t)\}^{F_j} \text{ et } f_j(t) = F_j \{S(t)\}^{F_j - 1} f(t), \\ j &= 1, ..., J, \quad \text{et} \quad \sum_{j=1}^{J} F_j = 1, \end{aligned} \tag{5.15}$$

où $F_j$ représente la probabilité de décéder de la cause $j$.

En substituant (5.15) dans les équations (5.13) et (5.14) et après avoir réarrangé ces dernières, on trouve la fonction de vraisemblance suivante :

$$L = L_1^{(1)} \times L_2, \tag{5.16}$$

où

$$\begin{aligned} L_1^{(1)} &= \prod_{j=1}^{J} (P_j F_j)^{n_j} \prod_{g} \left( \sum_{j \subset g} P_{g|j} F_j \right)^{n_g}, \\ L_2 &= \left\{ \prod_{i=1}^{n_c} S(t_i^{(c)}) \right\} \sum_{i=1}^{n-n_c} f(t_i). \end{aligned}$$

Notons que $L_2$ est la fonction de vraisemblance des données de survie du système et c'est la vraisemblance usuelle associée à l'estimation de survie de Kaplan-Meier. Par contre, $L_1^{(1)}$ est une fonction de vraisemblance profilée pour les paramètres $F_j, P_j$ et $P_{g|j}$. Ces derniers ont les contraintes suivantes :

$$0 \leq F_j \leq 1, \;\; 0 \leq P_j \leq 1, \;\; 0 \leq P_{g|j} \leq 1, \tag{5.17}$$

$$\sum_{j=1}^{J} F_j = 1, \tag{5.18}$$

$$P_j \; + \sum_{g \supset j} P_{g|j} = 1 \;\; (j = 1, ..., J). \tag{5.19}$$

Si on se base sur les contraintes (5.17) à (5.19), on peut déduire que

$$\sum_{j=1}^{J} P_j \; F_j \; + \sum_{g} \sum_{j \subset g} P_{g|j} \, F_j = \sum_{j=1}^{J} F_j \left( P_j + \sum_{g \supset j} P_{g|j} \right) = \sum_{j=1}^{J} F_j = 1. \tag{5.20}$$

Puisque tous les facteurs de $L_1^{(1)}$ sont situés entre 0 et 1 et que leurs sommes sont égales à 1, alors $L_1^{(1)}$ a la forme d'une vraisemblance multinomiale. Par contre, cette fonction de vraisemblance est surparamétrisée. Ainsi, les estimateurs du maximum de vraisemblance des paramètres ne peuvent pas être calculés si $n_g$ est différent de 0. Pour estimer les fonctions de survie spécifiques à chacun des risques concurrents, $S_j(t)$, on doit estimer la fonction de survie du système, $S(t)$, et les probabilités $F_j, \;\; j = 1, ..., J$ (voir l'équation (5.15)). L'estimateur de $S(t)$ de Kaplan-Meier peut être facilement obtenu, mais à cause de la surparamétrisation, l'estimation de $F_j$ devient problématique. Une solution envisagée et intéressante pour résoudre le problème est le passage à une deuxième étape pour un échantillon des systèmes masqués à la première étape.

Afin de pouvoir écrire la nouvelle fonction de vraisemblance avec les informations obtenues à la deuxième étape, on doit introduire de nouveaux éléments de notation.

$\mathcal{P}$ : La probabilité avec laquelle un échantillon particulier des systèmes masqués à l'étape 1 est choisi pour la deuxième étape. Cette probabilité ne dépend que des données observées à la première étape.

$n_{g,j}$ : Nombre de systèmes en panne pour lesquels la cause de panne est restreinte au groupe masqué $g$ à la première étape et qui est diagnostiquée comme étant la cause $j$ à la deuxième étape.

$n_g^+ = \sum_{j=1}^{J} n_{g,j}$ : Nombre total de systèmes avec le groupe masqué $g$ qui sont envoyés à la deuxième étape pour trouver la vraie cause de panne $j$.

$\tilde{n}_g$ : Nombre de systèmes en panne dont la cause de panne est restreinte au groupe masqué $g$ à la première étape et pour lesquels il n'y a pas de deuxième étape ($n_g^+ + \tilde{n}_g = n_g$).

$A_j = P_j F_j$ : Probabilité qu'une panne soit causée par le risque $j$ et qu'elle soit identifiée à l'étape 1.

$B_g = \sum_{j \subset g} P_{g|j} F_j$ : Probabilité qu'une panne soit masquée et qu'elle soit restreinte au groupe $g$ à l'étape 1.

$\pi_{j|g} = P_{g|j} F_j / B_g$ : Probabilité de diagnostic. C'est la probabilité qu'une panne restreinte au groupe $g$ à l'étape 1 soit en fait causée par le risque $j$.

Tout comme l'équation (5.20) mais sous une autre notation, on peut écrire que

$$\sum_{j=1}^{J} A_j + \sum_{g} B_g = 1. \tag{5.21}$$

De plus, pour tout $g$, on a la contrainte suivante :

$$\sum_{j \subset g} \pi_{j|g} = 1. \tag{5.22}$$

On remarque que l'étape 2 ne divulgue pas de nouvelle information concernant la durée de vie des systèmes. Ainsi, la partie $L_2$ de la fonction de vraisemblance reste la même qu'à la première étape. On a d'ailleurs déjà mentionné qu'on pouvait trouver l'estimateur $\hat{S}(t)$ de Kaplan-Meier à l'étape 1. Les informations recueillies à la deuxième étape sont utilisées pour modifier $L_1^{(1)}$ en $L_1$ comme suit :

$$\begin{aligned} L_1 &= \prod_{j=1}^{J} (P_j F_j)^{n_j} \prod_{g} \left( \sum_{j \subset g} P_{g|j} F_j \right)^{\tilde{n}_g} \prod_{j=1}^{J} \prod_{g \supset j} \left( P_{g|j} F_j \right)^{n_{g,j}} \times \mathcal{P} \\ &= \left( \prod_{j=1}^{J} A_j^{n_j} \prod_{g} B_g^{n_g} \right) \times \prod_{g} \left( \prod_{j \subset g} \pi_{j|g}^{n_{g,j}} \right) \times \mathcal{P}. \end{aligned} \tag{5.23}$$

En utilisant les contraintes (5.21) et (5.22) et la fonction (5.23), on peut maintenant écrire les estimateurs du maximum de vraisemblance de $A_j$, $B_g$ et $\pi_{j|g}$ comme suit :

$$\hat{A}_j = \frac{n_j}{n - n_c}, \ \hat{B}_g = \frac{n_g}{n - n_c}, \ \hat{\pi}_{j|g} = \frac{n_{g,j}}{n_g^+}.$$

De plus, de la contrainte (5.19), on trouve facilement que

$$1 = P_j + \sum_{g \supset j} P_{g|j} \quad \Leftrightarrow \quad F_j = P_j F_j + \sum_{g \supset j} P_{g|j} F_j$$

$$\Leftrightarrow \quad F_j = A_j \; + \sum_{g \supset j} B_g \pi_{j|g}.$$

Ainsi, on peut trouver les EMV de $F_j$, $P_j$ et $P_{g|j}$ par la propriété d'invariance du théorème 2.1. En effet,

$$\begin{aligned}
\hat{F}_j &= \hat{A}_j \; + \sum_{g \supset j} \hat{B}_g \hat{\pi}_{j|g} = \frac{n_j + \sum_{g \supset j} n_g n_{g,j}/n_g^+}{n - n_c}, \\
\hat{P}_j &= \frac{\hat{A}_j}{\hat{F}_j} = \frac{n_j}{n_j + \sum_{g \supset j} n_g n_{g,j}/n_g^+}, \\
\hat{P}_{g|j} &= \frac{\hat{\pi}_{j|g} \hat{B}_g}{\hat{F}_j} = \frac{n_{g,j}\, n_g}{n_g^+ (n_j + \sum_{g \supset j} n_g n_{g,j}/n_g^+)}.
\end{aligned}$$

On peut aussi estimer la fonction de survie pour chacun des risques de la façon suivante :

$$\hat{S}_j(t) = \hat{S}(t)^{\hat{F}_j}.$$

Ainsi, on voit que les données de deuxième étape permettent d'estimer les $F_j$, $j = 1, ..., J$, lorsqu'on fait l'hypothèse des risques concurrents proportionnels. En effet, les estimateurs de $F_j$ ne sont pas identifiables lorsqu'il n'y a pas de données de deuxième étape.

Flehinger, Reiser et Yashchin (1998) présentent un exemple sur le cancer du poumon en guise d'illustration. Il y a 2 causes de décès possibles : le cancer du poumon ou toute autre cause. On compare les estimations de $S_j(t) = S(t)^{F_j}$ pour 3 scénarios différents. Le premier envoie 20% des sujets masqués à la deuxième étape. Le second n'en envoie pas, il attribue donc tous les masques soit au cancer du poumon ou soit aux autres causes ; en effet, lorsqu'il n'y a pas de deuxième étape, avec le modèle des risques proportionnels de Flehinger, Reiser et Yashchin (1998), on doit distribuer les données masquées sur les causes de pannes pour pouvoir estimer $F_j$. Le dernier scénario choisit 5% des sujets pour la deuxième étape. Ces estimateurs sont comparés à ceux de Kaplan-Meier lorsqu'il n'y a pas de données masquées. Le but est donc d'être le plus près possible de cette estimation. Les auteurs concluent que même un petit échantillon de 5% envoyé à la deuxième étape améliore grandement l'estimation de $S_j(t)$ comparé à l'envoi d'aucun échantillon.

## 5.3 Modélisation paramétrique des données de survie

Étant donné que l'hypothèse des risques proportionnels est une hypothèse forte et qu'elle ne peut pas toujours être justifiée en pratique, Flehinger, Reiser et Yashchin (2002) ont développé un modèle avec une distribution de survie paramétrique. On cherche donc à trouver les EMV des paramètres de la distribution de survie ainsi que des probabilités d'identification et de masque à l'aide d'un modèle paramétrique basé sur les risques concurrents avec un échantillon de systèmes masqués envoyés à la deuxième étape.

On rajoute la notation suivante à celle déjà existante de la section 5.2 :

$n_j^*$ : Nombre de systèmes pour lesquels on a prouvé que les pannes venaient du risque $j$.

$t_i^{(j^*)}$ : $i$ème temps de panne parmi tous les temps de panne qui ont été identifiés comme étant causés par le risque $j$ soit à la première ou à la deuxième étape $(i = 1, ..., n_j^*)$.

$\tilde{t}_i^{(g)}$ : $i$ème temps de panne parmi tous les temps de panne non-résolus dans le groupe masqué $g$ $(i = 1, ..., \tilde{n}_g)$.

$\bar{t}_i$ : $i$ème temps de panne ou de censure $(i = 1, ..., n)$.

$t_i^{(g,j)}$ : $i$ème temps de panne des systèmes masqués dans le groupe $g$ à la première étape et dont la vraie cause $j$ est trouvée à la deuxième étape $(i = 1, ..., n_{g,j})$.

$\beta_j$ : Vecteur de paramètres inconnus associés à la distribution de survie de la cause $j$.

$\lambda_j(t|\beta_j)$ : Fonction de risque spécifique à la cause de décès $j$.

$f_j(t|\beta_j)$ : Fonction de densité marginale associée à la fonction de risque spécifique, $\lambda_j(t|\beta_j)$.

$S_j(t|\beta_j)$ : Fonction de survie marginale associée à la fonction de risque spécifique, $\lambda_j(t|\beta_j)$.

$f(t|\beta_1, ..., \beta_J)$ : Fonction de densité totale.

$S(t|\beta_1, ..., \beta_J)$ : Fonction de survie totale.

On a encore la relation

$$S(t|\beta_1, ..., \beta_J) = \prod_{j=1}^{J} S_j(t|\beta_j) \tag{5.24}$$

et sous l'hypothèse des causes de panne indépendantes, on a que

$$f_j(t|\beta_j) = \frac{-\partial}{\partial t} S_j(t|\beta_j) = \lambda_j(t|\beta_j) \exp\left\{ - \int_0^t \lambda_j(u|\beta_j) du \right\} \tag{5.25}$$

et donc que

$$\lambda_j(t|\beta_j) = \frac{f_j(t|\beta_j)}{S_j(t|\beta_j)}. \tag{5.26}$$

La fonction de risque cumulé pour la cause $j$ s'écrit

$$\Lambda_j(t|\beta_j) = \int_0^t \lambda_j(u|\beta_j) du. \tag{5.27}$$

Finalement, la probabilité de diagnostic, qui dépend aussi du temps $t$ lorsque les $\lambda_j(t|\beta_j)$, $j = 1, ..., J$, ne sont pas proportionnels, est donnée par

$$\pi_{j|g}(t) = \frac{P_{g|j} \lambda_j(t|\beta_j)}{\sum_{r \subset g} P_{g|r} \lambda_r(t|\beta_r)}. \tag{5.28}$$

En s'inspirant des fonctions de vraisemblance de la section 5.2 et en utilisant les informations des deux étapes, on écrit maintenant la fonction de vraisemblance de la façon suivante :

$$\begin{aligned} L \;=\; & \prod_{i=1}^{n_c} S(t_i^{(c)}|\beta_1, ..., \beta_J) \prod_{j=1}^{J} \prod_{i=1}^{n_j} \left\{ P_j f_j(t_i^{(j)}|\beta_j) \prod_{l \neq j} S_l(t_i^{(j)}|\beta_l) \right\} \\ & \times \prod_{g} \prod_{i=1}^{\tilde{n}_g} \left\{ \sum_{r \subset g} P_{g|r} f_r(\tilde{t}_i^{(g)}|\beta_r) \prod_{l \neq r} S_l(\tilde{t}_i^{(g)}|\beta_l) \right\} \\ & \times \prod_{j=1}^{J} \prod_{g \supset j} \prod_{i=1}^{n_{g,j}} \left\{ P_{g|j} f_j(t_i^{(g,j)}|\beta_j) \prod_{l \neq j} S_l(t_i^{(g,j)}|\beta_l) \right\}. \end{aligned} \tag{5.29}$$

Dans cette équation, la première ligne représente la contribution des systèmes qui ont soit des temps de censure ou des temps de panne dont la cause a été décelée à

la première étape. La deuxième ligne représente, pour sa part, les systèmes dont la cause de panne est masquée à l'étape 1 et dont la vraie cause de panne n'est jamais trouvée. La troisième ligne montre la contribution des systèmes dont les causes sont masquées à la première étape mais démasquées à la deuxième étape. Cette fonction de vraisemblance est similaire à celle trouvée pour la première étape, à l'équation (5.12), dans l'article de Flehinger, Reiser et Yashchin de 1998. Par contre, dans cette dernière ils ne tenaient pas compte des données de deuxième étape, alors la troisième ligne de l'équation (5.29) n'était pas présente.

La fonction de vraisemblance écrite ci-dessus peut se réécrire de la façon suivante en utilisant les équations (5.24) et (5.26) :

$$\begin{aligned} L \;=\; & \prod_{i=1}^{n} S(\bar{t}_i|\beta_1,...,\beta_J) \prod_{j=1}^{J} \left( P_j^{n_j} \prod_{g\supset j} P_{g|j}^{n_{g,j}} \right) \\ & \times \prod_{j=1}^{J} \prod_{i=1}^{n_j^*} \lambda_j(t_i^{(j^*)}|\beta_j) \prod_{g} \prod_{i=1}^{\tilde{n}_g} \sum_{r\subset g} \left( P_{g|r} \lambda_r(\tilde{t}_i^{(g)}|\beta_r) \right). \end{aligned} \tag{5.30}$$

Grâce à la relation $\ln(S_j(t|\beta_j)) = -\Lambda_j(t|\beta_j)$ et à l'équation (5.24), on peut exprimer la fonction de log-vraisemblance comme suit :

$$\begin{aligned} l \;=\; \ln(L) = \; & - \sum_{i=1}^{n} \sum_{j=1}^{J} \Lambda_j(\bar{t}_i|\beta_j) + \sum_{j=1}^{J} \left( n_j \ln(P_j) + \sum_{g\supset j} n_{g,j} \ln(P_{g|j}) \right) \\ & + \sum_{j=1}^{J} \sum_{i=1}^{n_j^*} \ln \left( \lambda_j(t_i^{(j^*)}|\beta_j) \right) \\ & + \sum_{g} \sum_{i=1}^{\tilde{n}_g} \ln \left( \sum_{r\subset g} P_{g|r} \lambda_r(\tilde{t}_i^{(g)}|\beta_r) \right). \end{aligned} \tag{5.31}$$

Lorsqu'aucun des paramètres ne peut être traité comme une quantité connue, on maximise (5.31) pour obtenir les EMV de $\beta_j, P_j$ et $P_{g|j}$, $j = 1,...,J$. Une approche itérative est utilisée pour ce problème. L'idée est de trouver d'abord les EMV de $P_j$ et de $P_{g|j}$ pour des valeurs données de $\beta_j$. Ensuite, on calcule les EMV des $\beta_j$ sachant $P_j$ et $P_{g|j}$. Pour estimer tous les paramètres à la fois, on alterne itérativement entre ces deux procédures. En effet, on assigne d'abord des valeurs initiales arbitraires à $P_j$ et à $P_{g|j}$. Ensuite, on calcule l'EMV de $\beta_j$ étant données ces valeurs initiales. Avec les EMV des $\beta_j$ qu'on vient de trouver, on calcule alors les EMV des $P_j$ et des $P_{g|j}$. On répète ces étapes de façon itérative jusqu'à ce que la différence entre les valeurs d'une nouvelle itération et de la précédente soit très petite.

