Afrika Statistika

**Afrika Statistika**
Vol. 13 (2), 2018, pages 1631–1643.
DOI: http://dx.doi.org/10.16929/as/1631.125

ISSN 2316-090X

# Global convergence and ascent property of a cyclic algorithm used for statistical analysis of crash data

**Issa Cherif Geraldo**[1,*], **Assi N'Guessan**[2] **and Kossi Essona Gneyou**[3]

[1]Ecole d'ingénieurs en Informatique et Technologies numériques, Université catholique de l'Afrique de l'Ouest - Unité universitaire du Togo, 01 B.P. 1502 Lomé 01, Lomé, Togo.
[2]Laboratoire Paul Painlevé UMR CNRS 8524, Université de Lille 1, 59655 Villeneuve d'Ascq Cedex, France.
[3]Département de Mathématiques, Faculté des Sciences, Université de Lomé, B.P. 1515 Lomé, Togo.

Received Febrary on 02, 2018, Accepted on March 17, 2018

**Abstract.** In this paper, we consider an estimation algorithm called cyclic iterative algorithm (CA) that is used in statistics to estimate the unknown vector parameter of a crash data model. We provide a theoretical proof of the global convergence of the CA that justifies the good numerical results obtained in early numerical studies of this algorithm. We also prove that the CA is an ascent algorithm, what ensures its numerical stability.

**Résumé.** Dans cet article, nous considérons un algorithme d'estimation appelé algorithme cyclic iteratif (CA) utilisé en statistique pour estimer le vecteur paramètre inconnu d'une modèle pour les données d'accidents. Nous donnons une preuve théorique de la convergence globale du CA qui justifie les excellents résultats numériques obtenues dans les études numériques antérieures dudit algorithme. Nous prouvons aussi que le CA augmente la log-vraisemblance à chaque itération, ce qui assure sa stabilité numérique.

**Key words:** Iterative method, Maximum likelihood, cyclic algorithm, global convergence, road safety.
**AMS 2010 Mathematics Subject Classification :** 62-04, 62F10, 62H12, 62P99.

*Corresponding Author: cherifgera@gmail.com
Assi.Nguessan@polytech-lille.fr
Kossi_gneyou@yahoo.fr

## 1. Introduction

In statistics, one is usually confronted to the problem of the estimation of the unknown parameter vector $\beta \in \mathbb{R}^d$ of a probability law $\mathcal{L}_\beta$. The first method and also the most used for this purpose is the Maximum Likelihood (ML) method (Cramer, 1946; Shao, 2003). If $Y_1, \ldots, Y_n$ are $n$ random variables considered as independent and identically distributed according to $\mathcal{L}_\beta$, the likelihood of observed data $y_1, \ldots, y_n$ is the product of the probabilities that each $Y_i$ takes the value $y_i$, $i = 1, \ldots, n$. The maximum likelihood method consists in maximizing the likelihood function $L(\beta)$ considered as a function of $\beta \in \mathbb{R}^d$. In practice, one prefers to maximize the log-likelihood function $\ell = \log L$. Except some school examples, the maximum likelihood estimation of $\beta$ often requires the use of a numerical iterative optimization method starting from an initial vector $\beta^{(0)}$ and computing successive approximations of the unknown solution by the recurrence formula

$$\beta^{(k+1)} = M(\beta^{(k)}) \tag{1}$$

where $M$ is a mapping from $\mathbb{R}^d$ into itself. Detailed reviews of modern numerical optimization methods can be found in Dennis and Schnabel (1996), Nocedal and Wright (2006), Lange (2013) and Lange et al. (2014).

The very first method that comes to mind is the Newton-Raphson's (NR) algorithm. The NR algorithm for the ML estimation of $\beta$ uses the iteration mapping

$$M(\beta^{(k)}) = -\left(\nabla^2 \ell(\beta^{(k)})\right)^{-1} \nabla \ell(\beta^{(k)}) \tag{2}$$

where $\nabla \ell$ is the gradient of $\ell$ and $\nabla^2 \ell$ is its Hessian matrix. Since the NR algorithm requires the inversion of the Hessian matrix at each iteration, its implementation may be difficult in large dimensions or if the Hessian matrix is ill-conditioned or singular at the point $\beta^{(k)}$. We can also mention the fact that the NR algorithm can diverge violently when the starting point $\beta^{(0)}$ is far from the true unknown value of $\beta$ (Dennis and Schnabel, 1996). When the NR algorithm fails, one can use one of the many algorithms that have been proposed in the literature as remedies. One can use the Fisher scoring algorithm which replaces the Hessian matrix by the expectation of its negative (Osborne, 1992) or quasi-Newton algorithms which compute an approximation of the inverse of the Hessian matrix using only the first derivatives (Nocedal and Wright, 2006). We can also mention Derivative Free Optimization (DFO) in which no derivative of the objective function is computed and the successive iterates are computed from the values of the objective function on a finite set of points (Rios and Sahinidis, 2013). DFO algorithms are of interest when $\ell$ is expensive to evaluate or non-differentiable. The recent decades have also seen the popularization of the Expectation Maximization (EM) algorithm (Dempster et al., 1977; McLachlan and Krishnan, 2008) which is considered as a special case of the more general class of MM (Minorization-Majorization) optimization algorithms (Hunter and Lange, 2004; Zhou and Lange, 2010). In maximization problems, the first M step consists in minorizing the objective function $\ell$ by a surrogate function $g(\beta|\beta^{(k)})$ and the second M step consists in maximizing this surrogate function with respect to $\beta$ to produce the next iterate $\beta^{(k+1)}$. MM algorithms are considered as effective algorithms for ML estimation because they consistently drive the likelihood uphill by maximizing a simple surrogate function for the log-likelihood. However, all these remedies brought by scientific results come at the cost of a greater and greater complexity and they are not always easy to implement.

