Revue de statistique appliquée

Didier Chauveau

Algorithmes EM et SEM pour un mélange censuré de distributions de défaillances. Application à la fiabilité d'équipements électroniques en période de jeunesse

Revue de statistique appliquée, tome 40, nº 2 (1992), p. 67-76

<http://www.numdam.org/item?id=RSA_1992__40_2_67_0>

Rev. Statistique Appliquée, 1992, **XXXX** (2), 67-76

ALGORITHMES EM ET SEM POUR UN MÉLANGE CENSURÉ DE DISTRIBUTIONS DE DÉFAILLANCES APPLICATION A LA FIABILITÉ D'ÉQUIPEMENTS ÉLECTRONIQUES EN PÉRIODE DE JEUNESSE

Didier CHAUVEAU

Université Paris Sud (ORSAY), Alcatel CIT, Dpt. ATC (ORMES)

RÉSUMÉ

Dans le cadre d'une étude statistique des techniques de déverminage en milieu industriel, nous présentons des méthodes d'estimation par Maximum de Vraisemblance des paramètres d'un mélange caché de lois de probabilité en présence de données censurées. Plus précisément, nous donnons des extensions de l'algorithme EM, et de sa version stochastique SEM, pour des données incomplètes du fait du modèle de mélange sous-jacent et du fait de la censure d'une partie des observations. De plus, nous traitons le cas de mélanges de lois de Weibull, n'appartenant pas à la famille des lois exponentielles.

Mots-clés : *Algorithmes EM, SEM, Apprentissage probabiliste, Données censurées, lois de Weibull, Maximum de Vraisemblance, Mélange de lois.*

1. Introduction

L'élaboration de techniques d'estimation pour des modèles de mélange dans lesquels les observations sont en partie censurées est motivée à l'origine par une problématique issue du domaine de l'industrie électronique. Après une brève présentation des aspects industriels et du problème d'ordre statistique associé (§3), nous nous intéressons (§4), aux algorithmes de type EM, particulièrement adaptés aux problèmes de données incomplètes.

Ces algorithmes, dont nous rappelons les principes généraux et les développements actuels, n'ont pas encore, à notre connaissance, été utilisés dans un contexte où les données manquantes sont de deux natures strictement différentes. Nous donnons donc au §5 les formulations de EM et SEM dans ce cadre général, puis particularisées aux distributions exponentielles et Weibull, les plus utilisées en fiabilité. Il faut à ce propos souligner le fait que EM et SEM ont toujours été auparavant utilisés pour des mélanges de lois appartenant à la famille exponentielle (essentiellement des mélanges de Gaussiennes) ; les propriétés de convergence ont donc encore à être précisées à la fois pour les mélanges censurés et pour les mélanges de lois de Weibull, qui ne sont pas issues d'une famille exponentielle.

2. Aspects industriels

Les fabricants de matériel électronique procèdent, au terme de leur processus de fabrication, à une étape dite de «déverminage», qui consiste à placer les équipements en condition de fonctionnement et à forte température pendant quelques jours afin de révéler les défauts de jeunesse des composants constitutifs des matériels (voir par exemple [2]), et de faire franchir à ceux-ci ce que l'on nomme couramment leur *période de jeunesse*. Cette étape est extrêmement coûteuse, car elle porte généralement sur toute la production, et allonge le cycle de fabrication sans garantir une amélioration de la fiabilité du matériel ayant subi sans défaillance le déverminage. *La durée du déverminage* est donc un paramètre augmentant de façon drastique le prix des produits, ce qui justifie la recherche d'une méthode d'optimisation de cette durée.

Les méthodes du type plan d'expérience avec échantillonnage au niveau du matériel complet sont à bannir dans ce contexte, car les produits sont variés et il n'est pas possible d'obtenir des données du type "life test" à chaque nouvel équipement produit. Il est donc indispensable de calculer la loi d'un système à partir de celle de ses composants (voir [8] ou [3]).