Si on n'a que des données de première étape, on doit poser $n_{g,j} = 0$. Ainsi, on a la fonction de vraisemblance appropriée et on peut estimer les paramètres avec la même

procédure. On doit faire attention lorsqu'on n'a que des données de première étape, car on ne peut pas toujours identifier le modèle. Par exemple, on ne peut pas estimer les paramètres de base lorsqu'il y a 2 causes de panne avec des risques proportionnels. De plus, avec des modèles paramétriques on sera généralement capable d'estimer les paramètres reliés à la cause $j$ même si aucune panne n'est identifiée à cette cause (que ce soit à la première ou à la deuxième étape).

Pour l'estimation des probabilités d'identification, $P_j$, et de masque, $P_{g|j}$, on maximise $l = \ln(L)$ de l'équation (5.31) pour ces probabilités lorsque les $\beta_j$ sont connus. Pour ce faire, on doit introduire le Lagrangien en utilisant la contrainte (5.19) :

$$G = l + \sum_{j=1}^{J} \alpha_j \left( 1 - P_j - \sum_{g \supset j} P_{g|j} \right). \tag{5.32}$$

Les contraintes $0 \leq P_j \leq 1$ et $0 \leq P_{g|j} \leq 1$ ne sont pas représentées explicitement dans le Lagrangien puisque la solution de celui-ci satisfait automatiquement ces contraintes.

La dérivée de $G$ est prise par rapport à $P_j$ et à $P_{g|j}$ et ensuite on égale ces équations à zéro pour tout $j$. On obtient donc les équations suivantes :

$$\begin{aligned} \frac{\partial G}{\partial P_j} &= \frac{n_j}{P_j} - \alpha_j = 0, \\ \frac{\partial G}{\partial P_{g|j}} &= \frac{n_{g,j}}{P_{g|j}} + \frac{1}{P_{g|j}} \sum_{i=1}^{\tilde{n}_g} \pi_{j|g}(\tilde{t}_i^{(g)}) - \alpha_j = 0, \end{aligned} \tag{5.33}$$

pour tout groupe $g$ contenant $j$, et $\alpha_j$ estime le nombre total de pannes qui sont dues à la cause $j$.

Sous l'hypothèse que les $\beta_j$ sont connus, on peut résoudre le système d'équations (5.33) à l'aide de la procédure itérative suivante :

On commence avec des valeurs initiales arbitraires de $P_j$ et $P_{g|j}$, $j = 1, ..., J$ (par exemple, pour tout $j$ et $g \supset j$, on peut poser $P_j = P_{g|j} = 1/(1 + \sum_{g \supset j} 1)$).

**1.** Pour tout $j = 1, 2, ..., J$, calculer la valeur estimée de $\alpha_j$ qui correspond aux valeurs courantes de $P_j$ et $P_{g|j}$ :

$$\alpha_j = n_j^* + \sum_{g \supset j} \sum_{i=1}^{\tilde{n}_g} \pi_{j|g}(\tilde{t}_i^{(g)}). \tag{5.34}$$

**2.** Mettre à jour $P_j$ et $P_{g|j}$ :

$$P_j^{(nouv.)} = n_j/\alpha_j, \tag{5.35}$$

$$P_{g|j}^{(nouv.)} = (1/\alpha_j)\left[n_{g,j} + \sum_{i=1}^{\tilde{n}_g} \pi_{j|g}(\tilde{t}_i^{(g)})\right], \tag{5.36}$$

et retourner à l'étape 1. On donne la solution lorsque la différence entre les nouvelles valeurs des probabilités d'identification et de masque et les précédentes sont assez petites.

Cette procédure itérative est une version de l'algorithme EM. Alors on peut être assuré que l'algorithme converge et que le maximum résultant de cet algorithme est unique puisque la fonction de log-vraisemblance est concave et que toutes les contraintes sont linéaires.

Afin d'obtenir une estimation des paramètres de distribution, $\beta_j$, sachant $P_j$ et $P_{g|j}$, on pose les dérivées de (5.31) par rapport aux $\beta_j$, $j = 1, ..., J$, égales à zéro. On trouve alors l'équation suivante :

$$\begin{aligned}
&\sum_{i=1}^{n_j^*} \frac{\partial}{\partial\beta_j} \ln\left(\lambda_j(t_i^{(j^*)}|\beta_j)\right) + \sum_{g\supset j}\sum_{i=1}^{\tilde{n}_g} \frac{P_{g|j}}{\sum_{r\subset g} P_{g|r}\lambda_r(\tilde{t}_i^{(g)}|\beta_r)} \frac{\partial}{\partial\beta_j}\lambda_j(\tilde{t}_i^{(g)}|\beta_j) \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial\beta_j}\Lambda_j(\bar{t}_i|\beta_j),\ j = 1, ..., J \\
\Leftrightarrow \\
&\sum_{i=1}^{n_j^*} \frac{\partial}{\partial\beta_j} \ln\left(\lambda_j(t_i^{(j^*)}|\beta_j)\right) + \sum_{g\supset j}\sum_{i=1}^{\tilde{n}_g} \frac{P_{g|j}\,\lambda_j(\tilde{t}_i^{(g)}|\beta_r)}{\sum_{r\subset g} P_{g|r}\lambda_r(\tilde{t}_i^{(g)}|\beta_r)} \frac{\partial}{\partial\beta_j} \ln\left(\lambda_j(\tilde{t}_i^{(g)}|\beta_j)\right) \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial\beta_j}\Lambda_j(\bar{t}_i|\beta_j),\ j = 1, ..., J \qquad (5.37) \\
\Leftrightarrow \\
&\sum_{i=1}^{n_j^*} \frac{\partial}{\partial\beta_j} \ln\left(\lambda_j(t_i^{(j^*)}|\beta_j)\right) + \sum_{g\supset j}\sum_{i=1}^{\tilde{n}_g} \pi_{j|g}(\tilde{t}_i^{(g)}) \frac{\partial}{\partial\beta_j} \ln\left(\lambda_j(\tilde{t}_i^{(g)}|\beta_j)\right) \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial\beta_j}\Lambda_j(\bar{t}_i|\beta_j),\ j = 1, ..., J. \qquad (5.38)
\end{aligned}$$

L'équivalence entre (5.37) et (5.38) vient de l'équation (5.28). On peut trouver les estimateurs $\hat{\beta}_j$, $j = 1, ..., J$ à l'aide d'un processus itératif. Le choix de ce dernier dépend de la forme paramétrique de la fonction de risque.

Pour leur part, les probabilités de diagnostic, $\pi_{j|g}(t)$, $j = 1, .., J$, peuvent être estimées en substituant $\hat{P}_{g|j}$ et $\hat{\beta}_j$ dans l'équation (5.28). Notons que sous l'hypothèse

de risques proportionnels, les probabilités de diagnostic ne dépendent pas du temps $t$ auquel la panne masquée est observée, comme on l'a vu dans la section 5.2.

On peut maintenant appliquer la méthode d'estimation de paramètres décrite dans cet article à une distribution en particulier. En connaissant la fonction de survie et la fonction de risque de la loi en question, il ne reste plus qu'à appliquer la méthode.

Voyons rapidement un exemple avec la loi de Weibull. Les fonctions de survie, de risque cumulé et de risque de la cause $j$ sont, respectivement,

$$S_j(t) = \exp\Big\{-(t/\theta_j)^{\delta_j}\Big\},\ j=1,...,J, \tag{5.39}$$

$$\Lambda_j(t) = \left(\frac{t}{\theta_j}\right)^{\delta_j},\ j=1,...,J, \tag{5.40}$$

$$\lambda_j(t) = \frac{\delta_j}{\theta_j}\left(\frac{t}{\theta_j}\right)^{\delta_j-1},\ j=1,...,J. \tag{5.41}$$

La fonction de log-vraisemblance peut donc être représentée de la façon suivante, en remplaçant les équations (5.39) à (5.41) dans l'équation (5.31) :

$$\begin{aligned} l = & -\sum_{i=1}^{n}\sum_{j=1}^{J}(\bar{t}_i/\theta_j)^{\delta_j} + \sum_{j=1}^{J}\left[n_j\ln(P_j) + \sum_{g\supset j} n_{g,j}\ln(P_{g|j})\right] \\ & + \sum_{j=1}^{J}\left[n_j^*\ln\left(\delta_j/\theta_j\right) + (\delta_j-1)\sum_{i=1}^{n_j^*}\ln\left(t_i^{(j^*)}/\theta_j\right)\right] \\ & + \sum_{g}\sum_{i=1}^{\tilde{n}_g}\ln\left(\sum_{r\subset g}P_{g|r}\left(\delta_r/\theta_r\right)\left(\tilde{t}_i^{(g)}/\theta_r\right)^{\delta_r-1}\right). \end{aligned} \tag{5.42}$$

On estime d'abord les probabilités d'identification, $P_j$, et de masque, $P_{g|j}$, pour la loi de Weibull. Sachant $(\theta_j,\delta_j),\ j=1,...,J$, on peut estimer $P_j$ et $P_{g|j}$ à l'aide de la procédure itérative décrite plus haut (équations (5.34) à (5.36)).

Dans ce cas, la probabilité de diagnostic s'écrit

$$\pi_{j|g} = \frac{P_{g|j}(\delta_j/\theta_j)(t/\theta_j)^{\delta_j-1}}{\sum_{r\subset g}P_{g|r}(\delta_r/\theta_r)(t/\theta_r)^{\delta_r-1}}. \tag{5.43}$$

On estime ensuite les paramètres de la distribution, $(\theta_j,\delta_j)$. Pour l'estimation des $\theta_j,\ j=1,...,J$, de l'équation (5.38), en remplaçant les fonctions de risque et de risque cumulée par celles de la loi de Weibull, on trouve facilement que

$$\alpha_j = \sum_{i=1}^{n}(\bar{t}_i/\theta_j)^{\delta_j},\ \ j=1,...,J, \tag{5.44}$$

où $\alpha_j$ est défini par l'équation (5.34).

En supposant que tous les autres paramètres sont connus, on peut isoler $\theta_j$, $j = 1, ..., J$, dans l'équation précédente et résoudre les équations grâce à une procédure itérative. On utilise alors l'équation suivante :

$$\theta_j^{(nouv.)} = \left( \frac{\sum_{i=1}^{n} (\bar{t}_i)^{\delta_j}}{\alpha_j} \right)^{1/\delta_j}, \quad j = 1, ..., J. \tag{5.45}$$

Pour estimer les $\delta_j$, $j = 1, ..., J$, les équations sont plutôt complexes. On suggère de se référer à l'article de Flehinger, Reiser et Yashchin (2002) pour plus de détails.

Finalement, l'approche basée sur les fonctions de survie paramétriques de Flehinger, Reiser et Yashchin (2002) a l'avantage de produire un modèle estimable sans faire d'hypothèses trop fortes sur les probabilités masquées telle que l'hypothèse de symétrie qui stipule que la probabilité qu'un système soit masqué à la première étape ne dépend pas de la vraie cause de panne. L'approche paramétrique est d'un côté plus générale que celle de la section 5.2 puisqu'on ne fait pas d'hypothèse de proportionnalité sur les risques concurrents, mais d'un autre côté moins générale puisqu'on force un modèle paramétrique. Par contre, plusieurs distributions sont possibles, alors il est souvent facile de trouver un modèle qui convient bien aux données.

## 5.4 Modèle de risques concurrents constants par intervalles

Dans l'article de Craiu et Duchesne (2004), on utilise l'algorithme EM pour estimer les paramètres d'un modèle de survie avec des risques concurrents constants par intervalles lorsque des causes de panne sont masquées et qu'il y a des données de deuxième étape. En faisant l'hypothèse que la fonction de risque spécifique est constante par paliers, on peut faire l'estimation du maximum de vraisemblance de certains paramètres sans émettre d'hypothèses trop fortes sur les données telles l'indépendance des risques concurrents, la proportionalité des risques ou la symétrie. Les paramètres à estimer sont les fonctions de risque spécifique, les probabilités de masque et les probabilités de diagnostic.

Tout comme dans les section 5.2 et 5.3, on se place dans un contexte où certains

systèmes qui ont une cause de panne masquée passent à une deuxième étape pour une investigation plus approfondie.

Supposons que $n$ systèmes indépendants sont observés sur une période de temps de longueur $\tau$ et que chacun de ces systèmes peut tomber en panne d'une des $J$ causes possibles. La collecte des données se fait en deux étapes. À la première étape, on observe un des résultats suivants pour chacun des $i$ systèmes, $i = 1, ..., n$ :

1. $i$ tombe en panne de la cause $j_i$ au temps $t_i$.
2. $i$ tombe en panne d'une des causes du groupe de causes $g_i \subset \{1, ...J\}$ au temps $t_i$.
3. $i$ n'est pas encore tombé en panne au temps $t_i$.

Un échantillon des systèmes faisant partie des groupes masqués, c'est-à-dire une partie des systèmes qui ont eu le résultat numéro 2, passe à la deuxième étape. Pour ces systèmes, la cause exacte de panne est déterminée.

On définit les « données complètes » comme étant les données qu'on obtiendrait si tous les systèmes avec un temps de panne masqué étaient envoyés à la deuxième étape. Alors l'observation du système $i$ dans ce jeu de données complètes est donnée par

$$(t_i, \gamma_{ig_1}, \ldots, \gamma_{ig_{G+J}}, \delta_{i1}, \ldots, \delta_{iJ}), \tag{5.46}$$

où les composantes de l'observation $i$ sont définies comme suit :

$t_i$ : Temps de panne ou de censure du système $i$.

$\gamma_{ig}$ : Indicatrice qui vaut 1 si la cause de panne du système $i$ est masquée dans le groupe $g$ à la première étape, 0 sinon. Si la cause de panne est connue à la première étape et que c'est la cause $j$, alors on dit qu'elle est masquée dans $g = \{j\}$.

$\delta_{ij}$ : Indicatrice qui vaut 1 si la cause de panne du système $i$ est $j$, 0 sinon. Si le système $i$ est censuré à droite, alors tous les $\delta_{ij},\ j = 1, ..., J$, prennent la valeur 0.

$G + J$ : Nombre total de groupes masqués, où les $G$ groupes masqués contiennent plus d'une cause de panne et les $J$ groupes consistent en une cause de panne individuelle.

Voici un exemple qui illustre la notation décrite ci-haut. Supposons qu'on a 3 causes de panne possibles. À la première étape, soit qu'on identifie la cause de panne (dans ce cas, la cause de panne est « masquée » dans l'un des 3 groupes suivants : $\{1\}$, $\{2\}$ ou

TAB. 5.3 – Exemple de notation pour quatre systèmes observés dans le temps, où · représente une donnée manquante.

| Système $i$ | $t_i$ | $\gamma_{i,\{1\}}$ | $\gamma_{i,\{2\}}$ | $\gamma_{i,\{3\}}$ | $\gamma_{i,\{1,2\}}$ | $\gamma_{i,\{1,2,3\}}$ | $\delta_{i,1}$ | $\delta_{i,2}$ | $\delta_{i,3}$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.5 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| 2 | 2.4 | 0 | 0 | 0 | 1 | 0 | · | · | 0 |
| 3 | 6.3 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| 4 | 4.1 | · | · | · | · | · | 0 | 0 | 0 |

$\{3\}$), soit que l'on sait que la panne est due à la cause 1 ou à la 2 (dans ce cas, la cause est masquée dans le groupe $\{1, 2\}$), ou soit que l'on sait que la panne est due à l'une des causes 1, 2 ou 3 (dans ce cas, la cause est masquée dans le groupe $\{1, 2, 3\}$). Pour le système 1, on observe une panne au temps 1.5 et on identifie dès la première étape que c'est la cause 3 qui est responsable de la panne. Ensuite, on observe que le système 2 tombe en panne au temps 2.4 et la cause de panne est masquée dans le groupe $\{1, 2\}$ à la première étape et il n'est pas envoyé à la deuxième étape. Le système 3 tombe en panne au temps 6.3 d'une cause masquée dans le groupe $\{1, 2, 3\}$ à la première étape et il est envoyé à la deuxième étape où on trouve que la panne est due à la cause 2. Finalement, le système 4 est censuré à droite au temps 4.1. Ces 4 observations peuvent être répertoriées comme dans le tableau 5.3.

Le modèle statistique utilisé est constitué de deux parties : une partie de risques concurrents qui implique les temps de panne et les causes de panne et une partie masquée qui comprend les probabilités de masque. La première partie est le modèle des risques concurrents expliqué au chapitre 4.

Soit $T$ le temps de panne et $J$ la cause de panne. Alors on peut écrire les définitions suivantes :

$N_{ij}(t) = I(T_i \leq t, J = j,\ i$ n'est pas censuré$)$ : Processus de dénombrement qui fait un saut de 1 au temps $t$ si une panne de cause $j$ du système $i$ est observée au temps $t$ $(i = 1, ..., n,\ j = 1, ..., J,\ t \in [0, \tau])$.