Regardless of the algorithm used, global convergence is necessary. An optimization algorithm is said to be globally convergent if the sequence of iterates $\beta^{(k)}$ generated by this algorithm converges to a stationary point of $\ell$ (a point where the gradient vanishes) from any starting $\beta^{(0)}$ (Dennis and Schnabel, 1996). Proving the global convergence is always a delicate exercise and this property may not always be satisfied. For example, the NR algorithm is not globally convergent because the convergence to a stationary point of the sequence $\beta^{(k)}$ generated by the NR algorithm is guaranteed only if the starting point $\beta^{(0)}$ is sufficiently close to the true value of $\beta$ (Lange, 2013). For MM algorithms, the global convergence has been established under some conditions (Lange, 2010). In order to ensure numerical stability, a maximum likelihood estimation algorithm should also be an ascent algorithm i.e. it should increase the log-likelihood at each step.

In the context of statistical analysis of crash data, N'Guessan and Truffier (2008) have considered a parameter vector $\beta = (\theta, \phi^T)^T$ where $\theta$ is the first component of $\beta$ and $\phi$ is a vector consisting in the remaining components of $\beta$. They proved the existence of a unique stationary point and proposed an estimation algorithm called Cyclic iterative Algorithm (CA) that cycles through the two subsets of components updating each from the other one. More specifically, this CA starts from an initial value $\phi^{(0)}$ and updates successively $\theta^{(1)}$ from $\phi^{(0)}$, $\phi^{(1)}$ from $\theta^{(1)}$, $\theta^{(2)}$ from $\phi^{(1)}$ and so on until a convergence criteria is satisfied. N'Guessan and Geraldo (2015) studied some numerical convergence properties of the CA using simulated accident datasets. Their simulation studies suggest that the CA converges to the maximum likelihood estimator (MLE) $\hat{\beta}$ of $\beta$ from any starting point $\beta^{(0)}$ and outperforms the classical optimization algorithms such as NR and MM. Geraldo et al. (2015) proved that the MLE $\hat{\beta}$ is strongly consistent, that is, it converges almost surely to the true value of $\beta$ when the sample size tends to $+\infty$.

In this paper, we provide convergence results that justify the good numerical results given by N'Guessan and Geraldo (2015) and which complete the stochastic convergence results of Geraldo et al. (2015). We prove that the CA is globally convergent (it converges to the MLE from any starting value $\beta^{(0)}$) and is also an ascent algorithm (the log-likelihood increases at each iteration of the algorithm i.e. $\ell(\beta^{(k+1)}) \geqslant \ell(\beta^{(k)})$ for any iteration $k \geqslant 0$).

The rest of the paper is structured as follows. Section 2 provides the state of the art of the cyclic algorithm. In section 3, we provide some intermediate lemmas that will be used to prove the main convergence theorems given in Section 4. More precisely, we prove that the sequence of iterates $\beta^{(k)}$ generated by the CA converges to the MLE from every starting point and that the CA enjoys the ascent property. The paper finishes with some concluding remarks.

## 2. An overview of the cyclic algorithm

### *2.1. Problem setting and statistical model*

Consider an experimental site where accidents can be classified into $r$ $(r > 0)$ mutually exclusive accidents types by increasing severity (for example, property damage, minor injury, severe injury, fatal). Assume that a road safety measure (transformation of intersections into roundabouts, installation of roundabouts, modification of the ground marking, etc. . . ) has been applied at this experimental site with the purpose of reducing the occurrence of accidents. After a certain period of application of the measure, it is certainly desirable to know if the latter had a significant effect.

There exists in the literature a plethora of statistical models for the evaluation of a road safety measure. In general, these models strongly depend on the available data. Lord and Mannering (2010) and Mannering and Bhat (2014) provide comprehensive reviews of contemporary thinking in the crash frequency-analysis field and give the advantages and disadvantages of each approach. In this work, we consider the before-after model proposed by N'Guessan et al. (2001). One of the main advantages of this before-after model is that it allows cause-effect interpretations (Hauer, 2010).

Let $X = (X_{11}, \ldots, X_{1r}, X_{21}, \ldots, X_{2r})^T$ be a random vector where $X_{1j}$ (resp. $X_{2j}$), $j = 1, \ldots, r$, represents the number of crashes of type $j$ occurred in the "before" (resp. "after") period. In order to take into account some external factors (traffic flow, speed limit variation, weather conditions, etc. . .), the experimental site is associated to a control area where the safety measure was not applied. Let $Z = (z_1, \ldots, z_r)^T$ be a vector such that $z_j$ denotes the ratio of the number of accidents of type $j$ for the period "after" to the period "before" in the control area over the same time period. The model of N'Guessan et al. (2001) (in the case of one experimental site) is described by the following assumptions:

$(A_1)$ The total number of accidents observed on the experimental site where the measure was applied is a fixed constant denoted $n$.

$(A_2)$ The control coefficients $z_1, \ldots, z_r$ are known and non-random.