Le principe retenu est de déterminer, pour une carte (Circuit Imprimé équipé, CIE) donnée, la loi de la v.a. T associée à sa durée de fonctionnement sans panne. Si nous supposons connues les lois des familles de composants constituant les différents produits fabriqués, la connaissance de la nomenclature[1] d'un CIE suffit à calculer la loi de sa v.a. T sans recourir à une expérimentation spécifique. On cherche alors à minimiser une fonction économique prenant en compte simultanément le coût $c(t_d)$ d'un déverminage de durée t_d, et le coût $a(.)$ d'une panne après mise en service réel du matériel, fonction de l'espérance de T sachant que l'on a appliqué un déverminage de durée t_d. Soit à déterminer

$$t_d^* \text{ tel que } \Phi(t_d^*) = \operatorname*{Min}_{t_d \in D} \{\Phi(t_d) = -a(E[T \mid T \geqslant t_d]) + c(t_d)\}$$

où le minimum est pris sur un ensemble D borné. Cette optimisation ne pose pas de problèmes autres que d'ordre numérique (calcul de l'espérance pour de nombreuses valeurs de t_d). Elle présente d'autre part l'avantage d'autoriser éventuellement une durée optimale t_d^* inférieure à celle de la période de jeunesse du CIE, si les coûts $c(t_d)$ et a le justifient, notamment si $c(t_d) \gg a$.

Le problème subsistant est donc de déterminer le type de la distribution de défaillance d'un composant électronique en début de vie, puis d'estimer cette distribution, ceci pour chaque famille de composants «à risques».

3. Modèles de mélange

Dans la littérature (par exemple [6] et [8]), le taux de défaillance instantané $\lambda(t)$ d'un composant électronique présente trois parties distinctes : jeunesse, vie

---

[1] Composition du CIE en types et quantités de composants.

utile et usure, caractérisées respectivement par des $\lambda(t)$ décroissant, constant (stabilisé), puis croissant (les trois périodes forment la célèbre courbe «en baignoire»). Il est classique d'utiliser dans ces modèles des lois de type Weibull qui, en raison du grand nombre de distributions qu'elles représentent, sont bien adaptées pour décrire chacune de ces périodes *prises séparément* (on joue pour cela sur le paramètre β de forme, cf. [6]). Mais nous nous intéressons ici à une modélisation portant sur l'ensemble des deux premières périodes de la vie du composant. D'autre part, l'apparition de «vices cachés» dans les composants d'une carte est due au fait que, dans la population d'où sont tirés ces composants, certains ont été fragilisés par leur propre processus de fabrication ou par d'autres stress extérieurs. Ces composants fragiles meurent durant ou peu après le déverminage, créant les pannes de jeunesse des systèmes. La population est donc structurellement composée de plusieurs *classes* : les composants sains, et les fragilisés par différentes causes. Dans les simulations, nous supposerons pour simplifier l'existence de deux sous-populations (fragilisés et sains).

Ceci nous conduit à considérer la distribution de l'instant de défaillance x d'un composant comme un *mélange de lois de probabilité* (cf. [11]), dans lequel chaque loi $p(x|\Phi_i)$, $i = 1 \dots m$ représente la distribution de défaillance associée à une sous-population i, et chaque poids α_i la probabilité qu'un composant tiré au hasard de la population totale appartienne à la population i. On considère donc la densité

$$p(x|\Phi) = \sum_{i=1}^{m} \alpha_i p(x|\Phi_i), \text{ avec } \Phi = (\alpha_1 \dots \alpha_{m-1}, \Phi_1 \dots \Phi_m) \text{ et } \sum_{i=1}^{m} \alpha_i = 1;$$

où le vecteur Φ des paramètres est à estimer à partir de l'observation d'un échantillon $\mathbf{x} = (x_1 \dots x_n)$ de données de type "life tests" pour ce composant.

Une «observation complète» issue d'un mélange serait la description (x_k, z_k) de l'instant de défaillance x_k et de la *provenance* de l'observation, représentée par un vecteur $z_k = (z_{k1} \dots z_{km})$, où $z_{ki} \in \{0, 1\}$ (cf. [1]). Comme on ne sait pas dire ici, lorsqu'on observe une défaillance, la sous-population d'où provient le composant, on est en présence d'un *mélange caché* (cf. [4] ou [9]) donc d'un problème de *données incomplètes du fait du mélange* (on n'observe pas les z_k). D'autre part, les expérimentations conduites en vue d'obtenir des instants de défaillances (life tests) à partir d'un échantillon de taille n sont fréquemment censurées, dans la mesure où il n'est pas possible, ou très coûteux, de mener un test jusqu'à l'obtention des n défaillances. Nous n'aurons affaire qu'à une censure de type I (cf. [6]), dans laquelle on arrête le test après un temps prédéterminé, obtenant ainsi r défaillances, $r \leqslant n$. Cela signifie que nous n'observons qu'un sous-échantillon de taille r issu de l'échantillon de départ, soit des *données incomplètes du fait de la censure.*