$Y_i(t) = I(T_i \geq t)$ : Indicatrice qui vaut 1 si le système $i$ est toujours à risque de tomber en panne au temps $t$ $(i = 1, ..., n,\ t \in [0, \tau])$.

$\bar{Y}(t) = \sum_{i=1}^{n} Y_i(t)$ : Nombre total de systèmes à risque de tomber en panne à $t$ $(t \in [0, \tau])$.

$\bar{N}_{\cdot j}(t) = \sum_{i=1}^{n} N_{ij}(t)$ : Processus de dénombrement qui compte le nombre total de pannes dues à la cause $j$ à ou avant $t$ $(j = 1, ..., J,\ t \in [0, \tau])$.

De l'équation (4.1), la fonction de risque spécifique à la cause $j$ est donnée par

$$\lambda_j(t) = \lim_{h \downarrow 0} \frac{P[t < T \leq t + h, J = j | T \geq t]}{h} \ , \ j = 1, ..., J. \tag{5.47}$$

Pour la partie masquée du modèle, étant donnée une cause de panne $j$, on suppose que les $\{\gamma_{ig} : g \in \{g_1, ..., g_{G+J}\}, g \ni j\}$ suivent une distribution multinomiale, avec un total de 1 et les probabilités données par les probabilités de masque définies comme suit :

$P_{g|j}(t)$ : Probabilité que la cause de panne soit masquée dans le groupe $g$ à la première étape sachant que la cause de panne au temps $t$ est $j$ ($j \in g$).

On aura aussi besoin des probabilités de diagnostic données par

$\pi_{j|g}(t)$ : Probabilité que le système soit tombé en panne de la cause $j$ sachant qu'il est tombé en panne au temps $t$ et qu'il est masqué dans le groupe $g$.

En utilisant la loi de Bayes, on obtient

$$\pi_{j|g}(t) = \frac{\lambda_j(t) P_{g|j}(t)}{\sum_{l \in g} \lambda_l(t) P_{g|l}(t)}. \tag{5.48}$$

Soit $\boldsymbol{\theta}$ le vecteur contenant les $\lambda_j(.)$, $j = 1, ..., J$ et les $P_{g_m|j}(.)$, $j = 1, ..., J$, $m = 1, ..., G + J$. Soit $\mathcal{G}_j = \{g : j \in g\}$ l'ensemble de tous les groupes masqués qui incluent la cause $j$ et $\mathcal{G}_j^* = \mathcal{G}_j / \{j\}$ l'ensemble de tous les groupes masqués qui incluent la cause $j$ sauf le groupe constitué du singleton $\{j\}$ lui-même.

La fonction de vraisemblance des données complètes est donnée par

$$\begin{aligned} L_C(\boldsymbol{\theta}) \quad = \quad & \prod_{i=1}^{n} \left\{ \left[ \prod_{j=1}^{J} \{\lambda_j(t_i)\}^{\delta_{ij}} \exp\left\{ - \int_0^{t_i} \lambda_j(t) dt \right\} \right] \right. \\ & \left. \times \prod_{j=1}^{J} \left[ \left(1 - \sum_{g \in \mathcal{G}_j^*} P_{g|j}(t_i)\right)^{\left(1 - \sum_{g \in \mathcal{G}_j^*} \gamma_{ig}\right)} \left(P_{g|j}(t_i)\right)^{\left(\sum_{g \in \mathcal{G}_j^*} \gamma_{ig}\right)} \right]^{\delta_{ij}} \right\}. \end{aligned} \tag{5.49}$$

La première partie de l'équation (5.49) représente la contribution des données de risques concurrents et elle est donnée par l'équation (4.11). La deuxième partie est la contribution des données masquées (la distribution multinomiale).

De l'équation (5.49), on trouve que la fonction de log-vraisemblance complète s'exprime de la façon suivante :

$$\begin{aligned}
l_C(\boldsymbol{\theta}) &= \sum_{i=1}^{n}\Bigg\{\Bigg[\sum_{j=1}^{J}\delta_{ij}\ln\Big(\lambda_j(t_i)\Big)-\sum_{j=1}^{J}\int_0^{t_i}\lambda_j(t)dt\Bigg] \\
&\qquad +\sum_{j=1}^{J}\delta_{ij}\Bigg[\Big(1-\sum_{g\in\mathcal{G}_j^*}\gamma_{ig}\Big)\ln\Big(1-\sum_{g\in\mathcal{G}_j^*}P_{g|j}(t_i)\Big)+\sum_{g\in\mathcal{G}_j^*}\gamma_{ig}\ln\Big(P_{g|j}(t_i)\Big)\Bigg]\Bigg\} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{J}\Bigg\{\Bigg[\ln\Big(\lambda_j(t_i)\Big)\int_0^{\tau}dN_{ij}(t)-\int_0^{\tau}Y_i(t)\lambda_j(t)dt\Bigg] \\
&\qquad +\Bigg[\Big(1-\sum_{g\in\mathcal{G}_j^*}\gamma_{ig}\Big)\ln\Big(1-\sum_{g\in\mathcal{G}_j^*}P_{g|j}(t_i)\Big) \\
&\qquad\quad +\sum_{g\in\mathcal{G}_j^*}\gamma_{ig}\ln\Big(P_{g|j}(t_i)\Big)\Bigg]\int_0^{\tau}dN_{ij}(t)\Bigg\} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{J}\int_0^{\tau}\Bigg\{\Bigg[\ln\Big(\lambda_j(t)\Big)dN_{ij}(t)-Y_i(t)\lambda_j(t)dt\Big)\Bigg] \\
&\qquad +\Bigg[\Big(1-\sum_{g\in\mathcal{G}_j^*}\gamma_{ig}\Big)\ \ln\Big(1-\sum_{g\in\mathcal{G}_j^*}P_{g|j}(t)\Big) \\
&\qquad\quad +\sum_{g\in\mathcal{G}_j^*}\gamma_{ig}\ln\Big(P_{g|j}(t)\Big)\Bigg]dN_{ij}(t)\Bigg\}. \qquad (5.50)
\end{aligned}$$

Si une donnée est censurée à droite, la partie après la première addition de l'équation (5.50) disparaît puisque $\int_0^{\tau} dN_{ij}(t)$ prend une valeur de 0 dans ce cas. Par conséquent, on n'a pas besoin des $\gamma_{ig}$ pour les données censurées à droite.

Des données sont masquées lorsque des $\delta_{ij}$ sont manquantes. Soit $\mathcal{M}$ l'ensemble des systèmes qui ont une cause de panne masquée à la première étape et qui ne vont pas à la deuxième étape. Pour tout $i \in \mathcal{M}$, les $\gamma_{ig}$ sont connus, alors supposons que $g_i$ est le groupe masqué du système $i$ (c'est-à-dire que $\gamma_{ig_i} = 1$). Puisqu'on suppose qu'il n'y a qu'une seule cause de panne, le vecteur $\{\delta_{ij} : j \in g_i\}$ suit une multinomiale avec un total de 1 et des probabilités données par $\pi_{j|g_i}(t_i)$, $j \in g_i$. Ainsi, la fonction de log-vraisemblance des données masquées sachant les données observées peut être écrite comme suit :

$$l_{\mathcal{M}}(\boldsymbol{\theta}) = \sum_{i\in\mathcal{M}}\sum_{j\in g_i}\delta_{ij}\ln\Big(\pi_{j|g_i}(t_i)\Big). \qquad (5.51)$$

En utilisant l'identité (3.19) de l'algorithme EM, les équations (5.50) et (5.51) et en dénotant l'espérance conditionnelle sur les données observées par $E(.|OBS)$, on peut

écrire la fonction de log-vraisemblance *observée* comme suit :

$$\begin{aligned} l_{OBS}(\boldsymbol{\theta}) &= Q(\boldsymbol{\theta}|\boldsymbol{\theta}') - H(\boldsymbol{\theta}|\boldsymbol{\theta}') \\ &= E_{\boldsymbol{\theta}'}[l_C(\boldsymbol{\theta})|OBS] - E_{\boldsymbol{\theta}'}[l_{\mathcal{M}}(\boldsymbol{\theta})|OBS] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{J}\int_0^\tau \Bigg\{\Bigg[\ln\Big(\lambda_j(t)\Big)E_{\boldsymbol{\theta}'}\Big[dN_{ij}(t)|OBS\Big] - Y_i(t)\lambda_j(t)dt\Bigg] \\ &\qquad + \Bigg[\Big(1 - \sum_{g\in\mathcal{G}_j^*}\gamma_{ig}\Big)\ln\Big(1 - \sum_{g\in\mathcal{G}_j^*}P_{g|j}(t)\Big) + \sum_{g\in\mathcal{G}_j^*}\gamma_{ig}\ln\Big(P_{g|j}(t)\Big)\Bigg] \\ &\qquad \times E_{\boldsymbol{\theta}'}\Big[dN_{ij}(t)|OBS\Big]\Bigg\} - \sum_{i\in\mathcal{M}}\sum_{j\in g_i}E_{\boldsymbol{\theta}'}\Big[\delta_{ij}|OBS\Big]\ln\Big(\pi_{j|g_i}(t_i)\Big), \end{aligned} \tag{5.52}$$

où $E_{\boldsymbol{\theta}'}\Big[dN_{ij}(t)|OBS\Big] = \begin{cases} E_{\boldsymbol{\theta}'}\Big[\delta_{ij}|OBS\Big] & \text{si } t = t_i, \\ 0 & \text{si } t \neq t_i \end{cases}$

avec

$$E_{\boldsymbol{\theta}'}\Big[\delta_{ij}|OBS\Big] = \begin{cases} 1 & \text{si on sait que la cause de panne de } i \text{ est } j, \\ 0 & \text{si on sait que la cause de panne de } i \text{ n'est pas } j, \\ \pi_{j|g_i}(t_i) & \text{si la cause de panne de } i \text{ est masquée dans } g_i \\ & \text{et qu'il n'y a pas d'étape 2 pour } i. \end{cases} \tag{5.53}$$

La fonction de vraisemblance donnée par l'équation (5.52) ne considère pas le processus de sélection de l'échantillon des systèmes avec une cause de panne masquée qui passent à la deuxième étape. Rubin (1987, p.53) mentionne que les inférences basées sur cette équation sont correctes seulement si les données sont manquantes aléatoirement (« missing at random »). C'est le cas lorsque $P$(le système $i$ est masqué $|OBS$) ne dépend pas des $\delta_{ij}$ manquants.

Si tous les systèmes masqués étaient envoyés à la deuxième étape, alors l'équation (5.50) permet d'écrire que $l_C(\boldsymbol{\theta}) = l_I(\lambda) + l_{II}(p)$, où

$$l_I(\lambda) = \sum_{i=1}^{n}\sum_{j=1}^{J}\int_0^\tau \Big[\ln\Big(\lambda_j(t)\Big)dN_{ij}(t) - Y_i(t)\lambda_j(t)dt\Big], \tag{5.54}$$

$$\begin{aligned} l_{II}(p) &= \sum_{i=1}^{n}\sum_{j=1}^{J}\int_0^\tau \Bigg[\Big(1 - \sum_{g\in\mathcal{G}_j^*}\gamma_{ig}\Big)\ln\Big(1 - \sum_{g\in\mathcal{G}_j^*}P_{g|j}(t)\Big) \\ &\qquad + \sum_{g\in\mathcal{G}_j^*}\gamma_{ig}\ln\Big(P_{g|j}(t)\Big)\Bigg]dN_{ij}(t), \end{aligned} \tag{5.55}$$

où $\lambda$ représente les paramètres de la partie des risques concurrents et $p$ représente les paramètres de la partie masquée. Avec les données complètes, les estimateurs du

maximum de vraisemblance de $\lambda$ et de $p$ sont obtenus indépendamment. Par contre, sous les données masquées, la séparation qui a été faite en (5.54) et (5.55) ne s'applique pas puisque les $E_{\boldsymbol{\theta}}\left\{\delta_{ij}|OBS\right\}$ sont fonction de $\lambda$ et de $p$ pour tout système $i$ qui a une cause de panne masquée.

Ce qu'on vient de voir était un modèle général et on s'intéresse maintenant à un modèle spécifique particulièrement pratique. En effet, on suppose que les fonctions de risque spécifiques sont constantes par intervalles, c'est-à-dire que

$$\lambda_j(t) = \sum_{k=1}^{K} \lambda_{jk} 1_k(t), j = 1, ..., J, \tag{5.56}$$

où $0 = a_0 < a_1 < a_2 < ... < a_K = \tau$, et $1_k(t)$ est l'indicatrice qui vaut 1 si $t \in (a_{k-1}, a_k]$, 0 sinon. On émet aussi l'hypothèse que les points de coupure $a_0, a_1, ..., a_K$ sont les mêmes pour toutes les causes de panne. Cette supposition n'est pas nécessaire, mais elle simplifie la notation et l'exposition. Les avantages d'un modèle avec des risques concurrents constants par intervalles sont les suivants :

- Les EMV sont exprimables de façon explicite sous les données complètes.
- L'algorithme EM converge sous de faibles hypothèses sur les données observées.
- La forme des fonctions de risque est flexible (croissante, décroissante, non-monotone).

Soit $e_k = \sum_{i=1}^{n} \int_0^{t_i} 1_k(u)du$, le temps total vécu par tous les systèmes dans l'intervalle $(a_{k-1}, a_k]$. La fonction de log-vraisemblance complète (5.50) est alors maximisée par

$$\hat{\lambda}_{jk} = \frac{\sum_{i=1}^{n} \delta_{ij} 1_k(t_i)}{e_k} \tag{5.57}$$

et pour le modèle avec les probabilités de masque constantes dans le temps ($P_{g|j}(t) \equiv P_{g|j}$),

$$\hat{P}_{g|j} = \frac{\sum_{i=1}^{n} \delta_{ij}\gamma_{ig}}{\sum_{i=1}^{n} \delta_{ij}}. \tag{5.58}$$

Du théorème 2.1, on peut trouver l'EMV de $\pi_{j|g}(t)$ en remplaçant les paramètres de l'équation (5.48) par les EMV de (5.57) et de (5.58).

Grâce à la simplicité de la fonction de log-vraisemblance complète (5.50) et à sa linéarité en les données manquantes ($\delta_{ij}$), l'algorithme EM est un bon outil pour ce problème. La fonction de vraisemblance observée (5.52) est maximisée en utilisant la

méthode expliquée au chapitre 3. Cette méthode itérative est utilisée de la façon suivante pour le modèle considéré :

On choisit d'abord des points de départ $\hat{\lambda}_{jk}^{(0)}$ et $\hat{P}_{g|j}^{(0)}$ et on exécute les étapes E et M itérativement jusqu'à ce que la règle d'arrêt soit satisfaite. Pour l'itération $(b+1)$, on a :

**Étape E**. Calculer $E_{\hat{\boldsymbol{\theta}}^{(b)}}(\delta_{ij}|OBS)$ en utilisant l'équation (5.53) et l'EMV de (5.48).

**Étape M**. Maximiser $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(b)}) = E_{\boldsymbol{\theta}^{(b)}}(l_C(\boldsymbol{\theta})|OBS)$ par rapport à $\boldsymbol{\theta}$ en utilisant les équations (5.57) et (5.58). Poser

$$\hat{\lambda}_{jk}^{(b+1)} = \frac{\sum_{i=1}^{n} E_{\hat{\boldsymbol{\theta}}^{(b)}}[\delta_{ij}|OBS]1_k(t_i)}{e_k} \tag{5.59}$$

et

$$\hat{P}_{g|j}^{(b+1)} = \frac{\sum_{i=1}^{n} E_{\hat{\boldsymbol{\theta}}^{(b)}}[\delta_{ij}|OBS]\ \gamma_{ig}}{\sum_{i=1}^{n} E_{\hat{\boldsymbol{\theta}}^{(b)}}[\delta_{ij}|OBS]}. \tag{5.60}$$

**Règle d'arrêt**. Lorsque $||\hat{\boldsymbol{\theta}}^{(b+1)} - \hat{\boldsymbol{\theta}}^{(b)}|| \leq \varepsilon$, pour une petite valeur de tolérance $\varepsilon$ présélectionnée.

On utilise les résultats de Wu (1983) pour montrer que l'algorithme EM expliqué ci-haut converge vers un point stationnaire dans le cas de risques concurrents constants par intervalles et de probabilités masquées.

Soit $\Omega \subset \mathcal{R}^d$, l'espace des paramètres du modèle. La dimension $d$ dépend du nombre d'intervalles sur lesquels les risques concurrents et les probabililités de masque sont constants, du nombre de groupes masqués et du nombre de causes de panne. Supposons que les points de coupure des intervalles des fonctions de risques concurrents et des probabilités de masque sont les mêmes. On suppose qu'on choisit ces points de coupure tels que pour chaque intervalle $k$ et chaque cause de panne $j$, il existe un $i$ tel que $j \in g_i$ et $1_k(t_i) = 1$.