$(A_3)$ The random vector $X$ is assumed to have the multinomial distribution

$$X \sim \mathcal{M}(n; \pi_1(\beta), \pi_2(\beta))$$

where $\beta = (\theta, \phi^T)^T \in \mathbb{R}^{1+r}$, $\theta > 0$, $\phi = (\phi_1, \ldots, \phi_r)^T$ belongs to the simplex

$$\mathbb{S}_r = \Big\{(\phi_1, \ldots, \phi_r)^T \in \mathbb{R}^r \mid \phi_i > 0, \quad 1 \leqslant i \leqslant r, \quad \sum_{i=1}^{r} \phi_i = 1\Big\}, \tag{3}$$

$\pi_i(\beta) = (\pi_{i1}(\beta), \dots, \pi_{ir}(\beta))^T, \quad i = 1, 2$ and

$$\pi_{ij}(\beta) = \begin{cases} \dfrac{\phi_j}{1 + \theta \sum_{m=1}^r z_m \phi_m}, & i = 1; \quad j = 1, \dots, r, \\ \\ \dfrac{\theta z_j \phi_j}{1 + \theta \sum_{m=1}^r z_m \phi_m}, & i = 2; \quad j = 1, \dots, r \end{cases} \tag{4}$$

The parameter $\theta$ represents the mean effect of the road safety measure while the components of $\phi = (\phi_1, \dots, \phi_r)^T$ represent the different accident risks.

### *2.2. The cyclic algorithm (CA) for maximum likelihood estimation of the parameters*

For an observed data $x = (x_{11}, \dots, x_{1r}, x_{21}, \dots, x_{2r})$ such that $\sum_{i=1}^2 \sum_{j=1}^r x_{ij} = n$, set $x_{\cdot j} = x_{1j} + x_{2j}$ $(j = 1, \dots, r)$ and $x_{i\cdot} = \sum_{j=1}^r x_{ij}$ $(i = 1, 2)$. Then the log-likelihood is defined up to an additive constant by

$$\ell(\beta) = \sum_{j=1}^r \left( x_{\cdot j} \log(\phi_j) + x_{2j} \log(\theta) - x_{\cdot j} \log(1 + \theta \sum_{m=1}^r z_m \phi_m) \right). \tag{5}$$

Provided that it exists, the Maximum Likelihood Estimator, $\hat{\beta} = (\hat{\theta}, \hat{\phi}^T)^T$, of $\beta = (\theta, \phi^T)^T$ is solution to the constrained optimization problem:

$$\begin{cases} \hat{\beta} = \underset{\beta \in \mathbb{R}^{1+r}}{\operatorname{argmax}} \, \ell(\beta) \\ \text{subject to} \\ \forall j = 1, \dots, r, \quad \phi_j > 0, \quad \theta > 0 \quad \text{and} \quad \sum_{j=1}^r \phi_j = 1. \end{cases} \tag{6}$$

N'Guessan et al. (2001) proved that the MLE $\hat{\beta}$ is solution to the non-linear system of equations

$$\begin{cases} \displaystyle\sum_{j=1}^r \left( x_{2j} - \frac{x_{\cdot r} \hat{\theta} \sum_{m=1}^r z_m \hat{\phi}_m}{1 + \hat{\theta} \sum_{m=1}^r z_m \hat{\phi}_m} \right) = 0 \\ \\ x_{\cdot j} - \dfrac{n \hat{\phi}_j (1 + \hat{\theta} z_j)}{1 + \hat{\theta} \sum_{m=1}^r z_m \hat{\phi}_m} = 0, \qquad j = 1, \dots, r. \end{cases} \tag{7}$$

N'Guessan (2010) proved that the non-linear system (7) accepts a solution $\hat{\beta} = (\hat{\theta}, \hat{\phi}^T)^T$ such that

$$\begin{cases} \hat{\theta} = \dfrac{x_{2\cdot}}{x_{1\cdot}} \dfrac{1}{\sum_{j=1}^r z_j \hat{\phi}_j} \\ \\ \hat{\phi}_j = \dfrac{1}{1 - \dfrac{1}{n} \displaystyle\sum_{m=1}^r \frac{\hat{\theta} z_m x_{\cdot m}}{1 + \hat{\theta} z_m}} \times \dfrac{x_{\cdot j}}{n(1 + \hat{\theta} z_j)}, \qquad j = 1, \dots, r. \end{cases} \tag{8}$$

He then proposed an iterative procedure alternating between updating $\theta$ holding $\phi$ fixed and vice-versa. Because of the link between $\hat{\theta}$ and $\hat{\phi}$, his procedure starts from an initial vector $\phi^{(0)} = (\phi_1^{(0)}, \ldots, \phi_r^{(0)})^T$ such that $\sum_{j=1}^r \phi_j^{(0)} = 1$. At the step $k+1$, $\theta^{(k+1)}$ is updated from $\phi^{(k)} = (\phi_1^{(k)}, \ldots, \phi_r^{(k)})^T$, afterwards $\phi^{(k+1)} = (\phi_1^{(k+1)}, \ldots, \phi_r^{(k+1)})^T$ is updated from the $\theta^{(k+1)}$, and so on until the convergence of the sequence of vectors $\beta^{(k)} = (\theta^{(k)}, {\phi^{(k)}}^T)^T$ is achieved. This strategy yields the following iterative algorithm (N'Guessan, 2010):