Le problème de l'estimation des paramètres d'un mélange, caché ou non, est largement traité pour les familles exponentielles de lois. La technique la plus employée et la plus efficace pour ce type de données incomplètes est l'algorithme EM ([4] et [9]) et, plus récemment ([1], [5]), sa version stochastique SEM, expérimentée dans [1] pour des mélanges de gaussiennes. Par contre, son utilisation dans le cas de données censurées est beaucoup plus rare : [4] donne un cadre général dans le cas d'une distribution simple censurée, mais pas d'utilisation réelle, ceci étant peut-être dû au fait que l'on dispose dans ce cas des méthodes classiques du Maximum de Vraisemblance ([6] et [8] donnent ces estimateurs pour des Weibull censurées de type I et II). Seul [7] traite des mélanges censurés dans le cas où l'on sait déterminer la provenance des observations; il donne une méthode itérative analogue à EM dans ce cas précis, bien qu'antérieure.

4. Algorithmes EM et SEM

4.1. EM (Expectation-Maximisation)

EM est, dans sa version la plus générale, un algorithme itératif de calcul des estimateurs du Maximum de Vraisemblance en présence de données incomplètes. L'idée intuitive est de remplacer la maximisation de la vraisemblance des données complètes, inobservée, par celle de son espérance conditionnelle aux observations. EM est particulièrement utile lorsque le Maximum de Vraisemblance sur les données complètes est facile à calculer. En voici la forme générale.

Soit $\mathbf{Y}$ l'espace des données complètes, et une application $\mathbf{y} \to \mathbf{x}(\mathbf{y})$ de $\mathbf{Y}$ sur l'espace $\mathbf{X}$ des données incomplètes. Soit $f(\mathbf{y}|\Phi)$ une famille de densités sur $\mathbf{Y}$, pour $\Phi \in \Omega$, et $g(\mathbf{x}|\Phi)$ une densité sur $\mathbf{X}$ induite par $f(\mathbf{y}|\Phi)$. EM a pour but de maximiser la log-vraisemblance des données incomplètes $L(\Phi) = \log\ g(\mathbf{x}|\Phi)$.

Posons $Q(\Phi|\Phi') = E[\log f(\mathbf{y}|\Phi)|\mathbf{x}, \Phi']$, et soit Φ^0 valeur d'initialisation du paramètre, Φ^p sa valeur à l'itération p. L'itération $p \to (p+1)$ se déroule comme suit :

étape E : Déterminer la fonction $Q(\cdot|\Phi^p)$.

étape M : Choisir Φ^{p+1} qui maximise $Q(\cdot|\Phi^p)$.

La propriété essentielle de EM est que toute suite (Φ^p) engendrée par cet algorithme vérifie $L(\Phi^{p+1}) \geqslant L(\Phi^p)$. Nous renvoyons à [4], [12] et principalement [9] pour une description complète et l'étude des propriétés de convergence dans le cas de familles exponentielles, à [1] pour la formulation dans le cas du mélange de deux gaussiennes. Nous allons donner pour notre part une version de EM pour un mélange *non censuré* de m distributions de Weibull à deux paramètres. Remarquons que, dans ce cas, l'étape M nécessite elle-même une méthode numérique itérative de recherche du zéro d'une fonction (comme d'ailleurs le Maximum de Vraisemblance usuel sur une simple loi de Weibull) en raison de la forme de la densité et de l'inexistence d'une statistique exhaustive autre que la statistique triviale.

Les observations incomplètes et complètes sont respectivement $\mathbf{x} = (x_1 \dots x_n) \in \Re_+^n$, et $\mathbf{y} = (\mathbf{x}, \mathbf{z})$, où $\mathbf{z} = (z_1 \dots z_n)$, z_k étant lui-même un vecteur à valeurs 0/1 indicateur de la provenance de x_k ($z_{ki} = 1$ si x_k provient de la composante i). Les couples (x_k, z_k), $1 \leqslant k \leqslant n$, sont indépendants et de même loi, et le modèle pour les x_k est $g(x|\Phi) = \sum_{i=1}^{m} \alpha_i p(x|\Phi_i)$, avec $\Phi_i = (\beta_i, \eta_i)$, et $p(x|\Phi_i) = \frac{\beta_i}{\eta_i} \left(\frac{x}{\eta_i}\right)^{\beta_i - 1} \exp\left[-\left(\frac{x}{\eta_i}\right)^{\beta_i}\right]$.

