HAL open science

Modèle de durée pour données multivoie

Alfred Baroulier, Mehdi Douch, Paul Messinesi, Laurent Le Brusquet, Gisela Lechuga, Arthur Tenenhaus

To cite this version:

Alfred Baroulier, Mehdi Douch, Paul Messinesi, Laurent Le Brusquet, Gisela Lechuga, et al.. Modèle de durée pour données multivoie. 49émes Journées de Statistique, May 2017, Avignon, France. hal-01474989

HAL Id: hal-01474989

https://hal.science/hal-01474989v1

Submitted on 23 Feb 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

MODELE DE DUREE POUR DONNEES MULTIVOIE

Alfred Baroulier[1], Mehdi Douch[1], Paul Messinesi[1], Laurent Le Brusquet[2(*)], Gisela Lechuga[(2)] & Arthur Tenenhaus[2,3(*)]

[1]*CentraleSupélec, département Signal et Statistiques, 3 rue Joliot Curie, 91192 Gif-sur-Yvette, prenom.nom@supelec.fr*
[2]*Laboratoire des Signaux et Systèmes, CentraleSupélec – CNRS – Univ. Paris-Sud, Université Paris-Saclay, 3 rue Joliot Curie, 91192 Gif-sur-Yvette, prenom.nom@centralesupelec.fr*
[3]*Bioinformatics/Biostatistics Platform IHU-A-ICM, Brain and Spine Institute, 47-83 bd de l'hôpital, 75013 Paris*

Résumé. Cet article étend le modèle de Cox au cas des données multivoie, c'est-à-dire aux données où chaque individu est décrit par plusieurs modalités de la même covariable. Imposer aux coefficients de régression une structure tensorielle identique à celle des données permet d'une part de restreindre le nombre de coefficients à estimer et donc la complexité calculatoire et d'autre part d'éviter le phénomène de sur-apprentissage. Cette nouvelle approche est évaluée et validée sur données simulées.

Mots-clés. Modèle de survie, modèle de Cox, données multivoie.

Abstract. . In this paper, we present a multiway extension of Cox proportional hazards model (Multiway Cox). Multi-way data refers to the case where several modalities of the same variable have been collected for each individual. The approach consists in imposing a tensor-like structure to the weight coefficients echoing the structure of the multi-way data set. This approach reduces the number of coefficients to estimate and limits overfitting issues. The proposed procedure is evaluated and validated on simulated data.

Keywords. Survival Analysis, Cox proportional hazards model, Multiway data set.

1 Introduction

Ce papier présente une extension du modèle de Cox au cas des données multivoie. La notion de données « multivoie » concerne les données explicatives qui ne sont pas représentées par une matrice mais par un tenseur.

Dans le contexte des modèles de durée, les données sont des instants de défaillance dont on cherche à expliquer la loi en fonction de covariables observées selon plusieurs modalités. Ce papier s'intéresse au cas des tenseurs d'ordre 3. Soit $\{\underline{\mathbf{Z}}_{ijk}\}_{1\leq i\leq n,1\leq j\leq J,1\leq k\leq K}$ un tenseur d'ordre 3 de dimension $n \times J \times K$ où n désigne le nombre d'individus, J le nombre de covariables et K le nombre de modalités. Pour chaque individu, au total JK variables sont donc observées (voir figure 1).

Une méthode simple pour traiter le cas de données multivoie consiste à « déplier » le tenseur en une matrice de taille $n \times JK$ que l'on note $\{\mathbf{Z}_{i,l}\}_{1\leq i\leq n,1\leq l\leq JK}$. On retrouve alors un problème adapté aux méthodes statistiques classiques : la régression de Cox consiste à rechercher un vecteur $\boldsymbol{\beta}$ de longueur JK représentant l'influence de chaque variable.

(*) co-auteurs à contacter pour toute correspondance.