Given an initial vector $\phi^{(0)} = (\phi_1^{(0)}, \ldots, \phi_r^{(0)})^T$ such that $\sum_{j=1}^r \phi_j^{(0)} = 1$ and for any $k \geqslant 0$,

$$\begin{aligned} \theta^{(k+1)} &= \frac{x_{2\cdot}}{x_{1\cdot}} \frac{1}{\sum_{j=1}^r z_j \phi_j^{(k)}} \\ \phi_j^{(k+1)} &= \frac{1}{1 - \dfrac{1}{n} \displaystyle\sum_{m=1}^r \frac{\theta^{(k+1)} z_m x_{\cdot m}}{1 + \theta^{(k+1)} z_m}} \times \frac{x_{\cdot j}}{n(1 + \theta^{(k+1)} z_j)}, \quad j = 1, \ldots, r. \end{aligned} \tag{9}$$

The aim of this paper is to prove that the sequence of vectors $\beta^{(k)} = (\theta^{(k)}, {\phi^{(k)}}^T)^T$ produced by Algorithm (9) converges to the MLE $\hat{\beta} = (\hat{\theta}, \hat{\phi}^T)^T$ of $\beta$. We also prove that Algorithm (9) is an ascent algorithm i.e. for all $k \geqslant 0$, $\ell(\beta^{(k+1)}) \geqslant \ell(\beta^{(k)})$.

## 3. Preliminary results

**Lemma 1.** *Let $\psi$ be the mapping defined on $\mathbb{R}_+$ by*

$$\psi(u) = -x_{1\cdot} + \sum_{m=1}^r \frac{x_{\cdot m}}{1 + u z_m}. \tag{10}$$

*i) There exists a unique real number $\theta^*$ such that $\psi(\theta^*) = 0$ and the MLE $\hat{\theta}$ of $\theta$ is equal to that unique root $\theta^*$ of $\psi$.*

*ii) For all $u > 0$, $\psi(u) \geqslant 0$ if $0 < u \leqslant \theta^*$ and $\psi(u) \leqslant 0$ if $u \geqslant \theta^*$.*

*Proof.*
*i)* One can easily check that $\psi$ is continuous and its derivative $\psi'(u)$ is strictly negative for every $u > 0$ and therefore $\psi$ is bijective. Moreover,

$$\lim_{u \to 0} \psi(u) \times \lim_{u \to +\infty} \psi(u) = (x_{2\cdot}) \times (-x_{1\cdot}) \;<\; 0.$$

Hence the equation $\psi(u) = 0$ has a unique solution denoted $\theta^*$. Let $j$ be an integer from the set $\{1, \ldots, r\}$. From the equalities

$$\hat{\phi}_j = \frac{1}{1 - \dfrac{1}{n} \displaystyle\sum_{m=1}^r \frac{\hat{\theta} z_m x_{\cdot m}}{1 + \hat{\theta} z_m}} \times \frac{x_{\cdot j}}{n(1 + \hat{\theta} z_j)}$$

and

$$1-\frac{1}{n}\sum_{m=1}^{r}\frac{\hat{\theta}z_m x_{\cdot m}}{1+\hat{\theta}z_m}=1-\frac{1}{n}\sum_{m=1}^{r}\left(x_{\cdot m}-\frac{x_{\cdot m}}{1+\hat{\theta}z_m}\right)=\frac{1}{n}\sum_{m=1}^{r}\frac{x_{\cdot m}}{1+\hat{\theta}z_m},$$

we may write

$$\hat{\phi}_j=\frac{x_{\cdot j}/(1+\hat{\theta}z_j)}{\sum_{m=1}^{r}x_{\cdot m}/(1+\hat{\theta}z_m)} \tag{11}$$

and then

$$\sum_{j=1}^{r}z_j\hat{\phi}_j=\frac{\sum_{j=1}^{r}\left(z_j x_{\cdot j}/(1+\hat{\theta}z_j)\right)}{\sum_{m=1}^{r}\left(x_{\cdot m}/(1+\hat{\theta}z_m)\right)}.$$

From the first line of (8), we deduce

$$\hat{\theta}=\frac{x_{2\cdot}}{x_{1\cdot}}\frac{\sum_{m=1}^{r}\left(x_{\cdot m}/(1+\hat{\theta}z_m)\right)}{\sum_{j=1}^{r}\left(z_j x_{\cdot j}/(1+\hat{\theta}z_j)\right)}.$$

This is equivalent to

$$\frac{x_{2\cdot}}{x_{1\cdot}}\sum_{m=1}^{r}\frac{x_{\cdot m}}{1+\hat{\theta}z_m}=\sum_{m=1}^{r}\frac{\hat{\theta}z_m x_{\cdot m}}{1+\hat{\theta}z_m}.$$

As we have $x_{1\cdot}+x_{2\cdot}=n$, we deduce the following equality:

$$\frac{n}{x_{1\cdot}}\sum_{m=1}^{r}\frac{x_{\cdot m}}{1+\hat{\theta}z_m}=n$$

which yields the equality $\psi(\hat{\theta})=0$. Since $\psi$ is bijective and $\psi(\theta^*)=0$ then $\hat{\theta}=\theta^*$.

*ii)* The function $\psi$ is a strictly decreasing function. Therefore,

$$\forall u\leqslant\theta^*,\quad \psi(u)\geqslant\psi(\theta^*)=0\quad\text{and}\quad\forall u\geqslant\theta^*,\quad\psi(u)\leqslant\psi(\theta^*)=0.$$

This completes the proof of Lemma 1.