Avec ce choix de $\{a_k,\ 0 \leq k \leq K\}$, pour tout $\boldsymbol{\theta}_0$ tel que $l_{OBS}(\boldsymbol{\theta}_0) > -\infty$, l'ensemble $\Omega_{\boldsymbol{\theta}_0} = \{\boldsymbol{\theta} \in \Omega : l_{OBS}(\boldsymbol{\theta}) \geq l_{OBS}(\boldsymbol{\theta}_0)\}$ est compact dans $\mathcal{R}^d$. Ceci implique que les conditions (5) à (7) de Wu (1983) sont respectées. Aussi, $Q(\boldsymbol{\theta}|\boldsymbol{\theta}')$ est continue en $\boldsymbol{\theta}$ et en $\boldsymbol{\theta}'$, ce qui implique (Wu, théorème 2) que les points limite de n'importe quelle suite $\left\{\boldsymbol{\theta}^{(b)},\ b = 0, 1, 2, ...\right\}$ de l'algorithme EM sont des points stationnaires de $l_{OBS}$, et

$l_{OBS}(\boldsymbol{\theta}^{(b)})$ converge de façon monotone vers $l_{OBS}(\boldsymbol{\theta}^*)$ pour un point stationnaire $\boldsymbol{\theta}^*$.

S'il n'y a pas de données de deuxième étape, l'algorithme proposé est encore applicable, sauf lorsque les fonctions de risque sont proportionnelles. Dans ce cas, la convergence de l'algorithme EM est extrêmement lente, ce qui requiert des milliers d'itérations. Intuitivement, on comprend que plus il y a de données manquantes, plus l'algorithme prend du temps à converger, comme on en a parlé à la section 3.8.2. C'est ce qui se passe lorsqu'aucun système n'est envoyé à la deuxième étape.

Dans l'article de Craiu et Duchesne (2004), on utilise l'algorithme SEM expliqué à la section 3.8.2 pour calculer la matrice de variance-covariance des paramètres estimés. De plus, les techniques de cet article peuvent être utilisées seules ou comme un complément des méthodes de Flehinger, Reiser et Yashchin (1998, 2002) pour, entre autres, guider vers la sélection d'un modèle paramétrique ou pour tester les hypothèses de symétrie ou de probabilités de masque constantes dans le temps. La méthode de Craiu et Duchesne est assez flexible pour fonctionner même lorsque les données sont groupées, lorsque les probabilités de masque dépendent du temps et même souvent lorsqu'il n'y a pas de données de deuxième étape. Cependant, Craiu et Duchesne ne considèrent pas la performance de leur méthode en l'absence de données de deuxième étape en détails. C'est ce que nous ferons au chapitre 6.

# Chapitre 6

# Absence de données de deuxième étape

En pratique, il peut fréquemment arriver qu'aucun des systèmes ayant leur cause de panne masquée à la première étape ne puisse avoir une investigation plus profonde à une deuxième étape. La deuxième étape peut être absente en pratique à cause du coût que cela entraîne ou à cause d'un manque de temps. Dans ce cas, seules les données de première étape sont disponibles.

On discutera de l'approche théorique qui a été développée jusqu'à maintenant sur ce sujet par les chercheurs à la section 6.1 ainsi que d'une approche par simulations à la section 6.2. Finalement, aux sections 6.3 et 6.4, on comparera les estimateurs obtenus dans des études de cas de Dinse (1986) et de Flehinger, Reiser et Yashchin (2002) aux estimateurs calculés avec l'algorithme EM et le modèle de Craiu et Duchesne (2004) lorsqu'il n'y a pas de données de deuxième étape.

## 6.1 Étude théorique

Dans cette section, on explique ce qui a déjà été étudié à propos de l'absence de données de deuxième étape. Tout d'abord, on présente à la sous-section 6.1.1 un bref

aperçu de ce qui a été dit sur l'évaluation des EMV lorsqu'on a un modèle de risques concurrents en l'absence de données de deuxième étape. Ensuite, on discute de l'identifiabilité des EMV pour un modèle de risques concurrents constants par paliers en l'absence de deuxième étape à la sous-section 6.1.2.

### 6.1.1 Études antérieures

Tel que vu à la section 5.2, Flehinger et al. (1998) font l'hypothèse des risques proportionnels. Une étude de deuxième étape, même de petite envergure, donne alors des estimateurs beaucoup plus précis que des données de première étape seulement. L'équation (5.12) contient la vraisemblance pour ces données. Flehinger et al. (1998) montrent que des données de deuxième étape sont nécessaires pour pouvoir estimer les probabilités de panne, $F_j$, associées à chacun des risques. Ainsi, les seules données de première étape ne permettent pas d'estimer les fonctions de survie marginales, $S_j(t)$, et les probabilités de diagnostic, $\pi_{j|g}$. Pour régler ce problème, s'il est impossible d'envoyer des systèmes à une deuxième étape, on peut, entre autres, émettre l'hypothèse de symétrie qui s'énonce comme suit : la probabilité qu'un système soit masqué à la première étape ne dépend pas de la vraie cause de panne, c'est-à-dire que $P_{g|j} = P_g$, pour tout $j \in g$, où $j$ représente la cause de panne et $g$ est le groupe masqué.

Dans l'article de Flehinger, Reiser et Yashchin de 2002 résumé à la section 5.3, on utilise un modèle paramétrique au lieu de faire l'hypothèse des risques proportionnels. S'il y a seulement des données de première étape, l'algorithme présenté dans cet article qui calcule les estimateurs peut être utilisé. Ces auteurs comparent un modèle dont seulement des données de première étape sont disponibles avec un autre dont un certain pourcentage des systèmes est envoyé à la seconde étape. Ils ont obtenu les résultats suivants : les estimateurs des probabilités de survie sont assez près de ceux obtenus lorsqu'il y a des données de deuxième étape. Par contre, les probabilités de masque et de diagnostic sont assez différentes.

À la section 5.4, Craiu et Duchesne (2004) ont utilisé un modèle de risques concurrents constants par paliers. Les EMV sont calculés avec l'algorithme EM et même en l'absence des données de deuxième étape, cet algorithme converge. Sous la proportionnalité des risques concurrents et en l'absence de deuxième étape, on sait de Flehinger, Reiser et Yashchin (1998) que les paramètres du modèle construit à partir des données observées ne sont pas identifiables. Par contre, l'algorithme EM converge quand même puisque les paramètres sont identifiables lorsque les données sont complètes, mais la

convergence se fait très lentement. Dans Dempster et coll. (1977), on affirme que lorsque les paramètres sont identifiables seulement dans le modèle des données complètes, il y a une chaîne de maxima locaux dans la surface de la vraisemblance observée et l'algorithme EM converge vers un des points de la chaîne, tout dépendant du point de départ. Dans ce genre de situation, il est suggéré d'utiliser une approche alternative pour passer par dessus la surparamétrisation, telle que l'hypothèse de symétrie.

### 6.1.2 Identifiabilité des EMV sous un modèle de risques concurrents constants par intervalles

Selon le modèle des risques concurrents constants par paliers discuté à la section 5.4, on fait l'étude de l'identifiabilité des EMV trouvés à partir de la fonction de vraisemblance construite sous ce modèle. On écrit cette fonction sous une forme différente de celle de l'équation (5.49). En effet, on emprunte plutôt le même genre de notation que celle des sections 5.2 et 5.3 pour pouvoir exprimer les EMV selon le nombre de pannes dues à la cause $j$ dans l'intervalle $k$, $n_{jk}$, et selon le nombre de systèmes masqués dans le groupe $g$ dans l'intervalle $k$, $n_{gk}$, $j = 1, ..., J$, $k = 1, ..., K$.

Craiu et Reiser (2004) discutent de ce modèle dans une section de leur article qui porte sur l'identifiabilité des paramètres lorsqu'il n'y a pas de systèmes envoyés à la deuxième étape. On montre que lorsque les mêmes points de coupure sont utilisés pour les intervalles des différentes causes de panne, on peut arriver à une solution non identifiable. Ceci est vrai même si les risques concurrents ne sont pas proportionnels.

Pour construire ce modèle, on reprend une partie de la notation des sections 5.2 et 5.3, c'est-à-dire que $n$ représente le nombre total de systèmes, tandis que $n_j$, $n_g$, $n_{g,j}$ et $n_c$ dénotent respectivement le nombre de systèmes tombés en panne de la cause $j$, masqués dans le groupe $g$, masqués dans le groupe $g$ à la première étape, mais diagnostiqués comme étant la cause $j$ à la deuxième étape et finalement, le nombre de systèmes censurés. Pour chacun de ces systèmes, on a soit un temps de panne dû à la cause $j$, $t_i^{(j)}$, un temps de panne masqué dans le groupe $g$, $t_i^{(g)}$, un temps de panne masqué dans le groupe $g$ à la première étape, mais dont la vraie cause $j$ est trouvée à la deuxième étape, $t_i^{(g,j)}$, ou un temps de censure, $t_i^{(c)}$. Finalement, on rappelle que $S(t)$ est la fonction de survie totale et que $\lambda_j(t)$ et $f_j(t)$ sont les fonctions de risque et de densité spécifiques à la cause $j$.

Pour faire un lien entre cette notation et celle de la section 5.4, on écrit $n_j$, $n_g$, $n_{g,j}$ et $n_c$ en fonction des quantités décrites à la section 5.4. Dans cette dernière, on a vu que l'observation pour le système $i$ est

$$(t_i, \gamma_{ig_1}, \ldots, \gamma_{ig_{G+J}}, \delta_{i1}, \ldots, \delta_{iJ}),$$

où $t_i$ est le temps de panne du système $i$, $\gamma_{ig}$ est l'indicatrice qui vaut 1 si le système $i$ est masqué dans le groupe $g$ à la première étape et $\delta_{ij}$ est l'indicatrice qui vaut 1 si la cause de panne du système $i$ est $j$. Ainsi, $n_j$, $n_g$, $n_{g,j}$ et $n_c$ peuvent s'écrire comme suit :

$$\begin{aligned}
n_j &= \textstyle\sum_{i=1}^{n} \gamma_{ij}, \text{ pour } j = 1, ..., J.\\
n_g &= \textstyle\sum_{i=1}^{n} \gamma_{ig_{J+g}}, \text{ pour } g = 1, ..., G.\\
n_{g,j} &= \textstyle\sum_{i=1}^{n} \gamma_{ig_{J+g}}\delta_{ij}, \text{ pour } g = 1, ..., G \text{ et } j = 1, ..., J.\\
n_c &= n - (n_j + n_g + n_{g,j}).
\end{aligned}$$

De plus, on a vu à l'équation (5.56) que $1_k(t)$ est l'indicatrice qui vaut 1 si $t \in (a_{k-1}, a_k]$. On a alors que

$$\begin{aligned}
n_{jk} &= \textstyle\sum_{i=1}^{n} \gamma_{ij} 1_k(t_i^{(j)}), \text{ pour } j = 1, ..., J.\\
n_{gk} &= \textstyle\sum_{i=1}^{n} \gamma_{ig_{J+g}} 1_k(t_i^{(g)}), \text{ pour } j = 1, ..., J \text{ et } g = 1, ..., G,
\end{aligned}$$

où $n_{jk}$ et $n_{gk}$ sont respectivement le nombre de pannes dues à la cause $j$ dans l'intervalle $k$ et le nombre de systèmes masqués dans le groupe $g$ dans l'intervalle $k$, $j = 1, ..., J$, $k = 1, ..., K$.

En se fiant à la fonction de vraisemblance de l'équation (5.12), on peut écrire la fonction de vraisemblance de la première étape de la façon suivante :

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{n_c} S(t_i^{(c)}) \prod_{j=1}^{J} \prod_{i=1}^{n_j} P_j f_j(t_i^{(j)}) \prod_{g} \prod_{i=1}^{n_g} \sum_{r \subset g} P_{g|r} f_r(t_i^{(g)}), \tag{6.1}$$

où $P_j = 1 - \sum_{g \in \mathcal{G}^*} P_{g|j}$.

En utilisant le fait que $\lambda_j(t) = \sum_{k=1}^{K} \lambda_{jk} 1_k(t)$, où $\lambda_{jk}$ est la fonction de risque spécifique à la cause $j$ dans l'intervalle $k$, et que $f_j(t) = \lambda_j(t)S(t)$, on définit la fonction de densité du temps de panne comme étant

$$f(t) = \sum_{j=1}^{J} f_j(t) = \sum_{j=1}^{J} \lambda_j(t)S(t) = \sum_{j=1}^{J} \sum_{k=1}^{K} \lambda_{jk} 1_k(t) S(t). \tag{6.2}$$

On définit aussi

$$\phi_{jk} = \frac{\lambda_{jk}}{\lambda_{.k}}, \; j = 1, ..., J, \; k = 1, ..., K, \tag{6.3}$$

comme étant la proportion du risque total de l'intervalle $k$ due à la cause $j$, où $\lambda_{\cdot k} = \sum_{j=1}^{J} \lambda_{jk}$.

En effectuant les opérations suivantes, on obtient une nouvelle façon d'écrire $f_j(t)$ :

$$\begin{aligned} f_j(t) = \lambda_j(t)S(t) &= \sum_{k=1}^{K} \lambda_{jk} 1_k(t) S(t) \\ &= \sum_{k=1}^{K} \phi_{jk} \lambda_{\cdot k} 1_k(t) S(t) \qquad \text{(de (6.3))} \\ &= \sum_{k=1}^{K} \phi_{jk} \lambda_{\cdot k} 1_k(t) \frac{f(t)}{\sum_{j=1}^{J} \sum_{k=1}^{K} \lambda_{jk} 1_k(t)} \qquad \text{(de (6.2))} \\ &= \frac{\sum_{k=1}^{K} \phi_{jk} \lambda_{\cdot k} 1_k(t)}{\sum_{k=1}^{K} \lambda_{\cdot k} 1_k(t)} f(t) \\ &= \phi_{jk} f(t), \quad \forall \ t \in (a_{k-1}, a_k]. \end{aligned} \tag{6.4}$$

De plus, si on définit $\phi_j(t) = \sum_{k=1}^{K} \phi_{jk} 1_k(t)$, alors on peut réécrire l'équation (6.4) ainsi :

$$f_j(t) = \phi_j(t) f(t). \tag{6.5}$$

La fonction de vraisemblance (6.1) peut maintenant s'écrire

$$\begin{aligned} L(\boldsymbol{\theta}) &= L_1 \times L_2 \\ &= \left\{ \prod_{i=1}^{n_c} S(t_i^{(c)}) \prod_{i=1}^{n-n_c} f(t_i) \right\} \\ &\quad \times \left\{ \prod_{j=1}^{J} \prod_{i=1}^{n_j} P_j \phi_j(t_i^{(j)}) \prod_{g} \prod_{i=1}^{n_g} \sum_{r \subset g} P_{g|r} \phi_r(t_i^{(g)}) \right\}. \end{aligned} \tag{6.6}$$

Dans l'équation (6.6), $L_1$ est la fonction de vraisemblance pour $n$ systèmes qui, dans le cas des risques concurrents constants par intervalle, détermine les EMV pour la fonction de risque total, $\lambda(t)$. Pour sa part, la fonction $L_2$ tient seulement compte du nombre de décès dû à chacune des causes dans chacun des intervalles, c'est-à-dire les $n_{jk}$. En effet, elle ne tient pas compte des temps de décès. Cette fonction peut se réécrire comme suit :

$$\begin{aligned} L_2 &= \prod_{k=1}^{K} \left\{ \prod_{j=1}^{J} \left( \prod_{i=1}^{n_{jk}} P_j \phi_{jk} \right) \prod_{g} \prod_{i=1}^{n_{gk}} \left( \sum_{r \subset g} P_{g|r} \phi_{rk} \right) \right\} \\ &= \prod_{k=1}^{K} \left\{ \prod_{j=1}^{J} ( P_j \phi_{jk})^{n_{jk}} \prod_{g} \left( \sum_{r \subset g} P_{g|r} \phi_{rk} \right)^{n_{gk}} \right\}. \end{aligned} \tag{6.7}$$

Dans le cas particulier où $J = 2$ et $K = 2$, $L_2$ s'écrit ainsi :

$$\begin{aligned} L_2 &= (P_1\phi_{11})^{n_{11}} (P_2\phi_{21})^{n_{21}} (P_{g|1}\phi_{11} + P_{g|2}\phi_{21})^{n_{g1}} \\ &\quad \times (P_1\phi_{12})^{n_{12}} (P_2\phi_{22})^{n_{22}} (P_{g|1}\phi_{12} + P_{g|2}\phi_{22})^{n_{g2}}, \end{aligned} \tag{6.8}$$

où $P_{g|1}\phi_{11} + P_{g|2}\phi_{21} = 1 - P_1\phi_{11} - P_2\phi_{21}$ et $P_{g|1}\phi_{12} + P_{g|2}\phi_{22} = 1 - P_1\phi_{12} - P_2\phi_{22}$.