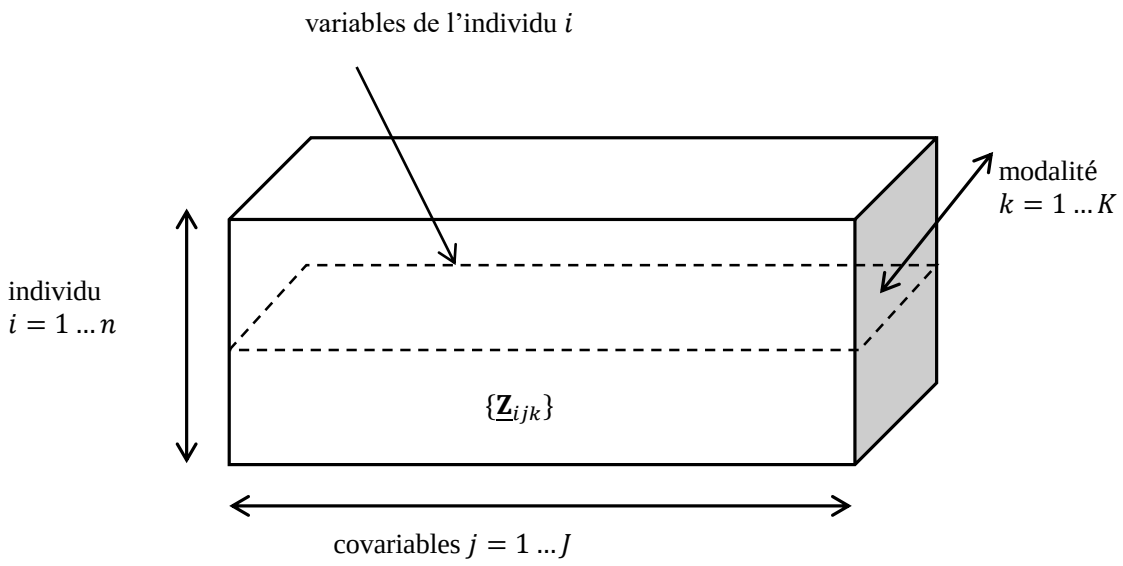

Figure 1 : Données tensorielles. Chaque individu est représenté par J covariables observées selon K modalités.

Lorsque les nombres de covariables (J) et/ou de modalités (K) sont grands, une telle approche peut conduire à la fois à des temps de calcul importants et à des problèmes de sur-apprentissage.

Afin de limiter ces défauts, on adopte ici l'hypothèse développée par Lechuga etal (2015) qui consiste à imposer une structure tensorielle au vecteur $\boldsymbol{\beta}$ cherché via le produit de Kronecker :

$$\boldsymbol{\beta} = \boldsymbol{\beta}^K \otimes \boldsymbol{\beta}^J$$

où $\boldsymbol{\beta}^K$ et $\boldsymbol{\beta}^J$ sont des vecteurs de taille K et J, avec $\boldsymbol{\beta}^K$ vecteur traduisant l'influence des modalités, et $\boldsymbol{\beta}^J$ vecteur traduisant l'influence des covariables. Le nombre de coefficients à estimer est ainsi restreint à $K + J$.

Après avoir brièvement présenté le modèle de Cox dans le contexte classique de données matricielles (section 2), la section 3 du papier présente une approche par directions alternées pour optimiser la vraisemblance partielle de Cox par rapport aux $K + J$ coefficients. Un exemple académique avec des données simulées est présenté en section 4 : la version standard de la régression de Cox (analyse sur données dépliées) et la version multivoie sont comparées en termes de temps de calcul et de qualité d'estimation du vecteur $\boldsymbol{\beta}$.

2 Modèle de Cox standard

Le modèle de Cox est ici brièvement exposé sur la base des travaux de Cox (1972). Le modèle de Cox étudie la probabilité de défaillance d'un individu en fonction de ses caractéristiques à différents instants t. Pour un individu i on note T_i sa date de défaillance.

On introduit la notion de censure C_i d'une donnée i si avant la fin de l'expérience on n'observe plus l'individu sans qu'il y ait eu de défaillance, ou si à la fin de l'expérience la défaillance n'a pas été observée. Une indicatrice indique alors si la donnée i est censurée :

$$\delta_i = \mathbb{1}_{\{T_i \leq C_i\}}.$$

Dans la théorie des modèles de survie on s'intéresse au taux de défaillance $\alpha(t)$. Il s'agit de la probabilité de défaillance à l'instant t sachant que rien ne s'est produit auparavant. Pour un individu dont la date de défaillance suit une loi de probabilité de densité f, le taux de défaillance est défini par $\alpha(t) = \frac{f(t)}{S(t)}$, avec $S(t) = 1 - F(t)$.