**Lemma 2.** *Let $\alpha_x=x_{2\cdot}/x_{1\cdot}$ and $\Psi_x$ be the function from $]0;+\infty[$ to $]0;+\infty[$ defined by:*

$$\Psi_x(u)=\alpha_x\left(\sum_{m=1}^{r}\frac{x_{\cdot m}}{1+uz_m}\right)\Big/\left(\sum_{m=1}^{r}\frac{z_m x_{\cdot m}}{1+uz_m}\right).$$

*i) The real $\theta^*$ defined by Lemma 1 is the unique fixed point of $\Psi_x$.*
*ii) The function $\Psi_x$ is such that*

$$\forall u\leqslant\theta^*,\quad\Psi_x(u)\geqslant u\quad\textit{and}\quad\forall u\geqslant\theta^*,\quad\Psi_x(u)\leqslant u.$$

*iii) The sequence of real numbers $(\theta^{(k)})$ generated by Algorithm (9) is monotonous and its monotony depends on $\theta^{(0)}$. If $\theta^{(0)}<\theta^*$ then it is an increasing sequence and if $\theta^{(0)}>\theta^*$, it is a decreasing sequence.*
*iv) The sequence $(\theta^{(k)})$ is also bounded. Then it is convergent and its limit is $\theta^*$.*

*Proof.*
*i)* The equation $\Psi_x(u) = u$ is equivalent to

$$\frac{x_{2\cdot}}{x_{1\cdot}} \sum_{m=1}^{r} \frac{x_{\cdot m}}{1+uz_m} = \sum_{m=1}^{r} \frac{uz_m x_{\cdot m}}{1+uz_m}.$$

After straightforward computations similar to those used in the proof of Lemma 1, one gets $\psi(u) = 0$. This latter equation has a unique solution $\theta^*$ that is also the unique solution of the equation $\Psi_x(u) = u$.

*ii)* Let $u > 0$. Then

$$\begin{aligned}
\Psi_x(u) - u &= \left(\frac{x_{2\cdot}}{x_{1\cdot}} \sum_{m=1}^{r} \frac{x_{\cdot m}}{1+uz_m} - \sum_{m=1}^{r} \frac{uz_m x_{\cdot m}}{1+uz_m}\right) \Bigg/ \left(\sum_{m=1}^{r} \frac{z_m x_{\cdot m}}{1+uz_m}\right) \\
&= \left(\frac{x_{2\cdot}}{x_{1\cdot}} \sum_{m=1}^{r} \frac{x_{\cdot m}}{1+uz_m} - n + \sum_{m=1}^{r} \frac{x_{\cdot m}}{1+uz_m}\right) \Bigg/ \left(\sum_{m=1}^{r} \frac{z_m x_{\cdot m}}{1+uz_m}\right) \\
&= \frac{n}{x_{1\cdot}} \left(\sum_{m=1}^{r} \frac{x_{\cdot m}}{1+uz_m} - x_{1\cdot}\right) \Bigg/ \left(\sum_{m=1}^{r} \frac{z_m x_{\cdot m}}{1+uz_m}\right) \\
&= \frac{n}{x_{1\cdot}} \left(\psi(u)\right) \Bigg/ \left(\sum_{m=1}^{r} \frac{z_m x_{\cdot m}}{1+uz_m}\right).
\end{aligned}$$

The sign of $\Psi_x(u) - u$ is obtained from that of $\psi(u)$.

*iii)* From the second line of Equation (9) and the equality

$$1 - \frac{1}{n} \sum_{m=1}^{r} \frac{\theta^{(k+1)} z_m x_{\cdot m}}{1+\theta^{(k+1)} z_m} = 1 - \frac{1}{n} \sum_{m=1}^{r} \left(x_{\cdot m} - \frac{x_{\cdot m}}{1+\theta^{(k+1)} z_m}\right) = \frac{1}{n} \sum_{m=1}^{r} \frac{x_{\cdot m}}{1+\theta^{(k+1)} z_m},$$

we deduce that the real sequence $\theta^{(k)}$ is given by :

$$\theta^{(k+1)} = \left(\sum_{m=1}^{r} \frac{x_{\cdot m}}{1+\theta^{(k)} z_m}\right) \Bigg/ \left(\sum_{m=1}^{r} \frac{z_m x_{\cdot m}}{1+\theta^{(k)} z_m}\right). \tag{12}$$

which is equivalent to $\theta^{(k+1)} = \Psi_x(\theta^{(k)})$. The function $\Psi_x$ is differentiable and its derivative has the form $\Psi_x'(u) = \alpha_x \times \Psi_{1,x}(u)/\Psi_{2,x}(u)$ where

$$\Psi_{2,x}(u) = \left(\sum_{m=1}^{r} \frac{z_m x_{\cdot m}}{1+uz_m}\right)^2 > 0$$

and

$$\Psi_{1,x}(u) = -\sum_{i=1}^{r} \frac{z_i x_{\cdot i}}{(1+uz_i)^2} \sum_{j=1}^{r} \frac{z_j x_{\cdot j}}{1+uz_j} + \sum_{i=1}^{r} \frac{x_{\cdot i}}{1+uz_i} \sum_{j=1}^{r} \frac{(z_j)^2 x_{\cdot j}}{(1+uz_j)^2}.$$