La fonction de log-vraisemblance pour cet exemple s'exprime donc de la façon suivante :

$$\begin{aligned} l_2 = \ln(L_2) &= n_{11}\ln(P_1\phi_{11}) + n_{21}\ln(P_2\phi_{21}) \\ &\quad + n_{12}\ln(P_1\phi_{12}) + n_{22}\ln(P_2\phi_{22}) \\ &\quad + n_{g1}\ln(1 - P_1\phi_{11} - P_2\phi_{21}) \\ &\quad + n_{g2}\ln(1 - P_1\phi_{12} - P_2\phi_{22}). \end{aligned} \tag{6.9}$$

Pour trouver les équations du maximum de vraisemblance pour le cas simple où $J = 2$ et $K = 2$, on calcule d'abord l'estimateur du maximum de vraisemblance de $P_1\phi_{11}$ :

$$\frac{\partial l_2}{\partial(P_1\phi_{11})} = \frac{n_{11}}{(P_1\phi_{11})} - \frac{n_{g1}}{(1 - P_1\phi_{11} - P_2\phi_{21})} = 0$$

$$\Leftrightarrow$$

$$\widehat{P_1\phi_{11}} = \frac{n_{11}\Big(1 - \Big[P_2\widehat{(1-\phi_{11})}\Big]\Big)}{n_{11} + n_{g1}}. \tag{6.10}$$

On calcule ensuite l'estimateur du maximum de vraisemblance de $P_2(1 - \phi_{11})$ :

$$\frac{\partial l_2}{\partial(P_2(1-\phi_{11}))} = \frac{n_{21}}{P_2(1-\phi_{11})} - \frac{n_{g1}}{(1 - P_1\phi_{11} - P_2(1-\phi_{11}))} = 0$$

$$\Leftrightarrow$$

$$P_2\widehat{(1-\phi_{11})} = \frac{n_{21}\Big(1 - [\widehat{P_1\phi_{11}}]\Big)}{n_{21} + n_{g1}}. \tag{6.11}$$

De la propriété d'invariance des EMV du théorème 2.1 et des équations (6.10) et (6.11), on trouve les deux équations suivantes :

$$\widehat{P_1\phi_{11}} = \frac{n_{11}}{n_{11} + n_{21} + n_{g1}}, \tag{6.12}$$

$$P_2\widehat{(1-\phi_{11})} = \frac{n_{21}}{n_{11} + n_{21} + n_{g1}}. \tag{6.13}$$

Par symétrie, on a pour $K = 2$ que

$$\widehat{P_1\phi_{12}} = \frac{n_{12}}{n_{12} + n_{22} + n_{g2}}, \tag{6.14}$$

$$P_2\widehat{(1-\phi_{12})} = \frac{n_{22}}{n_{12} + n_{22} + n_{g2}}. \tag{6.15}$$

Grâce aux équations (6.12) à (6.15), on peut écrire $a_1$ et $a_2$ comme suit :

$$a_1 = \frac{\hat{\phi}_{11}}{\hat{\phi}_{12}} = \frac{n_{11}(n_{12} + n_{22} + n_{g2})}{n_{12}(n_{11} + n_{21} + n_{g1})}, \tag{6.16}$$

$$a_2 = \frac{(1 - \hat{\phi}_{11})}{(1 - \hat{\phi}_{12})} = \frac{n_{21}(n_{12} + n_{22} + n_{g2})}{n_{22}(n_{11} + n_{21} + n_{g1})}. \tag{6.17}$$

À l'aide des équations (6.16) et (6.17), on exprime $\hat{\phi}_{11}$ et $\hat{\phi}_{12}$ en fonction de $a_1$ et de $a_2$ :

$$a_2 = \frac{(1 - \hat{\phi}_{11})}{(1 - \hat{\phi}_{12})} = \frac{(1 - \hat{\phi}_{11})}{(1 - \frac{\hat{\phi}_{11}}{a_1})}$$

$\Leftrightarrow$

$$\hat{\phi}_{11} = \frac{a_1(1 - a_2)}{a_1 - a_2}, \tag{6.18}$$

$$\hat{\phi}_{12} = \frac{\hat{\phi}_{11}}{a_1} = \frac{1 - a_2}{a_1 - a_2}. \tag{6.19}$$

Finalement, de (6.12) et (6.13), on trouve les estimateurs $\hat{P}_1$ et $\hat{P}_2$ :

$$\hat{P}_1 = \frac{1}{\hat{\phi}_{11}} \left[ \frac{n_{11}}{n_{11} + n_{21} + n_{g1}} \right], \tag{6.20}$$

$$\hat{P}_2 = \frac{1}{(1 - \hat{\phi}_{11})} \left[ \frac{n_{21}}{n_{11} + n_{21} + n_{g1}} \right]. \tag{6.21}$$

Ce sont les quatre paramètres distincts à estimer. Les quatre autres paramètres sont des combinaisons de ceux-ci. En effet,

$$\begin{aligned} \hat{\phi}_{21} &= 1 - \hat{\phi}_{11}, \\ \hat{\phi}_{22} &= 1 - \hat{\phi}_{12}, \\ \hat{P}_{g|1} &= 1 - \hat{P}_1, \\ \hat{P}_{g|2} &= 1 - \hat{P}_2. \end{aligned}$$

On remarque que tous ces estimateurs sont fonction seulement des $n_{jk}$ et non pas des temps de panne. Craiu et Reiser (2004) ont trouvé un exemple numérique pour lequel les estimateurs du maximum de vraisemblance de plusieurs des quatre paramètres tombent en dehors de l'espace paramétrique, c'est-à-dire à l'extérieur de $(0, 1)$. En effet, si on prend les valeurs de $n_{jk}$ suivantes : $n_{11} = n_{12} = n_{22} = 30$, $n_{21} = 20$, $n_{g1} = 10$ et $n_{g2} = 60$, alors on a des valeurs de $a_1 = 2$ et de $a_2 = 4/3$. Les quatre estimateurs du maximum de vraisemblance donnent des valeurs de $\hat{\phi}_{11} = -1$, $\hat{\phi}_{12} = -1/2$, $\hat{P}_1 = -1/2$ et $\hat{P}_2 = 1/6$.

Par contre, si on prend les $n_{jk}$ suivants : $n_{11} = n_{22} = 15$, $n_{12} = n_{21} = 30$ et $n_{g1} = n_{g2} = 15$, alors on a des valeurs de $a_1 = 1/2$ et de $a_2 = 2$. Les quatre estimateurs du maximum de vraisemblance donnent des valeurs de $\hat{\phi}_{11} = 1/3$, $\hat{\phi}_{12} = 2/3$, $\hat{P}_1 = 3/4$ et $\hat{P}_2 = 3/4$, toutes entre 0 et 1.

On remarque que selon les différentes valeurs de $a_1$ et de $a_2$, les valeurs de $\hat{\phi}_{11}$ et de $\hat{\phi}_{12}$ des équations (6.18) et (6.19) peuvent être soit entre 0 et 1 ou non. En effet, $\hat{\phi}_{11}$ et $\hat{\phi}_{12}$ tombent dans l'espace paramétrique si on a l'une des deux conditions suivantes :

1. $0 \leq a_1 \leq 1 \leq a_2$, sauf au point $a_1 = a_2 = 1$,
2. $0 \leq a_2 \leq 1 \leq a_1$, sauf au point $a_1 = a_2 = 1$.

Par les équations (6.20) et (6.21), on a aussi des valeurs entre 0 et 1 pour $\hat{P}_1$ et $\hat{P}_2$. Il s'agit maintenant de trouver des valeurs de $n_{jk}$ qui satisfont l'une de ces deux conditions.

Lorsque $J = 2$ et $K = 3$, le nombre d'équations est plus grand que le nombre d'inconnues et les paramètres ne sont pas complètement identifiables ou pas identifiables du tout. Craiu et Reiser (2004) ont généré des données qui ont permis de résoudre l'équation (6.7) exactement et d'autres données qui ont conduit à la non-identifiabilité. Avec trois intervalles, Craiu et Reiser ont noté que l'algorithme EM converge à des points très variables qui dépendent du point de départ utilisé dans l'algorithme. Ce comportement irrégulier semble être un signe d'absence d'identifiabilité sous les données observées et d'identifiabilité sous les données complètes. Pour contourner la non-identifiabilité, on peut choisir différents points de coupure pour les deux fonctions de risque spécifiques aux causes de panne. En effet, dans ce cas, les temps de panne contribuent à l'estimation des probabilités de masque, $P_{g|j},\ j = 1, 2$.

Sans les données de deuxième étape, le point stationnaire $\boldsymbol{\theta}^*$ discuté à la section 3.4 peut être un maximum local même s'il existe un maximum global identifiable. Ainsi, il est recommandé d'utiliser plusieurs points de départ dans l'algorithme EM.

## 6.2 Étude par simulations

Sous le modèle des risques concurrents constants par intervalles, on simule un jeu de données et on calcule les EMV avec l'algorithme EM lorsqu'il y a des causes de panne masquées et qu'aucun système n'est envoyé à la deuxième étape. On cherche à savoir si la moyenne des estimateurs, calculée à partir de plusieurs échantillons, peut être évaluée avec précision lorsqu'il n'y a pas de données de deuxième étape. On veut aussi savoir comment se comportent la moyenne et la variance de ces estimateurs lorsqu'on fait varier la taille des échantillons ou lorsque les risques sont proportionnels. Finalement, on regarde si les conditions d'identification des EMV selon la théorie de la section 6.1 peuvent être vérifiées à l'aide de ces simulations.

### 6.2.1 Création du jeu de données

On génère d'abord des données à l'aide du logiciel R. Celles-ci simulent des systèmes tombant en panne de 2 causes possibles, avec la possibilité de ne pas connaître la cause responsable de la panne. Ainsi, un temps et une cause de panne sont associés à chaque système de l'échantillon et un certain pourcentage de causes de panne sont masquées. Ce pourcentage est défini préalablement grâce aux probabilités de masque, $P_{g|1}$ et $P_{g|2}$, où $g = \{1, 2\}$. Les simulations sont effectuées avec plusieurs $P_{g|j}$, $j = 1, 2$ différents. Les résultats pour $P_{g|1} = 0.9$ et $P_{g|2} = 0.5$ sont présentés, les résultats pour les autres combinaisons de valeurs étant similaires.

Pour un système donné, on utilise l'algorithme de l'Annexe A pour générer un temps de panne $t$ dû à une cause $j$ avec des fonctions de risque spécifiques constantes par intervalles. Toutes les simulations sont faites avec des risques concurrents indépendants et avec les points de coupure $a_0 = 0$, $a_1 = 5$ et $a_2 = 10$. Ces derniers sont les mêmes pour les deux causes de panne. Les $K = 3$ intervalles sont donc les suivants : $[0, 5]$, $(5, 10]$ et $(10, \infty)$. Quelques essais ont été faits avec des risques non-proportionnels et d'autres avec des risques proportionnels. Les résultats pour les $\lambda_{jk}$ , $j = 1, 2$, $k = 1, 2, 3$, théoriques suivants sont présentés :

1. Pour les risques non-proportionnels :

$$\begin{array}{lll} \lambda_{11} = 0,003, & \lambda_{12} = 0,02, & \lambda_{13} = 0,012, \\ \lambda_{21} = 0,0045, & \lambda_{22} = 0,01, & \lambda_{23} = 0,03. \end{array}$$

2. Pour les risques proportionnels :

$$\begin{array}{lll} \lambda_{11} = 0,003, & \lambda_{12} = 0,02, & \lambda_{13} = 0,012, \\ \lambda_{21} = 0,006, & \lambda_{22} = 0,04, & \lambda_{23} = 0,024. \end{array}$$

On simule 1000 échantillons de 3 tailles différentes : 40, 100 et 1000. Afin de pouvoir comparer les valeurs des estimateurs trouvés lorsqu'il n'y a pas de systèmes masqués envoyés à la deuxième étape, on fait des essais avec quelques systèmes masqués envoyés à cette étape. Parmi les systèmes masqués à la première étape, on en résout 0%, 5%, 20%, 50%, et 100% à la deuxième étape pour les tailles de 40 et de 100. Pour la taille de 1000, seulement des essais avec 0% en deuxième étape sont étudiés.

### 6.2.2 Estimation des paramètres

Lorsque les jeux de données sont complétés, les paramètres peuvent être estimés avec l'algorithme EM. La programmation pour calculer ces derniers est développée en langage C. L'algorithme EM est généré à l'aide des étapes E et M énoncées à la section 5.4. Pour chaque échantillon, les estimateurs suivants sont calculés :

1. $\hat{P}_1$, $\hat{P}_{g|1}$, $\hat{P}_2$ et $\hat{P}_{g|2}$,
2. $\hat{\lambda}_{11}$, $\hat{\lambda}_{12}$, $\hat{\lambda}_{13}$, $\hat{\lambda}_{21}$, $\hat{\lambda}_{22}$ et $\hat{\lambda}_{23}$,

où $P_j = 1 - P_{g|j}$, $j = 1, 2$, et $g = \{1, 2\}$.

En théorie, si la taille des échantillons tendait vers l'infini, les valeurs de ces estimateurs devraient être égales aux valeurs des paramètres prédéfinis pour construire les jeux de données.

Puisqu'on obtient une valeur pour chaque estimateur de chacun des échantillons, une moyenne et un écart-type sont calculés pour chaque estimateur à partir des 1000 échantillons. Ce sont ces valeurs qui sont présentées dans les tableaux et les graphiques de la section 6.2.3.

TAB. 6.1 – Estimateurs des risques non-proportionnels pour 1000 échantillons de taille 100, où les valeurs théoriques sont données par : $P_{g|1} = 0.9$, $P_{g|2} = 0.5$, $\lambda_{11} = 0,003$, $\lambda_{12} = 0,02$, $\lambda_{13} = 0,012$, $\lambda_{21} = 0,0045$, $\lambda_{22} = 0,01$, et $\lambda_{23} = 0,03$.

| Estimateur | % de systèmes masqués envoyés à la 2ième étape | | | | | |
|---|---|---|---|---|---|---|
| | 0% | | 5% | | 20% | |
| | $j=1$ | $j=2$ | $j=1$ | $j=2$ | $j=1$ | $j=2$ |
| $\hat{P}_j$ | 0,159(0,216) | 0,676(0,278) | 0,136(0,162) | 0,532(0,160) | 0,105(0,070) | 0,503(0,086) |
| $\hat{P}_{g\|j}$ | 0,841(0,216) | 0,324(0,278) | 0,864(0,162) | 0,468(0,160) | 0,895(0,070) | 0,497(0,086) |
| $\hat{\lambda}_{j1}$ | 0,003(0,003) | 0,004(0,004) | 0,003(0,003) | 0,004(0,004) | 0,003(0,003) | 0,004(0,004) |
| $\hat{\lambda}_{j2}$ | 0,021(0,010) | 0,009(0,008) | 0,020(0,009) | 0,011(0,008) | 0,020(0,008) | 0,010(0,006) |
| $\hat{\lambda}_{j3}$ | 0,016(0,011) | 0,027(0,011) | 0,012(0,007) | 0,030(0,008) | 0,012(0,004) | 0,030(0,005) |

| Estimateur | % de systèmes masqués envoyés à la 2ième étape | | | |
|---|---|---|---|---|
| | 50% | | 100% | |
| | $j=1$ | $j=2$ | $j=1$ | $j=2$ |
| $\hat{P}_j$ | 0,099(0,053) | 0,499(0,070) | 0,098(0,051) | 0,498(0,064) |
| $\hat{P}_{g\|j}$ | 0,901(0,053) | 0,501(0,070) | 0,902(0,051) | 0,502(0,064) |
| $\hat{\lambda}_{j1}$ | 0,003(0,003) | 0,005(0,003) | 0,003(0,002) | 0,004(0,003) |
| $\hat{\lambda}_{j2}$ | 0,020(0,007) | 0,010(0,005) | 0,020(0,007) | 0,010(0,005) |
| $\hat{\lambda}_{j3}$ | 0,012(0,003) | 0,030(0,004) | 0,012(0,002) | 0,030(0,004) |

TAB. 6.2 – Estimateurs des risques proportionnels pour 1000 échantillons de taille 100, où les valeurs théoriques sont données par : $P_{g|1} = 0.9$, $P_{g|2} = 0.5$, $\lambda_{11} = 0,003$, $\lambda_{12} = 0,02$, $\lambda_{13} = 0,012$, $\lambda_{21} = 0,006$, $\lambda_{22} = 0,04$, et $\lambda_{23} = 0,024$.

| Estimateur | % de systèmes masqués envoyés à la 2ième étape | | | | | |
|---|---|---|---|---|---|---|
| | 0% | | 5% | | 20% | |
| | $j=1$ | $j=2$ | $j=1$ | $j=2$ | $j=1$ | $j=2$ |
| $\hat{P}_j$ | 0,211(0,297) | 0,679(0,284) | 0,180(0,247) | 0,539(0,175) | 0,109(0,067) | 0,503(0,088) |
| $\hat{P}_{g\|j}$ | 0,789(0,297) | 0,321(0,284) | 0,820(0,247) | 0,461(0,175) | 0,891(0,067) | 0,497(0,088) |
| $\hat{\lambda}_{j1}$ | 0,004(0,004) | 0,005(0,005) | 0,003(0,004) | 0,006(0,005) | 0,003(0,003) | 0,006(0,004) |
| $\hat{\lambda}_{j2}$ | 0,025(0,018) | 0,035(0,019) | 0,020(0,014) | 0,040(0,016) | 0,019(0,010) | 0,041(0,013) |
| $\hat{\lambda}_{j3}$ | 0,015(0,010) | 0,021(0,010) | 0,012(0,007) | 0,024(0,007) | 0,012(0,004) | 0,025(0,005) |

| Estimateur | % de systèmes masqués envoyés à la 2ième étape | | | |
|---|---|---|---|---|
| | 50% | | 100% | |
| | $j=1$ | $j=2$ | $j=1$ | $j=2$ |
| $\hat{P}_j$ | 0,103(0,056) | 0,499(0,068) | 0,101(0,054) | 0,498(0,060) |
| $\hat{P}_{g\|j}$ | 0,897(0,056) | 0,501(0,068) | 0,899(0,054) | 0,502(0,060) |
| $\hat{\lambda}_{j1}$ | 0,003(0,003) | 0,006(0,004) | 0,003(0,002) | 0,006(0,003) |
| $\hat{\lambda}_{j2}$ | 0,020(0,008) | 0,040(0,011) | 0,020(0,007) | 0,040(0,010) |
| $\hat{\lambda}_{j3}$ | 0,012(0,003) | 0,025(0,004) | 0,012(0,003) | 0,025(0,004) |

### 6.2.3 Résultats et discussion

Les tableaux 6.1 et 6.2 présentent les estimations de $P_j$, $P_{g|j}$ et $\lambda_{jk}$, $j = 1, 2$, $k = 1, 2, 3$ pour les risques concurrents non-proportionnels et proportionnels respectivement pour 1000 échantillons de taille 100. Dans ces tableaux, on présente les estimations faites pour un envoi de 0%, 5%, 20%, 50% et 100% des systèmes masqués à la deuxième étape. Seuls les résultats pour des échantillons de taille 100 sont présentés dans ces tableaux, même si on a aussi simulé des échantillons de taille 40. Pour la taille de 40, les résultats sont assez similaires à ceux avec une taille de 100. Il y a par contre une légère perte au niveau de la précision des estimations, les écarts-types étant un petit peu plus élevés. On pouvait s'y attendre puisqu'il y a généralement moins d'exactitude des estimateurs pour de plus petits échantillons.