Le modèle de Cox se définit alors ainsi :

$$\alpha(\, t \,|\mathbf{Z}_i\,) = \alpha_0(t)\exp(\mathbf{Z}_i^\top \boldsymbol{\beta})$$

avec $\alpha_0(t)$ le risque de base en l'absence d'effet des covariables, $\mathbf{Z}_i$ le vecteur des variables pour l'individu i, et $\boldsymbol{\beta}$ le vecteur des coefficients traduisant l'influence de chaque covariable. On considère que le modèle de Cox est à risques proportionnels : la dépendance du taux de défaillance par rapport aux variables explicatives s'explique par le terme multiplicatif $\exp(\mathbf{Z}_i^T \boldsymbol{\beta})$ indépendant du temps.

Le recours à la vraisemblance partielle $\mathcal{L}_p(\boldsymbol{\beta})$ (plutôt que l'utilisation de la fonction de vraisemblance classique) permet de s'affranchir de la connaissance de $\alpha_0(t)$ dans le calcul des coefficients $\boldsymbol{\beta}$ (voir Cox (1975)) :

$$\mathcal{L}_p(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left\{ \frac{\exp(\mathbf{Z}_i^\top \boldsymbol{\beta})}{\sum_{i'=1}^{n} Y_{i'}(T_i)\exp(\mathbf{Z}_i^\top \boldsymbol{\beta})} \right\}^{\delta_i}$$

où $Y_i(t) = \begin{cases} 1 \text{ si l'individu i est encore à risque à } t\ (T_i \geq t) \\ 0 \text{ sinon.} \end{cases}$

La maximisation par rapport à $\boldsymbol{\beta}$ de la vraisemblance $\mathcal{L}_p(\boldsymbol{\beta})$ est généralement résolue à l'aide d'algorithmes de type Newton-Raphson consistant à annuler le vecteur de score partiel $\mathbf{U}_p(\boldsymbol{\beta})$:

$$\mathbf{U}_p(\boldsymbol{\beta}) = \frac{\partial \log \mathcal{L}_p\,(\boldsymbol{\beta})}{\partial \beta} = \sum_{i=1}^{n} \delta_i \left(\mathbf{Z}_i - \frac{\sum_{i'=1}^{n} Y_{i'}(T_i)\,\mathbf{Z}_{i'}\exp(\mathbf{Z}_{i'}^\top \boldsymbol{\beta})}{\sum_{i'=1}^{n} Y_{i'}(T_i)\exp(\mathbf{Z}_{i'}^\top \boldsymbol{\beta})} \right)$$

Les algorithmes de type Newton-Raphson sont des algorithmes itératifs :

$$\widehat{\boldsymbol{\beta}}^{(q+1)} = \widehat{\boldsymbol{\beta}}^{(q)} + \mathbf{I}_n^{-1}\left(\widehat{\boldsymbol{\beta}}^{(q)}\right)\mathbf{U}_p\left(\widehat{\boldsymbol{\beta}}^{(q)}\right)$$

où $\mathbf{I}_n(\boldsymbol{\beta})$ est la matrice des dérivées secondes :

$$\mathbf{I}_n(\boldsymbol{\beta}) = -\ \frac{\partial^2 \log \mathcal{L}_p\,(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}$$

En pratique, cette matrice est soit calculée analytiquement (lorsque le nombre de variables est faible) soit approximée.
L'algorithme 1 résume les grandes lignes des algorithmes de régression de Cox de type Newton-Raphson pour des données standard, c'est-à-dire matricielles.

Algorithme 1 : Régression de Cox standard : $\widehat{\boldsymbol{\beta}}^{(q)} = COX(\mathbf{Z}, T, t_{censure})$

Require: $\epsilon > 0,\ \widehat{\boldsymbol{\beta}}^{(0)},\ T,\ \mathbf{Z},\ t_{censure}$

$q \leftarrow 0$

repeat:

Calcul (ou approximation) de $\mathbf{U}_p\left(\widehat{\boldsymbol{\beta}}^{(q)}\right)$ et de $\mathbf{I}_n\left(\widehat{\boldsymbol{\beta}}^{(q)}\right)$ à partir de $T,\ \mathbf{Z},\ t_{censure}$

$\widehat{\boldsymbol{\beta}}^{(q+1)} = \widehat{\boldsymbol{\beta}}^{(q)} + \mathbf{I}_n^{-1}\left(\widehat{\boldsymbol{\beta}}^{(q)}\right)\mathbf{U}_p\left(\widehat{\boldsymbol{\beta}}^{(q)}\right)$