By swapping indexes $i$ and $j$ in the second right-hand term, we get

$$\Psi_{1,x}(u) = \sum_{i=1}^{r}\sum_{j=1}^{r} \frac{(z_i)^2 x_{\cdot i}x_{\cdot j} - z_i z_j x_{\cdot i}x_{\cdot j}}{(1+uz_i)^2(1+uz_j)}.$$

After removing the zero terms corresponding to $i = j$, we get:

$$\Psi_{1,x}(u) = \sum_{1\leqslant i<j\leqslant r} \frac{(z_i)^2 x_{\cdot i}x_{\cdot j} - z_i z_j x_{\cdot i}x_{\cdot j}}{(1+uz_i)^2(1+uz_j)} + \sum_{1\leqslant j<i\leqslant r} \frac{(z_i)^2 x_{\cdot i}x_{\cdot j} - z_i z_j x_{\cdot i}x_{\cdot j}}{(1+uz_i)^2(1+uz_j)}.$$

By swapping indexes $i$ and $j$ in the second right-hand term, we get

$$\Psi_{1,x}(u) = \sum_{1\leqslant i<j\leqslant r} \frac{(z_i)^2 x_{\cdot i}x_{\cdot j} - z_i z_j x_{\cdot i}x_{\cdot j}}{(1+uz_i)^2(1+uz_j)} + \sum_{1\leqslant i<j\leqslant r} \frac{(z_j)^2 x_{\cdot i}x_{\cdot j} - z_i z_j x_{\cdot i}x_{\cdot j}}{(1+uz_j)^2(1+uz_i)}.$$

It is easy to check that

$$\begin{aligned}\Psi_{1,x}(u) &= \sum_{1\leqslant i<j\leqslant r} \left( \frac{(z_i)^2 x_{\cdot i}x_{\cdot j} - z_i z_j x_{\cdot i}x_{\cdot j}}{(1+uz_i)^2(1+uz_j)} + \frac{(z_j)^2 x_{\cdot i}x_{\cdot j} - z_i z_j x_{\cdot i}x_{\cdot j}}{(1+uz_i)(1+uz_j)^2} \right) \\ &= \sum_{1\leqslant i<j\leqslant r} \left( \frac{x_{\cdot i}x_{\cdot j}}{(1+uz_i)(1+uz_j)} \right) \left( \frac{(z_i)^2 - z_i z_j}{1+uz_i} + \frac{(z_j)^2 - z_i z_j}{1+uz_j} \right).\end{aligned}$$

Since

$$\frac{(z_i)^2 - z_i z_j}{1+uz_i} + \frac{(z_j)^2 - z_i z_j}{1+uz_j} = \frac{(z_i)^2 - 2z_i z_j + (z_j)^2}{(1+uz_i)(1+uz_j)} = \frac{(z_i - z_j)^2}{(1+uz_i)(1+uz_j)} \geqslant 0,$$

we get $\Psi_{1,x}(u) \geqslant 0$. Thus the function $\Psi_x$ is an increasing function.

If $\theta^{(0)} < \theta^*$, then by using the property ii) of Lemma 2, we will have $\theta^{(0)} < \Psi_x(\theta^{(0)}) = \theta^{(1)}$. As $\Psi_x$ is an increasing function, we will then have $\theta^{(1)} = \Psi_x(\theta^{(0)}) < \Psi_x(\theta^{(1)}) = \theta^{(2)}$, $\theta^{(2)} = \Psi_x(\theta^{(1)}) < \Psi_x(\theta^{(2)}) = \theta^{(3)}$ and so on. By a similar reasoning, one can prove by recurrence that if $\theta^{(0)} > \theta^*$ then $\theta^{(k)} > \theta^{(k+1)}$ for all $k = 0, 1, 2, \ldots$.

*iv)* For every $k \geqslant 0$, we have

$$0 < \theta^{(k)} \leqslant \max\left(\theta^{(0)}, \sup_{u>0} \Psi_x(u)\right)$$

where

$$\sup_{u>0} \Psi_x(u) = \lim_{u\to+\infty} \Psi_x(u) = \frac{\alpha_x}{n} \sum_{m=1}^{r} \frac{x_{\cdot m}}{z_m}. \tag{13}$$

The real sequence $(\theta^{(k)})$ is monotonous and bounded. Thus it converges to $\theta^*$ the only fixed point of the function $\Psi_x$ that is also equal to the MLE $\hat{\theta}$ (by Lemma 1). The proof of Lemma 2 is then completed.

Because of the partition of the parameter vector $\beta$ into two sub-parameters $\theta$ and $\phi$, we consider the concentrated (or profile) likelihood function that is also commonly used in maximum likelihood estimation (Monahan, 2011). For a given value of $\hat{\theta}$, the MLE of the sub-parameter $\phi$ is found as a function $\hat{\phi} = g(\hat{\theta}) = (g_1(\hat{\theta}), \ldots, g_r(\hat{\theta}))^T$ where

$$\begin{aligned} g_j(\hat{\theta}) &= \frac{1}{1 - \dfrac{1}{n}\displaystyle\sum_{m=1}^{r} \frac{\hat{\theta} z_m x_{\cdot m}}{1+\hat{\theta} z_m}} \times \frac{x_{\cdot j}}{n(1+\hat{\theta} z_j)}, \quad j = 1, \ldots, r \\ &= \frac{x_{\cdot j}/(1+\hat{\theta} z_j)}{\sum_{m=1}^{r} x_{\cdot m}/(1+\hat{\theta} z_m)}, \quad j = 1, \ldots, r \quad \text{(by expression (11))}. \end{aligned} \tag{14}$$

The likelihood function $\ell(\beta) = \ell(\theta, \phi)$ can be re-written as a function only of $\theta$,

$$\ell_c(\theta) = \ell(\theta, g(\theta))$$

that is called the concentrated (or profile) likelihood.