L'itération $p \longrightarrow (p+1)$ s'écrit de la façon suivante :

étape E : L'algorithme EM présente une forme particulièrement attractive lorsque le modèle est un mélange (non censuré), car alors la détermination de $Q(\Phi|\Phi')$ est particulièrement simple, les z_{ki} apparaissant linéairement dans l'espérance conditionnelle. $Q(\Phi|\Phi')$ est ainsi simplement la log-vraisemblance des données complètes, dans laquelle tout z_{ki} est remplacé par $E[Z_i|x_k, \Phi']$; il s'agit donc ici de calculer

$$Z_i^{(p+1)}(x_k) = E[Z_i|X = x_k, \Phi^{(p)}] = \frac{\alpha_i^{(p)} p(x_k|\Phi_i^{(p)})}{\sum_{t=1}^{m} \alpha_t^{(p)} p(x_k|\Phi_t^{(p)})},$$

pour $k = 1 \ldots n;\ i = 1 \ldots m.$

étape M :

$$\alpha_i^{(p+1)} = \frac{1}{n} \sum_{k=1}^{n} Z_i^{(p+1)}(x_k),\ i = 1 \ldots m-1; \quad \text{et} \quad \alpha_m^{(p+1)} = 1 - \sum_{i=1}^{m-1} \alpha_i^{(p+1)}$$

$\beta_i^{(p+1)}$ solution de :

$$\frac{1}{\beta} - \frac{\sum_{k=1}^{n} Z_i^{(p+1)}(x_k) \log(x_k) x_k^{\beta}}{\sum_{k=1}^{n} Z_i^{(p+1)}(x_k) x_k^{\beta}} + \frac{\sum_{k=1}^{n} Z_i^{(p+1)}(x_k) \log(x_k)}{\sum_{k=1}^{n} Z_i^{(p+1)}(x_k)} = 0;\ i = 1 \ldots m.$$

$$\text{et } \eta_i^{(p+1)} = \left(\frac{\sum_{k=1}^{n} Z_i^{(p+1)}(x_k) x_k^{\beta_i^{(p+1)}}}{\sum_{k=1}^{n} Z_i^{(p+1)}(x_k)} \right)^{\frac{1}{\beta_i^{(p+1)}}} ;\ i = 1 \ldots m.$$

Cette version, testée sur simulations pour $m = 2$, présente les problèmes de convergence et les limitations habituellement constatés avec EM (voir [12], [1] p.7). En plus, nous ne bénéficions pas avec les Weibull des propriétés liées aux familles exponentielles. Cependant, il paraît raisonnable en l'état actuel d'espérer une propriété de convergence locale.

4.2. SEM (Stochastic Expectation Maximisation)

Algorithme d'apprentissage probabiliste, SEM ([1]) repose sur un principe de tirage aléatoire des $Z_i^{(p+1)}(x_k)$ au cours d'une étape S insérée entre les étapes E et M. Les $Z_i^{(p+1)}(x_k)$ sont donc, comme pour les données complètes, à valeur 0/1. Dans les cas étudiés (mélanges de gaussiennes), il se comporte plus efficacement que EM, et évite notamment les convergences vers des points selle ou des «méplats» de la vraisemblance en raison de l'effet apporté par la perturbation aléatoire. SEM est directement applicable à notre mélange de Weibull comme suit :

étape E : Calculer $T_i^{(p+1)}(x_k) = E[Z_i|X = x_k, \Phi^{(p)}] = \dfrac{\alpha_i^{(p)} p(x_k|\Phi_i^{(p)})}{\sum_{t=1}^{m} \alpha_t^{(p)} p(x_k|\Phi_i^{(p)})}$, pour $k = 1 \dots n\,;\, i = 1 \dots m$.

étape S : Simuler $(Z_i^{(p+1)}(x_k)\,;\, i = 1 \dots m)$ suivant la multinomiale $\mathcal{M}(1; T_i^{(p+1)}$ $(x_k)\,;\, i = 1 \dots m)$, $k = 1 \dots n$.

étape M : Identique à EM.

Les premières simulations effectuées pour deux composantes donnent des estimations souvent meilleures que EM à taille d'échantillon égale (et assez grande), mais qui dépendent tout de même fortement du point initial. L'extension de propriétés de convergence au cas Weibull (censuré ou non) est actuellement à l'étude.