$q \leftarrow q + 1$

until $\left\| \widehat{\boldsymbol{\beta}}^{(q)} - \widehat{\boldsymbol{\beta}}^{(q-1)} \right\| < \epsilon$

return $\widehat{\boldsymbol{\beta}}^{(q)}$

---

3 Modèle de Cox multivoie

La version multivoie proposée consiste à maximiser la log-vraisemblance partielle en imposant la contrainte $\boldsymbol{\beta} = \boldsymbol{\beta}^K \otimes \boldsymbol{\beta}^J$, c'est-à-dire $\boldsymbol{\beta}_{jk} = \boldsymbol{\beta}_k^K \boldsymbol{\beta}_j^J$ avec $\boldsymbol{\beta}_k^K$ (resp. $\boldsymbol{\beta}_j^J$) le $k^{\text{ème}}$ (resp. $j^{\text{ème}}$) élément du vecteur $\boldsymbol{\beta}^K$ (resp $\boldsymbol{\beta}^J$), afin de (i) réduire le nombre de variables de JK à $K+J$ (ii) faciliter l'interprétation des coefficients du vecteur $\boldsymbol{\beta}$. En effet, avec la structure de Kronecker, l'interprétation de l'influence des covariables et des modalités peut se faire séparément : $\boldsymbol{\beta}^J$ pondère l'influence des covariables alors que $\boldsymbol{\beta}^K$ pondère celle des modalités.

La contrainte $\|\boldsymbol{\beta}^K\| = 1$ est également imposée afin de rendre la décomposition $\boldsymbol{\beta}^K \otimes \boldsymbol{\beta}^J$ unique (au signe près).

Comme précédemment, on maximise la vraisemblance partielle par rapport à $(\boldsymbol{\beta}^K, \boldsymbol{\beta}^J)$:

$$\mathcal{L}_p(\boldsymbol{\beta}^K, \boldsymbol{\beta}^J) = \prod_{i=1}^{n} \left\{ \frac{\exp(\mathbf{Z}_i^\top (\boldsymbol{\beta}^K \otimes \boldsymbol{\beta}^J))}{\sum_{i'=1}^{n} Y_{i'}(T_i) \exp\left(\mathbf{Z}_{i'}^\top (\boldsymbol{\beta}^K \otimes \boldsymbol{\beta}^J)\right)} \right\}^{\delta_i}$$

Le vecteur $\mathbf{Z}_i$ correspond aux JK variables explicatives du $i^{\text{ème}}$ individu concaténées dans un vecteur (c'est-à-dire la $i^{\text{ème}}$ ligne de la matrice dépliée).

Le produit scalaire $\mathbf{Z}_i^\top (\boldsymbol{\beta}^K \otimes \boldsymbol{\beta}^J)$ peut également s'écrire à partir du tenseur initial :

$$\begin{aligned} \mathbf{Z}_i^\top (\boldsymbol{\beta}^K \otimes \boldsymbol{\beta}^J) &= \left(\sum_{k=1}^{K} \underline{\mathbf{Z}}_{i,.,k} \boldsymbol{\beta}_k^K \right)^\top \boldsymbol{\beta}^J \\ &= \left(\sum_{j=1}^{j} \underline{\mathbf{Z}}_{i,j,.} \boldsymbol{\beta}_j^J \right)^\top \boldsymbol{\beta}^K \end{aligned}$$

où $\underline{\mathbf{Z}}_{i,.,k}$ est le vecteur $J \times 1$ des variables de l'individu i pour la modalité k. De même $\underline{\mathbf{Z}}_{i,j,.}$ est le vecteur $K \times 1$ des variables de l'individu i pour la covariable j.

Soit $\mathbf{Z}^K$ la matrice de taille $n \times J$ constituée, pour $i = 1, \dots, n$ des n lignes $\left(\sum_{k=1}^{K} \underline{\mathbf{Z}}_{i,.,k} \boldsymbol{\beta}_k^K\right)^\top$. Optimiser $\mathcal{L}_p(\boldsymbol{\beta}^K, \boldsymbol{\beta}^J)$ par rapport à $\boldsymbol{\beta}^J$ revient donc à réaliser une régression de Cox sur les données modifiées $(\mathbf{Z}^K, T)$. De même pour l'optimisation par rapport à $\boldsymbol{\beta}^K$: elle est réalisée via une régression de Cox opérée sur une matrice $\mathbf{Z}^J$ de taille $n \times K$.

L'optimisation par rapport à chacune des variables se ramenant à des analyses de Cox sur des données standard de taille réduite (le nombre de variables est soit J, soit K), un algorithme de directions alternées est utilisé pour optimiser $\mathcal{L}_p(\boldsymbol{\beta}^K, \boldsymbol{\beta}^J)$ conjointement par rapport aux 2 vecteurs (voir Algorithme 2).