**Lemma 3.** *The concentrated (or profile) likelihood function is defined up to an additive constant by*

$$\ell_c(\theta) = x_{2\cdot} \log\theta - \sum_{j=1}^{r} x_{\cdot j} \log(1+\theta z_j) \tag{15}$$

*Proof.* The expression (5) is equivalent to

$$\ell(\beta) = \sum_{j=1}^{r} x_{\cdot j} \log(\phi_j) + x_{2\cdot} \log(\theta) - n\log(1+\theta \sum_{m=1}^{r} z_m \phi_m)$$

and one can write

$$\begin{aligned} \ell_c(\theta) = \sum_{j=1}^{r} x_{\cdot j} \log\left(\frac{x_{\cdot j}}{1+\theta z_j}\right) - \sum_{j=1}^{r} x_{\cdot j} \log\left(\sum_{m=1}^{r} \frac{x_{\cdot m}}{1+\theta z_m}\right) + x_{2\cdot} \log(\theta) \\ - n\log\left(\sum_{m=1}^{r} \frac{x_{\cdot m}}{1+\theta z_m} + \sum_{m=1}^{r} \frac{\theta z_m x_{\cdot m}}{1+\theta z_m}\right) + n\log\left(\sum_{m=1}^{r} \frac{x_{\cdot m}}{1+\theta z_m}\right). \end{aligned}$$

This is equivalent to

$$\begin{aligned} \ell_c(\theta) = \sum_{j=1}^{r} x_{\cdot j} \log\left(\frac{x_{\cdot j}}{1+\theta z_j}\right) - n\log\left(\sum_{m=1}^{r} \frac{x_{\cdot m}}{1+\theta z_m}\right) + x_{2\cdot} \log(\theta) \\ - n\log n + n\log\left(\sum_{m=1}^{r} \frac{x_{\cdot m}}{1+\hat{\theta} z_m}\right). \end{aligned}$$

After removing the second and the fifth terms and the constants, we get the equality (15).

## 4. Main results

In this section we prove that the cyclic algorithm (9) satisfies two main properties generally required for ML estimation iterative algorithms. The first one is the convergence of the iterative scheme (9) to the MLE $\hat{\beta} = (\hat{\theta}, \hat{\phi}^T)^T$ from the algorithmic viewpoint. Indeed, there is no guarantee that an iterative algorithm will converge to the MLE. It should be noted that the convergence studied here is different from the consistency that was already studied by Geraldo et al. (2015). The first main result proved in this section is stated by the following theorem.

**Theorem 1.** *For all starting vector $\beta^{(0)} = (\theta^{(0)}, (\phi^{(0)})^T)^T$ where $\theta^{(0)} > 0$ and $\phi^{(0)} \in \mathbb{S}_r$, the sequence $\beta^{(k)} = (\theta^{(k)}, (\phi^{(k)})^T)^T$ generated by the cyclic algorithm (9) converges to the MLE $\hat{\beta} = (\hat{\theta}, \hat{\phi}^T)^T$ of $\beta$.*

*Proof.* From Lemma 2, we can conclude that the sequence $(\theta^{(k)})$ converges to $\hat{\theta}$. Since each $\phi_j^{(k)}$, $j = 1, \ldots, r$, is the image of $\theta^{(k)}$ by the continuous mapping $g_j$ defined by (14), the real sequence $\phi_j^{(k)}$ also has a limit that is $g_j(\hat{\theta}) = \hat{\phi}_j$. Thus, the vector $\beta^{(k)} = (\theta^{(k)}, (\phi^{(k)})^T)^T$ converges to the MLE $\hat{\beta} = (\hat{\theta}, \hat{\phi}^T)^T$. Theorem 1 is thus proved.

The second result that we prove is the ascent property of the cyclic algorithm (9) i.e. the fact that the log-likelihood is increased monotonically by the algorithm. That is given by the following theorem.

**Theorem 2.** *The cyclic algorithm (9) enjoys the ascent property i.e. the sequence $\beta^{(k)} = (\theta^{(k)}, (\phi^{(k)})^T)^T$ generated by the cyclic algorithm satisfies the property*

$$\ell(\beta^{(k+1)}) \geqslant \ell(\beta^{(k)}), \quad k = 0, 1, \ldots \tag{16}$$

*Proof.* The profile log-likelihood $\ell_c(\theta)$ is differentiable for every $\theta > 0$ and its derivative is

$$\ell_c'(\theta) = \frac{x_{2\cdot}}{\theta} - \sum_{j=1}^{r} \frac{x_{\cdot j} z_j}{1 + \theta z_j} = \frac{1}{\theta}\left(x_{2\cdot} - \sum_{j=1}^{r}\left(x_{\cdot j} - \frac{x_{\cdot j}}{1 + \theta z_j}\right)\right) = \frac{1}{\theta}\left(x_{2\cdot} - n + \sum_{j=1}^{r} \frac{x_{\cdot j}}{1 + \theta z_j}\right).$$

Since $x_{1\cdot} + x_{2\cdot} = n$, we have

$$\ell_c'(\theta) = \theta^{-1}\, \psi(\theta)$$

where the function $\psi$ is defined by equation (10) (Lemma 1). Using this lemma, we deduce that

$$\forall \theta \leqslant \theta^*, \quad \ell_c'(\theta) \geqslant 0 \quad \text{and} \quad \forall \theta \geqslant \theta^*, \quad \ell_c'(\theta) \leqslant 0$$

where $\theta^*$ is the MLE of $\theta$ and also the unique root of $\psi$. Hence the function $\ell_c$ is increasing on the interval $]0, \theta^*]$ and decreasing on $[\theta^*, +\infty[$. To finish the proof, we consider the two cases $\theta^{(0)} < \theta^*$ and $\theta^{(0)} > \theta^*$.