5. Mélanges censurés

5.1. EM_{CM} (Censored Mixture)

Nous nous plaçons dans le cadre général du §4, sans préciser pour l'instant les distributions. Posons $\Re_+ = R_1 \cup R_2$, chaque observation x issue du mélange n'étant connue que si $x \in R_1$ (censurée sur R_2); ainsi, pour une censure droite en c (cas qui nous intéresse ici) $R_1 = [0, c[$. L'échantillon complet sera noté $\mathbf{y} = (\tilde{\mathbf{x}}, \mathbf{z})$, avec $\tilde{\mathbf{x}} = (x_1 \dots x_n)$ et $\mathbf{z} = (z_1 \dots z_n)$ où $z_k = (z_{k1} \dots z_{km})$. Soit $C = \{k : x_k \in R_1\}$ l'ensemble des indices k pour lesquels la censure n'est pas intervenue, et $r = \mathrm{Card}(C)$. Les données incomplètes peuvent alors s'écrire traditionnellement $\mathbf{x} = \{x_k : k \in C\}$. Dans ce cadre la densité des données complètes s'écrit

$$f(\mathbf{y}|\Phi) = \prod_{k=1}^{n} \prod_{i=1}^{m} (\alpha_i^{z_{ki}} p(x_k|\Phi_i)^{z_{ki}}); \text{ et } Q(\Phi|\Phi^{(p)}) = E[\log\ f(\mathbf{y}|\Phi)|\mathbf{x}, \Phi^{(p)}]$$

on notera $Z_{1,i}^{(p+1)}(x) = E[Z_i|X = x, \Phi^{(p)}]$, et $Z_{2,i}^{(p+1)} = E[Z_i|X \in R_2, \Phi^{(p)}]$.

L'itération $p \longrightarrow (p+1)$ de $\mathrm{EM_{CM}}$ a la forme générale suivante :

étape E : on a

$$Q(\Phi|\Phi^{(p)}) = \sum_{k\in C}\sum_{i=1}^{m} Z_{1,i}^{(p+1)}(x_k)\log(\alpha_i) + \sum_{k\in C}\sum_{i=1}^{m} Z_{1,i}^{(p+1)}(x_k)\log(p(x_k|\Phi_i))$$

$$+(n-r)\sum_{i=1}^{m} Z_{2,i}^{(p+1)}\log(\alpha_i) + (n-r)\sum_{i=1}^{m} Z_{2,i}^{(p+1)} \frac{\int_{R_2}\log(p(x|\Phi_i))p(x|\Phi_i^{(p)})dx}{\int_{R_2} p(x|\Phi_i^{(p)})dx}.$$

il s'agit donc ici de calculer

$$Z_{1,i}^{(p+1)}(x_k) = E[Z_i|X = x_k, \Phi^{(p)}] = \frac{\alpha_i^{(p)}p(x_k|\Phi_i^{(p)})}{\sum_{t=1}^{m}\alpha_t^{(p)}p(x_k|\Phi_t^{(p)})}, \text{ pour } k \in C;\ i = 1\dots m.$$

$$\text{et } Z_{2,i}^{(p+1)} = E[Z_i|X \in R_2, \Phi^{(p)}] = \frac{\alpha_i^{(p)}\int_{R_2} p(x|\Phi_i^{(p)})dx}{\sum_{t=1}^{m}\alpha_t^{(p)}\int_{R_2} p(x|\Phi_t^{(p)})dx}, \text{ pour } i = 1\dots m.$$

étape M : Comme dans les cas non censurés, on peut maximiser séparément en α_i et en Φ_i.

$$\alpha_i^{(p+1)} = \frac{1}{n}\left(\sum_{k\in C} Z_{1,i}^{(p+1)}(x_k) + (n-r)Z_{2,i}^{(p+1)}\right),\ i = 1\dots m-1;$$

$$\text{et } \alpha_m^{(p+1)} = 1 - \sum_{i=1}^{m-1}\alpha_i^{(p+1)}$$

La maximisation en Φ_i, elle, dépend des distributions utilisées. Le problème est le calcul de $\int_{R_2}\log(p(x|\Phi_i))p(x|\Phi_i^{(p)})dx$ de manière à obtenir une fonction de Φ_i que l'on puisse maximiser. Dans le cas de lois exponentielles, cela ne pose pas de problème et nous donnons ci-dessous EM_{CM} sous forme explicite; par contre, dans le cas de lois de Weibull, il faut recourir à des étapes d'intégration numérique, ce qui alourdit considérablement les calculs. Il faut souligner que la version stochastique (§5.2) ne présente pas ces inconvénients d'ordre numérique et est donc bien plus facilement programmable.