4 Exemple illustratif

La version multivoie de la régression de Cox a été testée sur des données simulées pour des problèmes de tailles différentes afin d'évaluer le comportement en termes de temps de calcul et en termes d'erreurs d'estimation.

Les covariables et les modalités jouant un rôle symétrique, une seule valeur a été considérée pour le nombre de modalités ($K = 10$) alors que plusieurs valeurs du nombre de covariables ont été utilisés (J allant de 2 à 50).

Algorithme 2 : Régression de Cox multivoie, algorithme des directions alternées
Require: $\epsilon > 0,\ \boldsymbol{\beta}^{K(0)},\ T,\ \underline{\mathbf{Z}},\ t_{censure}$

$q \leftarrow 0$

repeat:

Construction de $\mathbf{Z}^K$: $\mathbf{Z}_i^K = \left(\sum_{k=1}^{K} \underline{\mathbf{Z}}_{i,.,k} \boldsymbol{\beta}_k^{K(q)}\right)^\top$

$\boldsymbol{\beta}^{J(q)} \leftarrow COX(\mathbf{Z}^K, T, t_{censure})$

Construction de $\mathbf{Z}^J$: $\mathbf{Z}_i^J = \left(\sum_{j=1}^{J} \underline{\mathbf{Z}}_{i,j,.} \boldsymbol{\beta}_j^{J(q)}\right)^\top$

$\boldsymbol{\beta}^{K(q+1)} \leftarrow COX(\mathbf{Z}^J, T, t_{censure})$

$\boldsymbol{\beta}^{K(q+1)} \leftarrow \frac{\boldsymbol{\beta}^{K(q+1)}}{\|\boldsymbol{\beta}^{K(q+1)}\|}$

$q \leftarrow q + 1$

until $\|\boldsymbol{\beta}^{K(q)} - \boldsymbol{\beta}^{K(q-1)}\| < \epsilon$ et $\|\boldsymbol{\beta}^{J(q)} - \boldsymbol{\beta}^{J(q-1)}\| < \epsilon$

return $\left(\boldsymbol{\beta}^{K(q)}, \boldsymbol{\beta}^{J(q)}, \widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{K(q)} \otimes \boldsymbol{\beta}^{J(q)}\right)$

Pour chaque valeur de J, 100 jeux de données ont été générés aléatoirement de la manière suivante :

(i) $\boldsymbol{\beta}^K$ et $\boldsymbol{\beta}^J$ tirés aléatoirement selon la loi gaussienne $N(0,1)$ (les éléments des vecteurs sont générés indépendamment les uns des autres). Le vecteur inconnu à identifier est donc : $\boldsymbol{\beta} = \boldsymbol{\beta}^K \otimes \boldsymbol{\beta}^J$.

(ii) $n = 1000$ individus ont été simulés. Chaque individu est représenté par J covariables observées selon K modalités et par un instant de défaillance T_i. Les JK variables explicatives du vecteur $\mathbf{Z}_i$ ont été générées aléatoirement indépendamment les unes des autres selon la loi $N(0,1)$. L'instant T_i a été généré selon une loi de Weibull $W(p, \lambda)$ correspondant au taux de défaillance suivant :

$$\alpha(t) = -\frac{f(t)}{1-F(t)} = \frac{\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k}}{e^{-\left(\frac{x}{\lambda}\right)^k}} = \frac{2x}{\lambda^2} = 2xe^{-2\ln(\lambda)}$$

Le paramètre p a été fixé à 2 alors que le paramètre λ dépend des JK variables explicatives :

$$\lambda = \exp\left(\mathbf{Z}_i^\top (\boldsymbol{\beta}^K \otimes \boldsymbol{\beta}^J)\right)$$

La version multivoie de la régression de Cox est comparée à la version standard appliquée à la matrice dépliée de taille $n \times (JK)$. Les temps de calcul ainsi que l'erreur d'estimation du vecteur $\boldsymbol{\beta}$ sont donnés Figure 2.

Le graphique de gauche montre que la version multivoie permet de réduire significativement les erreurs d'estimation. Ce résultat s'explique aisément par le fait qu'imposer une contrainte de Kronecker à $\boldsymbol{\beta}$ revient à injecter un fort a priori. Cet a priori est particulièrement utile, en particulier lorsque le nombre de variables devient grand et que la méthode standard est sujette au phénomène de sur-apprentissage. Ces bons résultats montrent ainsi que l'algorithme des directions alternées converge vers une solution de bonne qualité.