- If $\theta^{(0)} < \theta^*$, then we have proved that the sequence $\theta^{(k)}$ is increasing and still belongs to the interval $]0, \theta^*]$. Then $\theta^{(k)} \leqslant \theta^{(k+1)}$ and $\ell_c(\theta^{(k)}) \leqslant \ell_c(\theta^{(k+1)})$ because $\ell_c$ is increasing on $]0, \theta^*]$.

- If $\theta^{(0)} > \theta^*$, then the sequence $\theta^{(k)}$ is decreasing and still belongs to the interval $[\theta^*, +\infty[$. Then $\theta^{(k+1)} \leqslant \theta^{(k)}$ and $\ell_c(\theta^{(k+1)}) \geqslant \ell_c(\theta^{(k)})$ because $\ell_c$ is decreasing on $[\theta^*, +\infty[$.

In all the cases, we have

$$\ell(\beta^{(k+1)}) - \ell(\beta^{(k)}) = \ell_c(\theta^{(k+1)}) - \ell_c(\theta^{(k)}) \geqslant 0$$

and the proof of Theorem 2 is complete.

## 5. Concluding remarks

In this paper, we gave some theoretical convergence theorems for a cyclic algorithm (CA) for the maximum likelihood estimation of the vector parameter of a statistical model for crash data. The vector parameter is partitioned under the form $\beta = (\theta, \phi^T)^T$ where $\theta$ is a positive real number while $\phi$ is a vector belonging to a multidimensional simplex. We proved the global convergence of the CA (i.e. the CA converges to the MLE of $\beta$ from every initial point) and we also proved that it enjoys the ascent property. Future research will include the study of the convergence rate as well as the extension of our results to the statistical model proposed in N'Guessan et al. (2001) which is a generalization of the model studied in this paper.

**Acknowledgment** The author wish to thank the unknown reviewer for the insightful comments and recommendations that helped to improve the quality and the editors for a high professional handling of the publication process.

## References

Cramer, H. (1946). *Mathematical methods of statistics*. Princeton University Press, Princeton.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

Dennis, Jr, J. E. and Schnabel, R. B. (1996). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics.

Geraldo, I. C., N'Guessan, A., and Gneyou, K. E. (2015). A note on the strong consistency of a constrained maximum likelihood estimator used in crash data modelling. *Comptes Rendus Mathematique*, 353(12):1147–1152.

Hauer, E. (2010). Cause, effect and regression in road safety: A case study. *Accident Analysis and Prevention*, 42:1128–1135.

Hunter, D. R. and Lange, K. (2004). A Tutorial on MM Algorithms. *The American Statistician*, 58(1):30–37.

Lange, K. (2010). *Numerical Analysis for Statisticians*. Springer.

Lange, K. (2013). *Optimization*. Springer, New York, 2nd edition.

Lange, K., Chi, E. C., and Zhou, H. (2014). A Brief Survey of Modern Optimization for Statisticians. *International Statistical Review*, 82(1):46–70.

Lord, D. and Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A*, 44:291–305.

Mannering, F. L. and Bhat, C. R. (2014). Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research*, 1:1–22.

McLachlan, G. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. Wiley series in probability and statistics. John Wiley & Sons, Inc., Hoboken, New Jersey, 2nd edition.

Monahan, J. F. (2011). *Numerical Methods of Statistics*. Cambridge University Press, 2nd edition.

N'Guessan, A. (2010). Analytical Existence of solutions to a system of non-linear equations with application. *Journal of Computational and Applied Mathematics*, 234:297–304.

N'Guessan, A., Essai, A., and Langrand, C. (2001). Estimation multidimensionnelle des contrôles et de l'effet moyen d'une mesure de sécurité routière. *Revue de statistique appliquée*, 49(2):85–102.

N'Guessan, A. and Geraldo, I. C. (2015). A cyclic algorithm for maximum likelihood estimation using Schur complement. *Numerical Linear Algebra with Applications*, 22(6):1161–1179.

N'Guessan, A. and Truffier, M. (2008). Impact d'un aménagement de sécurité routière sur la gravité des accidents de la route. *Journal de la Société Française de Statistique*, 149(3):23–41.

Nocedal, J. and Wright, S. J. (2006). *Numerical optimization*. Springer, second edition.

Osborne, M. R. (1992). Fisher's method of scoring. *International Statistical Review / Revue Internationale de Statistique*, 60(1):99–117.

Rios, L. M. and Sahinidis, N. V. (2013). Derivative-free optimization: a review of algorithms and comparison of software implementations. *Journal of Global Optimization*, 56(3):1247–1293.

Shao, J. (2003). *Mathematical Statistics*. Springer, second edition.

Zhou, H. and Lange, K. (2010). MM algorithms for some discrete multivariate distributions. *Journal of Computational and Graphical Statistics*, 19(3):645–665.