Dans les deux tableaux, on remarque que les estimations des $\lambda_{jk}$, $j = 1, 2$, $k = 1, 2, 3$ sont assez près des valeurs théoriques, mais avec des écarts-types assez élevés. Lorsqu'il y a des systèmes envoyés à la deuxième étape, les moyennes des estimateurs sont presque toujours pareilles aux $\lambda_{jk}$ théoriques. Si aucun système n'est envoyé à la seconde étape, une légère différence s'installe entre les moyennes des estimateurs et les valeurs théoriques. Il semble que le fait d'en envoyer ne serait-ce que 5% à la deuxième étape améliore la valeur des estimations des $\lambda_{jk}$. La variance des estimateurs lorsqu'aucun système n'est investigué à l'étape 2 semble encore plus grande pour les risques proportionnels que pour les risques non-proportionnels.

Les figures 6.1 et 6.2 représentent la deuxième ligne des tableaux 6.1 et 6.2, c'est-à-dire les estimations des probabilités de masque, $\hat{P}_{g|j}$, $j = 1, 2$. Ce sont les graphiques des estimateurs $\hat{P}_{g|1}$ et $\hat{P}_{g|2}$ en fonction du pourcentage de systèmes masqués envoyés à la deuxième étape pour 1000 échantillons de taille 100. La ligne horizontale jaune représente la valeur théorique des probabilités, tandis que les lignes verticales bleues et rouges représentent les valeurs des estimateurs plus ou moins un écart-type pour les risques non-proportionnels et proportionnels, respectivement.

On constate dans ces 2 graphiques que plus il y a de systèmes masqués investigués à la deuxième étape, plus la valeur moyenne de l'estimateur est près de la valeur théorique et plus l'écart-type diminue. De plus, il ne semble pas y avoir de différence significative entre les estimateurs obtenus avec les risques non-proportionnels et proportionnels lorsqu'il y a 20% ou plus de systèmes masqués acheminés à l'étape 2. Pour les pourcentages de 0 et de 5, les valeurs sont assez différentes entre les estimations des risques non-proportionnels et proportionnels pour la figure 6.1. Dans le cas de la figure 6.2,

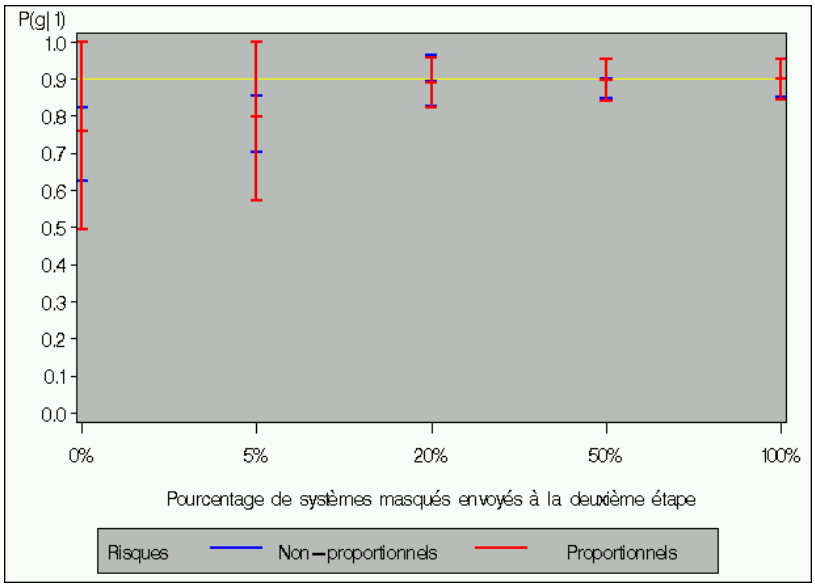


FIG. 6.1 – Probabilité de masque, $\hat{P}_{g|1}$, en fonction du pourcentage de systèmes masqués envoyés à la deuxième étape pour 1000 échantillons de taille 100, où la valeur théorique de l'estimateur est de 0.9.

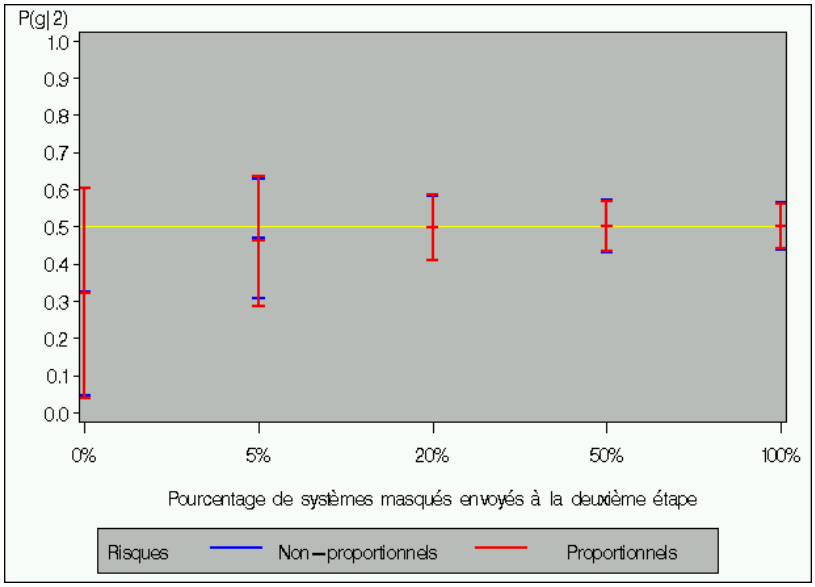


FIG. 6.2 – Probabilité de masque, $\hat{P}_{g|2}$, en fonction du pourcentage de systèmes masqués envoyés à la deuxième étape pour 1000 échantillons de taille 100, où la valeur théorique de l'estimateur est de 0.5.

l'estimation varie un peu entre les deux types de risques lorsqu'il y a 5% des systèmes masqués transmis à la deuxième étape. Dans les deux cas, c'est l'estimation sous les risques proportionnels qui dévie le plus de la valeur théorique.

Le tableau 6.3 indique les estimations de $P_j$, $P_{g|j}$, et $\lambda_{jk}$, $j = 1, 2$, $k = 1, 2, 3$, pour les risques non-proportionnels et proportionnels pour 1000 échantillons de tailles 40, 100 et 1000 lorsqu'aucun système masqué n'est transféré à la deuxième étape. En général, on remarque dans ce tableau que les estimations des $\lambda_{jk}$ sont plus précises pour les échantillons de taille 1000 que pour les deux autres tailles. Il ne semble par contre pas y avoir de différence importante entre les valeurs de ces estimateurs pour les tailles de 40 et de 100. Ce tableau indique aussi que les estimations des $P_j$ et $P_{g|j}$ sous les risques proportionnels ont une grande variabilité, peu importe la taille. Pour les risques non-proportionnels, les valeurs des $\hat{P}_j$ et $\hat{P}_{g|j}$ avec les échantillons de taille 1000 ont une moyenne plus près des valeurs théoriques et un écart-type plus petit que celles des tailles de 40 et de 100. De plus, il semble que sous les risques non-proportionnels, les estimations de $P_{g|1}$ soient moins variables que celles de $P_{g|2}$ pour les trois tailles. Ce résultats n'est pas surprenant étant donné qu'une probabilité de 0.9 est plus facile à estimer que 0.5. En effet, la probabilité de 0.5 est en théorie celle qui présente une plus grande variance.

Les figures 6.3 et 6.4 contiennent les estimateurs des probabilités de masque, $\hat{P}_{g|1}$ et $\hat{P}_{g|2}$, respectivement, pour 1000 échantillons de taille 40, 100 et 1000 lorsqu'aucun système n'est envoyé à la deuxième étape. La légende des couleurs est la même que pour les figures 6.1 et 6.2. Les figures 6.3 et 6.4 représentent la deuxième ligne de chacune des tailles du tableau 6.3.

En regardant ces graphiques, on voit bien que les risques proportionnels ont presque toujours la même variabilité pour les 3 tailles et pour les deux estimateurs. Cette variabilité est assez élevée, car les moyennes plus ou moins un écart-type couvrent environ la moitié de l'espace paramétrique. La figure 6.3 montre que pour la taille 1000 il y a un très grand contraste entre les estimations pour les risques proportionnels et non-proportionnels. Il semble y avoir eu un énorme dérapage des estimations pour les risques proportionnels. Dans tous les cas, l'algorithme EM sous-estime en moyenne les estimateurs $\hat{P}_{g|j}$ par rapport à la valeur théorique.

Il est aussi intéressant de noter que le nombre d'itérations nécessaires pour calculer les estimateurs avec l'algorithme EM lorsque les risques sont proportionnels est en moyenne beaucoup plus grand que lorsque les risques ne sont pas proportionnels. De plus, plus on envoie de systèmes masqués à la deuxième étape, moins il y a d'itérations effectuées par l'algorithme. Ce constat va dans le même sens que la théorie qui dit que

TAB. 6.3 – Estimateurs pour les risques non-proportionnels et proportionnels pour 1000 échantillons **sans deuxième étape**. Les valeurs théoriques sont données par : $P_{g|1} = 0.9$, $P_{g|2} = 0.5$, $\lambda_{11} = 0,003$, $\lambda_{12} = 0,02$ et $\lambda_{13} = 0,012$. Pour les risques non-proportionnels, $\lambda_{21} = 0,0045$, $\lambda_{22} = 0,01$, et $\lambda_{23} = 0,03$. Pour les risques proportionnels, $\lambda_{21} = 0,006$, $\lambda_{22} = 0,04$, et $\lambda_{23} = 0,024$.

| | | Risques non-proportionnels | | Risques proportionnels | |
|---|---|---|---|---|---|
| Taille | Estimateur | $j = 1$ | $j = 2$ | $j = 1$ | $j = 2$ |
| 40 | $\hat{P}_j$ | 0,150 (0,246) | 0,688 (0,280) | 0,178 (0,277) | 0,663 (0,270) |
| | $\hat{P}_{g\|j}$ | 0,850 (0,246) | 0,312 (0,280) | 0,822 (0,277) | 0,337 (0,270) |
| | $\hat{\lambda}_{j1}$ | 0,004 (0,005) | 0,003 (0,005) | 0,005 (0,005) | 0,004 (0,006) |
| | $\hat{\lambda}_{j2}$ | 0,021 (0,013) | 0,009 (0,011) | 0,024 (0,022) | 0,036 (0,024) |
| | $\hat{\lambda}_{j3}$ | 0,017 (0,011) | 0,026 (0,012) | 0,015 (0,010) | 0,022 (0,010) |
| 100 | $\hat{P}_j$ | 0,159 (0,216) | 0,676 (0,278) | 0,211 (0,297) | 0,679 (0,284) |
| | $\hat{P}_{g\|j}$ | 0,841 (0,216) | 0,324 (0,278) | 0,789 (0,297) | 0,321 (0,284) |
| | $\hat{\lambda}_{j1}$ | 0,003 (0,003) | 0,004 (0,004) | 0,004 (0,004) | 0,005 (0,005) |
| | $\hat{\lambda}_{j2}$ | 0,021 (0,010) | 0,009 (0,008) | 0,025 (0,018) | 0,035 (0,019) |
| | $\hat{\lambda}_{j3}$ | 0,016 (0,011) | 0,027 (0,011) | 0,015 (0,010) | 0,021 (0,010) |
| 1000 | $\hat{P}_j$ | 0,104 (0,044) | 0,574 (0,188) | 0,329 (0,376) | 0,618 (0,287) |
| | $\hat{P}_{g\|j}$ | 0,896 (0,044) | 0,426 (0,188) | 0,671 (0,376) | 0,382 (0,287) |
| | $\hat{\lambda}_{j1}$ | 0,003 (0,002) | 0,004 (0,002) | 0,003 (0,003) | 0,006 (0,003) |
| | $\hat{\lambda}_{j2}$ | 0,020 (0,004) | 0,010 (0,003) | 0,020 (0,016) | 0,040 (0,016) |
| | $\hat{\lambda}_{j3}$ | 0,014 (0,007) | 0,028 (0,007) | 0,012 (0,010) | 0,024 (0,010) |

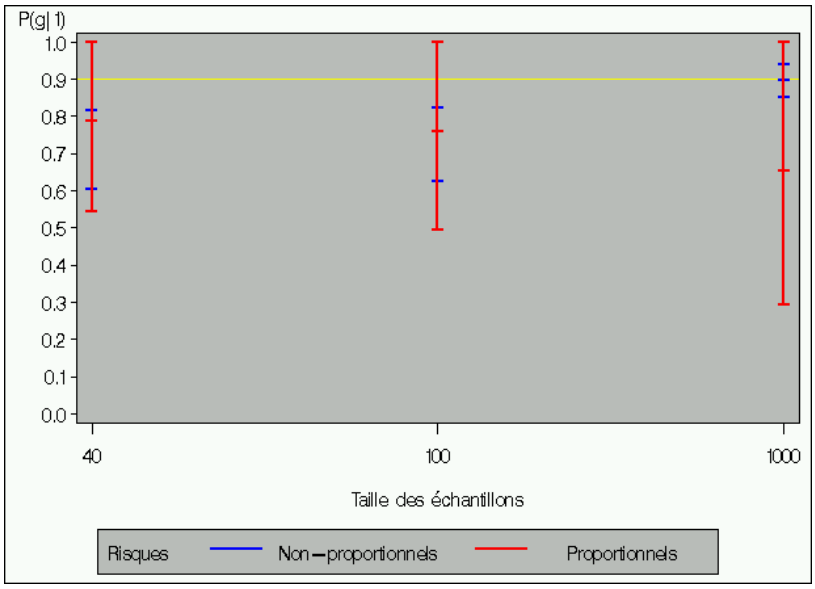


FIG. 6.3 – Probabilité de masque, $\hat{P}_{g|1}$, en fonction de la taille des échantillons simulés lorsqu'il y a absence de deuxième étape, où la valeur théorique de l'estimateur est 0.9.

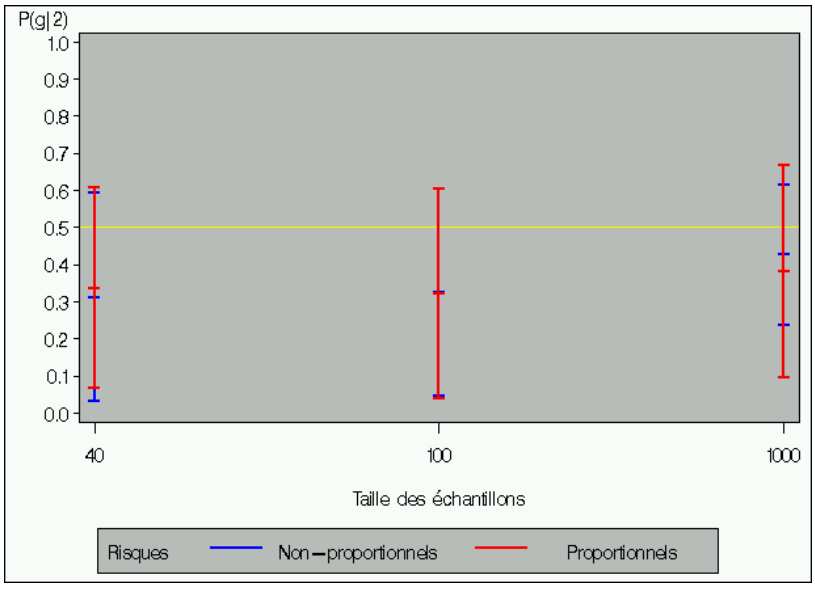


FIG. 6.4 – Probabilité de masque, $\hat{P}_{g|2}$, en fonction de la taille des échantillons simulés lorsqu'il y a absence de deuxième étape, où la valeur théorique de l'estimateur est 0.5.