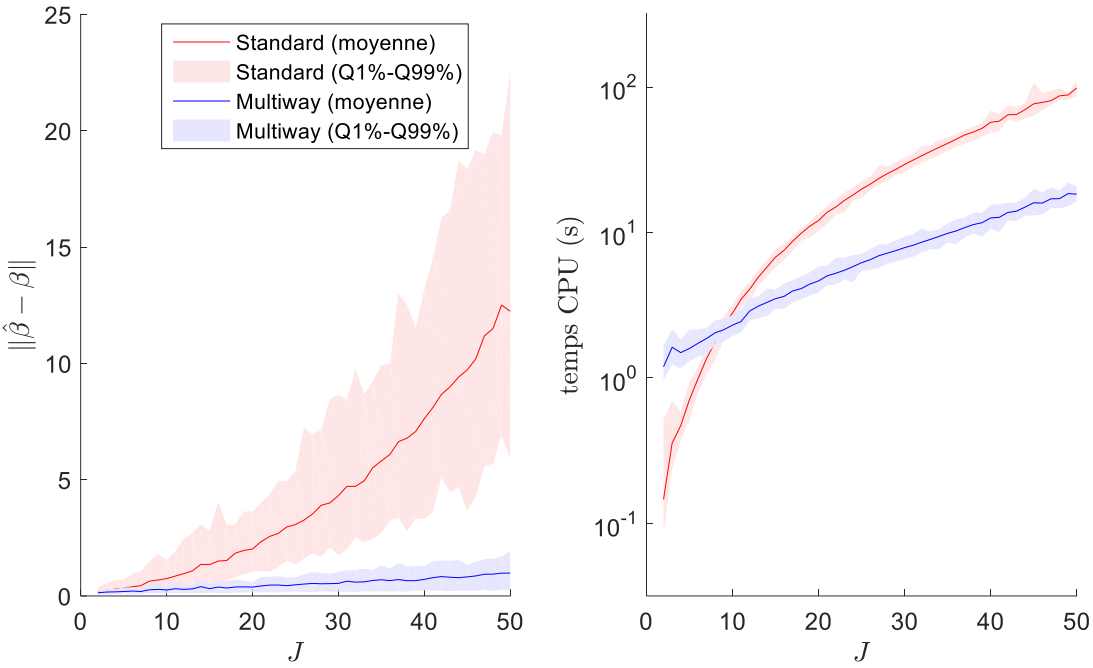

Figure 2 : Evolution des performances en fonction du nombre J de covariables. Pour chaque figure, les traits pleins représentent la moyenne obtenue à partir des 100 simulations, les zones de couleur les intervalles entre le quantile à 1% et le quantile à 99%. A gauche les erreurs d'estimation du vecteur $\boldsymbol{\beta}$, à droite le temps de calcul.

Le graphique de droite illustre la vitesse de convergence des deux versions. Le temps de calcul augmente plus vite avec la taille du problème pour la version standard de la régression de Cox que pour la version multivoie. Ainsi, dès que le nombre J de covariables dépasse quelques unités, la version multivoie se comporte mieux que la version standard. Ce résultat s'explique par le fait que la version multivoie repose sur un algorithme itératif faisant appel à des régressions de Cox standard et que le nombre d'itérations varie peu en fonction de la taille du problème.

Cet exemple académique montre ainsi que la méthode des directions alternées, bien que réputée converger lentement, conduit ici à de bons résultats. En effet, imposer une structure de Kronecker conduit ici à résoudre, à chaque itération de l'algorithme des directions alternées, des problèmes de taille restreinte et faciles à résoudre.

A noter que la version multivoie permet une interprétation séparée de l'influence des covariables et des modalités. Cet aspect, peu exploité sur les données académiques ici utilisées, sera particulièrement intéressant pour des problèmes réels, en particulier si le nombre de covariables et de modalités sont importants.

Bibliographie

[1] Lechuga G., Le Brusquet L., Perlbarg V., Puybasset L., Galanaud D, Tenenhaus A. (2015), Proceedings in Mathematics and Statistics, chapter Discriminant Analysis for Multiway Data. Springer Verlag.
[2] Cox D. R. (1972), Regression Models and Life-Tables, Journal of the Royal Statistical Society, Series B (Methodological), vol. 34, no. 2, p. 187-220.
[3] Cox D. R. (1975), Partial Likelihood, *Biometrika*, vol. 62, p. 269-276.