Itération $p \to (p+1)$ de EM_{CM} pour des lois exponentielles :

on a $p(x|\Phi_i) = p(x|\eta_i) = \frac{1}{\eta_i}\exp\left(\frac{-x}{\eta_i}\right)$, et $R_1 = [0, c[$

étape E :

$$Z_{1,i}^{(p+1)}(x_k) = \frac{\alpha_i^{(p)} p(x_k|\Phi_i^{(p)})}{\sum_{t=1}^{m} \alpha_t^{(p)} p(x_k|\Phi_t^{(p)})},$$

$$\text{et } Z_{2,i}^{(p+1)} = \frac{\alpha_i^{(p)} \exp\left(\frac{-c}{\eta_i^{(p)}}\right)}{\sum_{t=1}^{m} \alpha_t^{(p)} \exp\left(\frac{-c}{\eta_t^{(p)}}\right)}, \text{ pour } k \in C; \; i = 1 \ldots m.$$

étape M :

$$\alpha_i^{(p+1)} = \frac{1}{n}\left(\sum_{k \in C} Z_{1,i}^{(p+1)}(x_k) + (n-r)Z_{2,i}^{(p+1)}\right), \; i = 1 \ldots m-1;$$

$$\text{et } \alpha_m^{(p+1)} = 1 - \sum_{i=1}^{m-1} \alpha_i^{(p+1)}$$

$$\eta_i^{(p+1)} = \frac{\sum_{k \in C} Z_{1,i}^{(p+1)}(x_k)x_k + (n-r)Z_{2,i}^{(p+1)}(c + \eta_i^{(p)})}{\sum_{k \in C} Z_{1,i}^{(p+1)}(x_k) + (n-r)Z_{2,i}^{(p+1)}}, \; i = 1 \ldots m.$$

5.2. SEM_{CM}

Comme dans le cas classique, il s'agit de compléter des données manquantes à l'aide de générations aléatoires. Nous avons choisi ici de simuler la totalité des données manquantes, c'est-à-dire les z_k (la loi de tirage variant selon que $k \in C$ ou $k \notin C$) et, pour tout $k \notin C$, x_k (la loi de tirage dépendant du z_k simulé correspondant). L'étape M se ramène alors à une simple maximisation de la log-vraisemblance des données complètes sur l'échantillon «complété». Remarquons qu'un autre choix aurait été de simuler uniquement les z_k et de garder la logique EM d'usage de l'espérance mathématique en ce qui concerne les x_k censurés.

L'itération $p \to (p+1)$ s'écrit :

étape E : Calculer

$$T_{1,i}^{(p+1)}(x_k) = E[Z_i|X = x_k, \Phi^{(p)}] = \frac{\alpha_i^{(p)} p(x_k|\Phi_i^{(p)})}{\sum_{t=1}^{m} \alpha_t^{(p)} p(x_k|\Phi_t^{(p)})}, \text{ pour} k \in C; \; i = 1 \ldots m.$$

$$\text{et } T_{2,i}^{(p+1)} = E[Z_i | X \in R_2, \Phi^{(p)}] = \frac{\alpha_i^{(p)} \int_{R_2} p(x|\Phi_i^{(p)})dx}{\sum_{t=1}^{m} \alpha_t^{(p)} \int_{R_2} p(x|\Phi_t^{(p)})dx}, \text{ pour } i = 1 \dots m.$$

étape S : Simuler r v.a. $(Z_i^{(p+1)}(x_k);\ i = 1 \dots m)$ suivant $\mathcal{M}(1; T_{1,i}^{(p+1)}(x_k),\ i = 1 \dots m)$, pour $k \in C$.

Simuler $(n-r)$v.a. $(Z_{ij}^{(p+1)};\ i = 1 \dots m)$ suivant $\mathcal{M}(1; T_{2,i}^{(p+1)},\ i = 1 \dots m)$, $j = 1 \dots (n-r)$.