TAB. 6.4 – Nombre d'itérations de l'algorithme EM pour 1000 échantillons de taille 100.

| Risques | % de systèmes masqués envoyés à la 2ième étape | | | | |
|---|---|---|---|---|---|
| | 0% | 5% | 20% | 50% | 100% |
| Non-proportionnels | 2425 (6249) | 214 (160) | 92 (155) | 38 (65) | 1 (0) |
| Proportionnels | 3935 (24348) | 290 (515) | 102 (180) | 43 (96) | 1 (0) |

TAB. 6.5 – Nombre d'itérations de l'algorithme EM pour 1000 échantillons **sans deuxième étape**.

| Risques | Taille des échantillons | | |
|---|---|---|---|
| | 40 | 100 | 1000 |
| Non-proportionnels | 860 (2287) | 2425 (6249) | 4696 (12484) |
| Proportionnels | 941 (5964) | 3935 (24348) | 23057 (48686) |

plus il manque de données, plus l'algorithme EM converge lentement (voir équation (3.52)). Le tableau 6.4 montre ces résultats. Notons que les nombres présentés dans ce tableau sont les moyennes du nombre d'itérations pour les 1000 échantillons de taille 100 accompagnés de leurs écarts-types entre parenthèses. On remarque que ces derniers augmentent lorsque les risques sont proportionnels et lorsqu'il y a moins de systèmes masqués envoyés à la deuxième étape. Par contre, on voit au tableau 6.5 que plus la taille des échantillons est grande, plus il y a d'itérations en moyenne. Ce résultat est logique puisqu'avec une grande taille d'échantillon, la fonction de log-vraisemblance a une plus grande sensibilité aux valeurs des paramètres.

Comme on a pu le remarquer dans les tableaux 6.1 à 6.3 et les graphiques 6.1 à 6.4, lorsqu'il n'y a pas de systèmes masqués envoyés à la deuxième étape, les estimateurs sont plutôt variables. À cause de ces résultats douteux, on s'est demandé si c'était le choix des points de départ dans l'algorithme EM qui pouvait faire converger l'estimateur vers des points stationnaires différents. Des tests ont donc été faits pour vérifier cette hypothèse avec 100 échantillons de tailles 100 et 1000.

Pour chacun des 100 échantillons, on a fait rouler l'algorithme EM 3 fois avec des points de départ tout à fait différents, mais toujours situés dans l'espace paramétrique. Si les estimations sont les mêmes pour les 3 points de départ, on garde la valeur de ces estimations et l'échantillon est retenu. Sinon, le résultat est invalide et l'échantillon est rejeté. Le tableau 6.6 présente le pourcentage d'échantillons retenus avec cette méthode.

TAB. 6.6 – Pourcentage d'échantillons dont les estimations sont les mêmes pour les 3 points de départ différents.

| Risques | Taille des échantillons | |
|---|---|---|
| | 100 | 1000 |
| Non-proportionnels | 53% | 95% |
| Proportionnels | 62% | 72% |

On remarque que pour une taille d'échantillon de 1000, il y a un plus grand pourcentage d'échantillons retenus que pour une taille de 100. Ce résultat est logique puisqu'en général plus la taille des échantillons est élevée, plus les estimateurs sont précis. De plus, pour une taille de 1000 la convergence vers un même point se fait dans plus de cas lorsque les risques sont non-proportionnels par rapport aux risques proportionnels. On pouvait s'y attendre étant donné que les résultats sous les risques proportionnels étaient généralement moins précis dans l'étude de simulation développée précédemment. Par contre, pour la taille de 100, le pourcentage est plus élevé sous les risques proportionnels, mais il n'y a pas une grande différence entre les pourcentages des deux types de risques. De plus, l'estimation d'une proportion n'est généralement pas aussi précise pour une taille de 100 que pour une taille de 1000, c'est ce qui peut expliquer que le pourcentage des risques proportionnels semble plus élevé pour la taille de 100. On a aussi fait le constat suivant pour chaque échantillon rejeté : pour au moins un des 3 points de départ, au moins un des estimateurs est sur la frontière de l'espace paramétrique.

On a tenté de trouver une caractéristique commune des $n_{jk}$, $j = 1, 2$, $k = 1, 2, 3$, lorsqu'il y avait convergence vers un même point pour les 3 points de départ. On a fait des analyses en composantes principales (ACP) et des arbres de décision, mais rien d'intéressant n'en est ressorti.

Par contre, dans le but de vérifier si les conditions sur $a_1$ et $a_2$ énoncées à la section 6.1.2 sont atteintes en pratique, on a fait des simulations avec $J = 2$ causes de pannes et $K = 2$ intervalles. On a pris le jeu de données qui a été utilisé pour le calcul du tableau 6.6 avec une taille de 100 et des risques non-proportionnels. On a fait le calcul des estimateurs avec l'algorithme EM en supposant qu'on n'a que 2 intervalles, soit $[0, 10]$ et $(10, \infty)$. Les estimateurs calculés sont

1. $\hat{P}_1$, $\hat{P}_{g|1}$, $\hat{P}_2$ et $\hat{P}_{g|2}$,
2. $\hat{\lambda}_{11}$, $\hat{\lambda}_{12}$, $\hat{\lambda}_{21}$, $\hat{\lambda}_{22}$.

Pour un échantillon donné, si les estimateurs sont les mêmes pour 3 points de départ différents choisis arbitrairement, alors l'échantillon est qualifié de convergent et il est conservé. Sinon, l'échantillon est rejeté. On a obtenu un pourcentage de 62% des échantillons qui ont été retenus.

Comme on l'a mentionné plus haut, on veut vérifier si les conditions de la section 6.1.2 sur $a_1$ et $a_2$ sont bel et bien atteintes. Si elles le sont, cela veut dire que $\hat{\phi}_{11}$ et $\hat{\phi}_{12}$ sont dans leur espace paramétrique et ainsi que tous les autres estimateurs le sont aussi. Pour arriver à vérifier les critères, on a calculé les $n_{jk}, j = 1, 2, k = 1, 2$, et on a calculé $a_1$ et $a_2$ pour les 38% d'échantillons qui étaient rejetés indépendamment des 62% qui ne l'étaient pas. Pour les échantillons rejetés, aucun échantillon n'a atteint les critères, comme on s'y attendait. Pour les échantillons retenus, 97% (60 échantillons sur 62) ont atteint les conditions sur $a_1$ et $a_2$.

On a essayé d'autres points de départ pour les 3% restants (2 échantillons sur 62), mais tous les essais convergent vers les mêmes estimateurs. Par contre, en regardant ces estimateurs calculés avec l'algorithme EM, on s'aperçoit que ceux des 2 échantillons concernés sont sur la frontière de l'espace paramétrique, alors que les 60 autres ne le sont pas. Ceci est un signe que les EMV seraient en-dehors de l'espace paramétrique, comme on s'y attendait selon les $n_{jk}, j = 1, 2, k = 1, 2$, et donc des $a_1$ et $a_2$, de cet échantillon.

## 6.3 Calcul d'estimateurs à partir des données de Dinse (1986)

Dans l'article de Dinse (1986) vu à la section 5.1, les estimateurs $\hat{\xi}(t)$ et $\hat{\psi}(t)$ ont été calculés pour des données d'une expérience sur 58 souris dont 33 étaient atteintes du NRVD. Dans la section 5.1, on a vu que ces estimateurs représentaient les proportions d'animaux ayant une cause de décès masquée parmi tous les animaux mourant au temps $t$ soit d'une autre cause que le NRVD ($\hat{\xi}(t)$) ou du NRVD ($\hat{\psi}(t)$). Dinse a obtenu des résultats sous plusieurs contraintes différentes sur ces estimateurs. La contrainte pour laquelle on a de l'intérêt ici est celle qui stipule que les estimateurs sont constants dans le temps, c'est-à-dire que $\hat{\xi}(t) = \hat{\xi}$ et $\hat{\psi}(t) = \hat{\psi}$.

C'est sous cette contrainte que les deux estimateurs de Dinse sont le plus près des estimateurs qu'on veut calculer avec l'algorithme EM, soit $\hat{P}_{g|1}$ et $\hat{P}_{g|2}$, où $g = \{1, 2\}$. En effet, ces derniers sont indépendants du temps. Le lien entre les estimateurs de Dinse et les estimateurs $\hat{P}_{g|1}$ et $\hat{P}_{g|2}$ se traduit comme suit si on suppose qu'un décès dû au NRVD est un décès de la cause $j = 1$ :

$$\begin{aligned}
\hat{\psi}(t) &= P(U = 1|D = 1, Y = 1, T = t) \\
\Leftrightarrow \quad \hat{\psi} &= P(U = 1|D = 1, Y = 1) \\
&= P(g = \{1, 2\}|j = 1, Y = 1) \\
&= \hat{P}^*_{g|1} \approx \hat{P}_{g|1},
\end{aligned}$$

$$\begin{aligned}
\hat{\xi}(t) &= P(U = 1|D = 0, Y = 1, T = t) \\
\Leftrightarrow \quad \hat{\xi} &= P(U = 1|D = 0, Y = 1) \\
&= P(g = \{1, 2\}|j = 2, Y = 1) \\
&= \hat{P}^*_{g|2} \approx \hat{P}_{g|2}.
\end{aligned}$$

On rappelle que $U$, $D$ et $Y$ sont des indicatrices définies comme suit : $U = 1$ si la cause de décès est inconnue, $D = 1$ si la cause de décès est le NRVD et $Y = 1$ si la souris avait la maladie lors du décès.

Pour les deux équations ci-haut, s'il n'y avait pas la condition $Y = 1$, on aurait une égalité stricte au lieu d'une égalité approximative. Cette condition signifie que les souris

prises en considération dans $\hat{P}^*_{g|1}$ et $\hat{P}^*_{g|2}$ sont atteintes de la maladie. Ainsi, au lieu de considérer les 58 souris dans les calculs des estimateurs obtenus avec l'algorithme EM, on doit prendre seulement les 33 souris atteintes du NRVD.

Les estimateurs $\hat{P}^*_{g|1}$ et $\hat{P}^*_{g|2}$ décris précédemment sont calculés grâce au modèle et à l'algorithme EM de Craiu et Duchesne (2004) avec les données de Dinse (1986). Les 33 souris considérées sont atteintes de la maladie. Parmi celles-ci, 8 souris sont décédées de la cause 1 (NRVD), 19 de la cause 2 (autre cause) et 6 souris ont une cause de décès masquée ($g = \{1, 2\}$). Il n'y a pas de deuxième étape pour les données masquées. Pour calculer les estimateurs, on utilise l'algorithme EM avec des fonctions de risque constantes par paliers. Les essais ont été faits avec 4 intervalles ((0,548], (549,730], (731,913] et (914,1095] jours) et 3 intervalles ((0,730], (731,913] et (914,1095] jours). Plusieurs points de départ différents ont été testés pour exécuter l'algorithme EM. Pour tous ces points, on a obtenu les mêmes estimateurs. De plus, des écarts-types ont pu être calculés pour chacun des estimateurs. Par contre, l'algorithme SEM n'a pu être utilisé pour faire ces calculs. En effet, cet algorithme s'arrête au milieu des calculs et on suppose que ce problème est dû à un nombre insuffisant de données. On a donc dû se tourner vers une autre méthode pour le calcul des écarts-types : la méthode du bootstrap.

La méthode d'échantillonnage du bootstrap consiste en un rééchantillonnage avec remise des $n = 33$ observations qu'on avait déjà dans notre échantillon de départ. Afin de pouvoir calculer les écarts-types des estimateurs, on a simulé 300 échantillons de taille 33 avec la méthode du bootstrap. On a alors utilisé l'algorithme EM pour calculer 300 estimateurs pour chacune des causes de décès. On a pu ainsi estimer les écarts-types entre ces estimations par les écarts-types échantillonnals. Les estimateurs calculés avec cette méthode sont plus conservateurs que ceux de l'algorithme SEM puisque la méthode du bootstrap est non-paramétrique tandis que pour l'algorithme SEM, la paramétrisation d'une distribution en escalier est considérée.

Le tableau 6.7 présente les résultats obtenus par Dinse en comparaison avec ceux obtenus par l'algorithme EM. Pour les estimateurs trouvés avec l'algorithme EM, les écarts-types calculés avec la méthode du boostrap sont entre parenthèses à côté de la valeur de l'estimateur. Par contre, pour les estimateurs de Dinse, il n'y a pas d'écarts-types puisque Dinse (1986) n'en donne pas.

En regardant le tableau 6.7, on remarque que les estimateurs trouvés avec l'algorithme EM sont très près de ceux trouvés avec la méthode de Dinse (1986). En effet, lorsque la fonction de risque est constante sur 4 intervalles, les valeurs des estimateurs sont exactement les mêmes que celles de Dinse, mais les écarts-types sont assez grands.

TAB. 6.7 – Comparaison entre les estimateurs de Dinse (1986) et de l'algorithme EM pour 33 souris femelles atteintes du NRVD, où $\hat{\psi}$ et $\hat{\xi}$ sont les estimateurs de Dinse pour $j = 1$ et $j = 2$ respectivement et où les chiffres entre parenthèses représentent les écart-types des estimateurs.

| Cause ($j$) | Algorithme EM ($\hat{P}^*_{g\|j}$) | | Dinse ($\hat{\psi}$ et $\hat{\xi}$) |
|---|---|---|---|
| | 4 intervalles | 3 intervalles | |
| 1 | 0.259 (0.211) | 0.289 (0.204) | 0.259 |
| 2 | 0.144 (0.118) | 0.126 (0.111) | 0.144 |

Lorsque la fonction de risque est constante sur 3 intervalles, il y a une petite différence entre les estimateurs de l'algorithme EM et ceux de Dinse et les écarts-types sont encore assez élevés.

Il semble donc que les estimateurs trouvés avec l'algorithme EM lorsqu'il y a 4 intervalles sont les plus près de ceux de Dinse. Il n'y a cependant pas de différence significative non plus entre les valeurs des estimateurs de l'algorithme EM et de Dinse lorsqu'il y a 3 intervalles.

## 6.4 Calcul d'estimateurs à partir des données de Flehinger, Reiser et Yashchin (2002)

Le premier exemple de l'article de Flehinger, Reiser et Yashchin de 2002 présente des données d'une compagnie qui construit des disques durs d'ordinateurs. La compagnie tente d'analyser les causes de panne des disques durs. Flehinger, Reiser et Yashchin considèrent qu'il y a 3 causes de panne possibles et qu'elles surviennent de façon indépendante.

Il y a quatre ans, 10 000 disques durs ont été mis en service et depuis ce temps, les informations sur les pannes ont été enregistrées. Pendant cette période, 172 disques durs sont tombés en panne. À la première étape, on a observé que 15 disques durs sont tombés en panne de la cause 1, 13 de la cause 2 et 11 de la cause 3, tandis que 64 ont été masqués dans le groupe $\{1, 3\}$ et 69 dans le groupe $\{1, 2, 3\}$. Dans l'article de

Flehinger, Reiser et Yachshin (2002), les estimateurs ont été calculés avec un modèle de risques concurrents paramétrisé par la loi de Weibull. 67 des 133 systèmes qui étaient masqués à la première étape ont été analysés plus profondément à la deuxième étape. Les estimateurs ont aussi été trouvés lorsqu'aucune donnée n'était envoyée à la deuxième étape. Dans cet article, les estimateurs calculés sont les paramètres de la Weibull ainsi que les probabilités de masque $\hat{P}_{\{1,3\}|1}$, $\hat{P}_{\{1,3\}|3}$, $\hat{P}_{\{1,2,3\}|1}$, $\hat{P}_{\{1,2,3\}|2}$ et $\hat{P}_{\{1,2,3\}|3}$.

Ces données ont été analysées de nouveau dans l'article de Craiu et Duchesne (2004), mais cette fois, c'est un modèle avec des fonctions de risque constantes par paliers qui a été utilisé. Les $K = 4$ intervalles pour les fonctions de risque sont les suivants : $[0, 1], (1, 2], (2, 3]$ et $(3, 4]$ (en années). Le calcul des estimateurs s'est fait avec l'algorithme EM et l'étude a été faite seulement dans le cas où les 67 données de deuxième étape sont disponibles.

Dans ce mémoire, les estimateurs sont maintenant calculés avec le modèle des risques constants par intervalles, mais lorsqu'aucune donnée de deuxième étape n'est disponible. L'algorithme EM est utilisé et il y a 133 données masquées sur 172. Les intervalles utilisés pour les fonctions de risque sont exactement les mêmes que ceux de l'article de Craiu et Duchesne (2004) mentionnés ci-haut. Les écarts-types sont calculés avec l'algorithme SEM (voir section (3.8)). De plus, des essais ont été faits avec plusieurs points de départ différents pour les paramètres de l'algorithme EM. Dans chaque cas, on a obtenu les mêmes estimateurs.