Simuler $(n-r)$v.a. $x_j^{(p+1)}$ suivant la loi déduite de $\prod_{i=1}^{m} p(x|\Phi_i^{(p)})^{Z_{ij}^{(p+1)}}$ par conditionnement par $\{x \in R_2\}$, $j = 1 \dots (n-r)$.

étape M : Il s'agit de maximiser séparément en α_i et en Φ_i la fonction

$$Q(\Phi|\Phi^{(p)}) = \sum_{i=1}^{m} \left(\sum_{k \in C} Z_i^{(p+1)}(x_k) + \sum_{j=1}^{n-r} Z_{ij}^{(p+1)} \right) \log(\alpha_i)$$
$$+ \sum_{i=1}^{m} \left(\sum_{k \in C} Z_i^{(p+1)}(x_k) \log\ p(x_k|\Phi_i) + \sum_{j=1}^{n-r} Z_{ij}^{(p+1)} \log\ p(x_j^{(p+1)}|\Phi_i) \right)$$

En α_i, on a :

$$\alpha_i^{(p+1)} = \frac{1}{n} \left(\sum_{k \in C} Z_i^{(p+1)}(x_k) + \sum_{j=1}^{n-r} Z_{ij}^{(p+1)} \right),\ i = 1 \dots m-1;$$

$$\text{et } \alpha_m^{(p+1)} = 1 - \sum_{i=1}^{m-1} \alpha_i^{(p+1)}$$

En Φ_i, on retrouve le même type de calcul que dans le cas de la maximisation de la log-vraisemblance des données complètes, en présence d'un mélange non censuré. En particulier pour un mélange de lois de Weibull, les résultats sont similaires à ceux du §4.1.

6. Conclusion

En conclusion, nous insisterons simplement sur le fait que ces algorithmes itératifs sont facilement utilisables pour des mélanges censurés de lois exponentielles. En revanche, l'utilisation de lois de Weibull implique toujours un recours à l'étape M à une méthode elle-même itérative, du type Newton-Raphson (NR). EM_{CM} pour des lois de Weibull nécessite en plus une intégration numérique par itération de l'algorithme NR, et peut donc s'avérer désespérément long, d'autant que la convergence de EM est elle-même lente. Les versions stochastiques semblent ainsi préférables autant du point de vue du calcul numérique que du point de vue de leur efficacité.

Références

[1] CELEUX G., DIEBOLT J. - *A Random Imputation Principle : The Stochastic EM Algorithm.* INRIA, rapport de recherche n° 901 (1988).

[2] Commission électrotechnique internationale - *Déverminage sous contraintes de dispositifs réparables fabriqués en lot.* Comité d'Etude n° 56 - Fiabilité et Maintenabilité, Part. I et II (1989).

[3] CORAZZA M. - *Techniques Mathématiques de la Fiabilité Prévisionnelle des Systèmes.* Cepadues Editions (1975).

[4] DEMPSTER A.P., LAIRD N.M., RUBIN D.B. - *Maximum Likelihood from Incomplete Data via the EM Algorithm.* J.R. Statis. Soc. B, **39**, 1-38 (1977).

[5] DIEBOLT J. - *Extrêmes et Fluctuations de Processus Aléatoires Stationnaires.* Thèse de Doctorat d'état, Université Pierre et Marie Curie, Paris VI (1989).

[6] MANN N.R., SCHAFER R.E., SINGPURWALLA N.D. - *Methods for Statistical Analysis of Reliability and Life Data.* J. Wiley & Sons (1973).

[7] MENDENHALL W., HADER R.J. - *Estimation of Parameters of Mixed Exponentially Distributed Failure Time Distributions from Censored Life Test Data.* Biometrika **45**, 504-520 (1958).

[8] PAGÈS A., GONDRAN M., MAGNIEN M. - *Fiabilité des Systèmes.* Eyrolles (1980).

[9] REDNER R.A., WALKER H.F. - *Mixture Densities, Maximum Likelihood and the EM Algorithm.* SIAM Review, **26**, n° 2, 195-239 (1984).

[10] SUNDBERG R. - *Maximum Likelihood Theory for Incomplete Data from an Exponential Family.* Scand J Statist **1**, 49-58 (1974).

[11] TITTERINGTON D.M., SMITH A.F.M., MAKOV U.E. - *Statistical Analysis of Finite Mixture Distribution.* New York, Wiley (1985).

[12] WU C.F. - *On the Convergence Properties of the EM Algorithm.* Ann. Statis. **11**, 95-103 (1983).