Le tableau 6.8 présente les estimations des probabilités de masque obtenues par les 4 méthodes différentes. On remarque de façon générale que $\hat{P}_{g|3}$ est précisément estimée. Cet estimateur est très semblable d'une méthode à l'autre et les écarts-types sont très petits. Pour l'estimateur $\hat{P}_{g|1}$, si on compare aux résultats obtenus lorsqu'il y a une deuxième étape, les estimateurs obtenus par l'algorithme EM (sans 2ième étape) semblent être mieux évalués que ceux de Flehinger, Reiser et Yashchin (sans 2ième étape). En effet, l'estimation de $\hat{P}_{\{1,2,3\}|1}$ pour Flehinger, Reiser et Yashchin (sans 2ième étape) semble avoir été impossible. Par contre, pour l'estimateur $\hat{P}_{\{1,2,3\}|2}$, il semble cette fois que ce soit l'algorithme EM (sans 2ième étape) qui n'ait pas fonctionné.

Dans le tableau 6.9, on présente les estimateurs des fonctions de risque, $\hat{\lambda}_{jk}$, $j = 1, 2, 3$, $k = 1, ..., 4$, de Craiu et Duchesne (2004) et de l'algorithme EM sans deuxième étape. On remarque que lorsqu'il n'y a pas de données de deuxième étape, les estimateurs ne sont pas souvent semblables à ceux calculés lorsqu'il y a une deuxième étape et les écarts-types sont généralement plus élevés sans la deuxième étape. De plus, on constate que les valeurs des $\hat{\lambda}_{1k}$ et des $\hat{\lambda}_{2k}$, $k = 1, ..., 4$, sans les données de deuxième étape sont toujours beaucoup plus petits que les estimateurs de Craiu et Duchesne

TAB. 6.8 – Comparaison entre les estimateurs des probabilités de masque, $\hat{P}_{g|j}$, de Flehinger, Reiser et Yashchin (2002), de Craiu et Duchesne (2004) et de l'algorithme EM sans deuxième étape pour 10 000 disques durs.

| Source | Groupe ($g$) | Cause ($j$) | | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| Flehinger et al. | $\{1,3\}$ | 0,412 | – | 0,446 |
| | $\{1,2,3\}$ | 0,310 | 0,469 | 0,436 |
| Flehinger et al. (pas de 2ième étape) | $\{1,3\}$ | 0,530 | – | 0,450 |
| | $\{1,2,3\}$ | 0,000 | 0,620 | 0,450 |
| Craiu et Duchesne | $\{1,3\}$ | 0,410 (0,0789) | – | 0,445 (0,0563) |
| | $\{1,2,3\}$ | 0,308 (0,0766) | 0,457 (0,1190) | 0,439 (0,0565) |
| Algorithme EM (pas de 2ième étape) | $\{1,3\}$ | 0,243 (0,2212) | – | 0,445 (0,0614) |
| | $\{1,2,3\}$ | 0,310 (0,2156) | 0,000 (0,0000) | 0,467 (0,0626) |

(2004). Pour la troisième cause, les estimateurs sont plus similaires.

Il faut préciser qu'il y a 133 données sur 172 qui sont masquées à la première étape, ce qui représente 77% des observations. Lorsqu'il y a une investigation à la deuxième étape, il n'y a que 38% des observations qui sont masquées. C'est environ deux fois moins de données masquées que lorsqu'il n'y a pas de deuxième étape. Cet écart peut expliquer en grande partie la difficulté de l'algorithme EM à trouver les estimateurs de façon précise.

En général, on a donc décelé que les estimateurs des tableaux 6.8 et 6.9 obtenus avec l'algorithme EM sans deuxième étape sont mieux estimés pour la cause 3 que pour les deux autre causes. Il y avait pourtant moins de systèmes dont la vraie cause de panne était dévoilée à la première étape pour le risque 3 que pour les deux autres causes. Mais ce phénomène est peut-être expliqué par le fait que si un système tombe en panne dans les intervalles 2 à 4, il y a une très bonne chance que cette panne soit due à la cause 3 (voir tableau 6.9) ; il est par contre très difficile de départager entre les causes 1 et 2 dans ce cas.

TAB. 6.9 – Comparaison entre les estimateurs des fonctions de risques, $\hat{\lambda}_{jk}$, de Craiu et Duchesne (2004) et de l'algorithme EM sans deuxième étape pour 10 000 disques durs.

| Source | Intervalle ($k$) | Cause ($j$) | | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| Craiu et Duchesne | 1 | 0,00204 (0,0005) | 0,00095(0,0003) | 0,00032 (0,0002) |
| | 2 | 0,00120 (0,0004) | 0,00026(0,0002) | 0,00205 (0,0005) |
| | 3 | 0,00083 (0,0004) | 0,00053(0,0003) | 0,00329 (0,0006) |
| | 4 | 0,00129 (0,0004) | 0,00067(0,0003) | 0,00392 (0,0007) |
| Algorithme EM (pas de 2ième étape) | 1 | 0,00136 (0,0010) | 0,00070(0,0003) | 0,00125 (0,0010) |
| | 2 | 0,00066 (0,0006) | 0,00010(0,0001) | 0,00275 (0,0007) |
| | 3 | 0,00062 (0,0005) | 0,00030(0,0002) | 0,00372 (0,0008) |
| | 4 | 0,00074 (0,0007) | 0,00020(0,0001) | 0,00494 (0,0010) |

# Chapitre 7

# Conclusion

Dans ce mémoire, on a d'abord expliqué la théorie se rapportant à l'algorithme EM et on a vu que cette technique est utilisée pour calculer les estimateurs du maximum de vraisemblance lorsque des données sont manquantes. On a aussi parlé du modèle des risques concurrents ; celui-ci se résume en un modèle de survie dont les sujets ou les systèmes à l'étude décédent ou tombent en panne dans le temps d'une cause parmi plusieurs possibles. Dans ce contexte, on a énoncé une fonction de vraisemblance qui permet de calculer des estimateurs tels les fonctions de risque et les fonctions de survie.

Il arrive parfois que la cause de panne des systèmes soit masquée dans un sous-groupe des causes possibles. Dans cette situation, la fonction de vraisemblance doit être réécrite en tenant compte que de l'information est manquante. Pour ce faire, différents modèles on été élaborés par des chercheurs. Parmi ceux-ci, on a expliqué le modèle nonparamétrique de Dinse (1986), le modèle des risques concurrents proportionnels et la modélisation paramétrique de Flehinger, Reiser et Yashchin (1998, 2002), ainsi que le modèle de risques concurrents constants par intervalles de Craiu et Duchesne (2004). Ce sont ces derniers auteurs qui exploitent le plus l'algorithme EM pour le calcul de leurs estimateurs ; ceux-ci étant les fonctions de risques spécifiques à chacune des causes de panne, les probabilités de masque ainsi que les probabilités de diagnostic.

De plus, on a vu que si la cause de panne de certains systèmes est masquée, il peut arriver qu'un échantillon de ces systèmes soit envoyé à une deuxième étape pour déterminer la vraie cause de panne. Avec leur modèle de risques proportionnels, Flehinger, Reiser et Yashchin (1998) ont un problème d'identification de certains estimateurs lorsque cette deuxième étape est absente. Par contre, sous d'autres modèles, tels ceux

de Flehinger, Reiser et Yashchin (2002) et de Craiu et Duchesne (2004), les estimateurs peuvent être calculés en l'absence de cette étape. De plus, à partir de Craiu et Reiser (2004), on a trouvé que, pour $J = 2$ et $K = 2$, certaines conditions sur le nombre de systèmes tombant en panne dans chaque intervalle pour chacune des causes, $n_{jk},\ j = 1, 2,\ k = 1, 2$, fait en sorte que les estimateurs sont identifiables ou non.

Sous le modèle de Craiu et Duchesne (2004), on a simulé un jeu de données avec deux causes de panne possibles et avec des fonctions de risque spécifiques aux causes de panne qui sont constantes sur 3 intervalles. Les estimateurs des probabilités de masque, $\hat{P}_{g|j},\ g = \{1, 2\},\ j = 1, 2$, ainsi que des fonctions de risque constantes par intervalles, $\hat{\lambda}_{jk},\ j = 1, 2,\ k = 1, 2, 3$, ont été calculés avec l'algorithme EM lorsqu'aucun système n'est envoyé à la deuxième étape. Dans cette étude, on a remarqué qu'en l'absence de deuxième étape, les estimateurs ont une grande variabilité. Ce phénomène semble encore plus prononcé lorsque les risques concurrents sont proportionnels. On a vu aussi que plus il y a de systèmes dont les causes de panne sont masquées, plus l'algorithme EM est lent à converger et plus la variance des estimateurs est élevée. On a aussi remarqué que plus la taille des échantillons est élevée, plus le nombre d'itérations est élevé.

Pour vérifier les conditions d'identifiabilité de Craiu et Reiser (2004), on a simulé un jeu de données avec 2 causes de panne et 2 intervalles. On a obtenu des résultats qui répondaient à nos attentes. En effet, pour chaque échantillon, on roulait 3 fois l'algorithme EM avec des points de départ différents à chaque fois. On s'est aperçu que les estimateurs étaient les mêmes pour les 3 points de départ lorsqu'on était sous les conditions d'identifiabilité pour presque tous les échantillons, et qu'ils étaient différents en dehors des conditions.

Il serait intéressant de développer ces conditions d'identifiabilité des estimateurs à toutes les situations. On pourrait donc savoir, selon les $n_{jk},\ j = 1, ..., J,\ k = 1, ..., K$, pour des $J \geq 2$ et $K \geq 1$ donnés, si les estimateurs vont être identifiables ou non en l'absence de deuxième étape. Il aurait été intéressant aussi de faire des simulations avec des intervalles différents pour chacune des causes. Selon Craiu et Reiser (2004), cette façon de faire permet de contourner des problèmes de non-identifiabilité en l'absence des données de deuxième étape.

Finalement, dans les études où il est possible d'envoyer des systèmes à la deuxième étape, il serait intéressant d'optimiser la sélection de l'échantillon choisi pour être investigué à cette étape. Ainsi, peut-être que la deuxième étape pourrait être encore plus importante qu'elle ne l'est présentement pour calculer les estimateurs avec précision.

# Bibliographie


[1] Casella, G., Berger, R.L. (2002), *Statistical Inference*, Second Edition. Pacific Grove : Duxbury.

[2] Craiu, R.V., Duchesne, T. (2004), Inference based on the EM algorithm for the competing risk model with masked causes of failure, *Biometrika*, **91**, 543-558.

[3] Craiu, R.V., Reiser, B. (2004), Inference for the dependent competing risks model with masked causes of failure, Rapport technique, University of Toronto.

[4] Dempster, A.P., Laird, N.M et Rubin, D.B. (1977), Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society. Series B*, **39**, 1-38.

[5] Dinse, G.E. (1986), Nonparametric prevalence and mortality estimators for animal experiments with incomplete cause-of-death data, *Journal of the American Statistical Association*, **81**, 328-335.

[6] Flehinger, B. J., Reiser, B., et Yashchin, E. (1996), Inference about defects in the presence of masking, *Technometrics*, **38**, 247-255.

[7] Flehinger, B. J., Reiser, B., et Yashchin, E. (1998), Survival with competing risks and masked causes of failure, *Biometrika*, **85**, 247-256.

[8] Flehinger, B. J., Reiser, B., et Yashchin, E. (2002), Parametric modeling for Survival with competing risks and masked causes of failure, *Lifetime Data Analysis*, **8**, 177-203.

[9] Goetghebeur, E. et Ryan, L. (1995), Analysis of competing risks survival data when some failure types are missing, *Biometrika*, **82**, 821-833.

[10] Hogg, R.V., Craig, A.T. (1995), *Introduction to Mathematical Statistics*, Fifth Edition. New Jersey : Prentice Hall.

[11] Kalbfleisch, J.D. et Prentice, R.L. (2002), *The statistical analysis of failure time data*, Second Edition. New York : Wiley.

[12] Klein, J.P. et Moeschberger M.L. (2003), *Survival Analysis : Techniques for Censored and Truncated Data*, Second Edition. New York : Springer.

[13] Little, R.J.A. et Rubin, D.B. (2002), *Statistical Analysis with Missing Data*, Second Edition. New Jersey : Wiley.

[14] Louis, T.A. (1982), Finding the observed information matrix when using the EM algorithm, *Journal of the Royal Statistical Society B*, **44**, 226-233.

[15] McLachlan, G.J. et Krishnan, T. (1997), *The EM Algorithm and Extensions*. New Jersey : Wiley.

[16] Meng, X.L. et Rubin, D.B. (1991), Using EM to obtain asymptotic variance-covariance matrices : the SEM algorithm, *Journal of the American Statistical Association*, **86**, 899-909.

[17] Prentice, R.L., Kalbfleisch, J.D., Peterson, AV Jr., Flournoy, N., Farewell, V.T., Breslow, N.E. (1978), The analysis of failure times in the presence of competing risks, *Biometrics*, **34**, 541-554.

[18] Raguet, C., *Tout sur l'économétrie : Les tests bootstrap*. En ligne, page consultée le 9 janvier 2005. http ://membres.lycos.fr/raguet/econometrie/bootstrap1.html.

[19] Ross, S.M. (1996), *Initiation aux probabilités*, Traduction de la quatrième édition américaine. Lausanne : Presses polytechniques et universitaires romandes.

[20] Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Survey*. New York : Wiley.

[21] Smith, C.A.B. (1977), discussion of Maximum Likelihood from Incomplete Data via the EM Algorithm, par A.P. Dempster, N.M. Laird et D.B. Rubin, *Journal of the Royal Statistical Society. Series B*, **39**, 24-25.

[22] Wu, C. F. (1983), On the convergence properties of the EM algorithm, *The Annals of Statistics*, **11**, 95-103.

# Annexe A

# Algorithme de simulation des temps de panne d'une loi à risques constants par palliers

Voici un algorithme pour générer des temps provenant de fonctions de risques spécifiques constantes par intervalles (dans le cas où $K = 3$ intervalles).

1. ENTRÉE :
   Les points de coupure $a_0 = 0$, $a_1$, $a_2$ et $a_K = \tau$.
   La cause de panne $j$.
   Les fonctions de risques spécifiques : $\lambda_{j1}$, $\lambda_{j2}$ et $\lambda_{j3}$.

2. $\lambda_j(t) \leftarrow \begin{cases} \lambda_{j1} & \text{si } \ a_0 < t \leq a_1, \\ \lambda_{j2} & \text{si } \ a_1 < t \leq a_2, . \\ \lambda_{j3} & \text{si } \ a_2 < t < \tau \end{cases}$

3. Générer un intervalle $k$ ($k = 1, 2$ ou $3$) :
   - $S_j(t) \leftarrow \exp\{-\int_0^t \lambda_j(u)\, du\}$, $t = a_1, a_2, \tau$.
   - $F_j(t) = 1 - S_j(t), \ \ t = a_1, a_2, \tau$.
   - Générer $U \sim U(0, 1)$.
   - $k \leftarrow \begin{cases} 1 & \text{si } \ 0 \leq U < F_j(a_1), \\ 2 & \text{si } \ F_j(a_1) \leq U < F_j(a_2), \\ 3 & \text{si } \ F_j(a_2) \leq U < F_j(\tau) = 1. \end{cases}$

4. Générer $t$ :
   - Si $k = 1$ $(0 < t \leq a_1)$ :

$$\begin{aligned} F_j(t) &= 1 - S_j(t) = 1 - \exp\Big\{ - \int_0^t \lambda_j(u)\, du \Big\} = U \\ &\Leftrightarrow \; 1 - \exp\{-t\, \lambda_{j1}\} = U \\ &\Leftrightarrow \; t = \frac{-\ln(1-U)}{\lambda_{j1}}. \end{aligned}$$

   - Si $k = 2$ $(a_1 < t \leq a_2)$ :

$$\begin{aligned} F_j(t) &= 1 - S_j(t) = 1 - \exp\Big\{ - \int_0^t \lambda_j(u)\, du \Big\} = U \\ &\Leftrightarrow \; 1 - \exp\{-a_1\lambda_{j1} - (t-a_1)\lambda_{j2}\} = U \\ &\Leftrightarrow \; t = a_1 - \frac{\ln(1-U)}{\lambda_{j2}} - \frac{a_1\lambda_{j1}}{\lambda_{j2}}. \end{aligned}$$

   - Si $k = 3$ $(a_2 < t \leq \tau)$ :

$$\begin{aligned} F_j(t) &= 1 - S_j(t) = 1 - \exp\Big\{ - \int_0^t \lambda_j(u)\, du \Big\} = U \\ &\Leftrightarrow \; 1 - \exp\{-a_1\lambda_{j1} - a_1\lambda_{j2} - (t-a_2)\lambda_{j3}\} = U \\ &\Leftrightarrow \; t = a_2 - \frac{\ln(1-U)}{\lambda_{j3}} - \frac{a_1(\lambda_{j1}+\lambda_{j2})}{\lambda_{j3}}. \end{aligned}$$

5. SORTIE : Le temps de panne $t$.