Prix Norbert Marx 2019

présenté le 05/06/2019 par

Simon Bussy

pour l'article

C-mix: a high dimensional mixture model for censored durations, with applications to genetic data

Statistical Methods in Medical Research, 2017

S. Bussy [1], A. Guilloux [2], S. Gaïffas [1,3], A.S. Jannot [4,5]

[1] LPSM, UMR 8001, Sorbonne Université, Paris, France.
[2] LAMME, Université Evry, CNRS, Université Paris-Saclay, Paris, France.
[3] CMAP, UMR 7641, École Polytechnique CNRS, Paris, France.
[4] APHP, Département d'Informatique Biomédicale et de Santé Publique, HEGP, Paris, France.
[5] INSERM, UMRS 1138, Eq22, Centre de Recherche des Cordeliers, Université Paris Descartes, Paris, France.

Le modèle

- Contexte de l'analyse de survie

$$Y = \min(T, C) \quad \text{et} \quad \Delta = \mathbb{1}_{\{T \leq C\}}$$

Le modèle

- Contexte de l'analyse de survie

$$Y = \min(T, C) \quad \text{et} \quad \Delta = \mathbb{1}_{\{T \leq C\}}$$

- Variable latente $Z \in \{0, \ldots, K-1\}$

Le modèle

- Contexte de l'analyse de survie

$$Y = \min(T, C) \quad \text{et} \quad \Delta = \mathbb{1}_{\{T \leq C\}}$$

- Variable latente $Z \in \{0, \ldots, K-1\}$
- Modèle de mélange $f(t|X = x) = \sum_{k=0}^{K-1} \pi_{\beta_k}(x) f_k(t; \alpha_k)$

Le modèle

- Contexte de l'analyse de survie

$$Y = \min(T, C) \quad \text{et} \quad \Delta = \mathbb{1}_{\{T \leq C\}}$$

- Variable latente $Z \in \{0, \dots, K-1\}$
- Modèle de mélange $f(t|X = x) = \sum_{k=0}^{K-1} \pi_{\beta_k}(x) f_k(t; \alpha_k)$
- $\pi_{\beta_k}(x) = \mathbb{P}[Z = k | X = x] = \frac{e^{x^\top \beta_k}}{\sum_{k=0}^{K-1} e^{x^\top \beta_k}}$ (softmax)

Le modèle

- Contexte de l'analyse de survie

$$Y = \min(T, C) \quad \text{et} \quad \Delta = \mathbb{1}_{\{T \leq C\}}$$

- Variable latente $Z \in \{0, \ldots, K-1\}$
- Modèle de mélange $f(t|X=x) = \sum_{k=0}^{K-1} \pi_{\beta_k}(x) f_k(t; \alpha_k)$
- $\pi_{\beta_k}(x) = \mathbb{P}[Z = k | X = x] = \frac{e^{x^\top \beta_k}}{\sum_{k=0}^{K-1} e^{x^\top \beta_k}}$ (softmax)
- Hypothèses : (i) $T|Z, X \perp\!\!\!\perp C|Z, X$ [2] (ii) $C \perp\!\!\!\perp Z$ [3]

Le modèle

- Contexte de l'analyse de survie

$$Y = \min(T, C) \quad \text{et} \quad \Delta = \mathbb{1}_{\{T \leq C\}}$$

- Variable latente $Z \in \{0, \dots, K-1\}$
- Modèle de mélange $f(t|X=x) = \sum_{k=0}^{K-1} \pi_{\beta_k}(x) f_k(t; \alpha_k)$
- $\pi_{\beta_k}(x) = \mathbb{P}[Z=k|X=x] = \frac{e^{x^\top \beta_k}}{\sum_{k=0}^{K-1} e^{x^\top \beta_k}}$ (softmax)
- Hypothèses : (i) $T|Z,X \perp\!\!\!\perp C|Z,X$ [2] (ii) $C \perp\!\!\!\perp Z$ [3]
- Échantillon *i.i.d.* $(x_1, y_1, \delta_1), \dots, (x_n, y_n, \delta_n) \in \mathbb{R}^d \times \mathbb{R}_+ \times \{0, 1\}$

Le modèle

- Contexte de l'analyse de survie

$$Y = \min(T, C) \quad \text{et} \quad \Delta = \mathbb{1}_{\{T \leq C\}}$$

- Variable latente $Z \in \{0, \dots, K-1\}$
- Modèle de mélange $f(t|X=x) = \sum_{k=0}^{K-1} \pi_{\beta_k}(x) f_k(t; \alpha_k)$
- $\pi_{\beta_k}(x) = \mathbb{P}[Z=k|X=x] = \frac{e^{x^\top \beta_k}}{\sum_{k=0}^{K-1} e^{x^\top \beta_k}}$ (softmax)
- Hypothèses : (i) $T|Z,X \perp\!\!\!\perp C|Z,X$ [2] (ii) $C \perp\!\!\!\perp Z$ [3]
- Échantillon *i.i.d.* $(x_1, y_1, \delta_1), \dots, (x_n, y_n, \delta_n) \in \mathbb{R}^d \times \mathbb{R}_+ \times \{0,1\}$
- $\theta = (\alpha_0, \dots, \alpha_K, \beta_0, \dots, \beta_K)^\top$, log-vraissemblance du C-mix

$$\ell_n(\theta) = n^{-1} \sum_{i=1}^{n} \Big\{ \delta_i \log \Big[\bar{G}(y_i^-) \sum_{k=0}^{K-1} \pi_{\beta_k}(x_i) f_k(y_i; \alpha_k) \Big]$$

$$+ (1-\delta_i) \log \Big[g(y_i) \sum_{k=0}^{K-1} \pi_{\beta_k}(x_i) \bar{F}_k(y_i^-; \alpha_k) \Big] \Big\}$$

L'algorithme QNEM

- Objectif pénalisé

$$\ell_n^{\text{pen}}(\theta) = -\ell_n(\theta) + \sum_{k=0}^{K-1} \gamma_k \big((1-\eta)\|\beta_k\|_1 + \frac{\eta}{2}\|\beta_k\|_2^2\big) \quad (1)$$

L'algorithme QNEM

- Objectif pénalisé

$$\ell_n^{\text{pen}}(\theta) = -\ell_n(\theta) + \sum_{k=0}^{K-1} \gamma_k \big((1-\eta) \|\beta_k\|_1 + \frac{\eta}{2} \|\beta_k\|_2^2 \big) \qquad (1)$$

- $\ell_n^{\text{comp}}(\theta)$ log-vraissemblance complétée (négative)

$$-n^{-1} \sum_{i=1}^{n} \Big\{ \delta_i \Big[\sum_{k=0}^{K-1} \mathbb{1}_{\{z_i=k\}} \big(\log \pi_{\beta_k}(x_i) + \log f_k(y_i; \alpha_k) \big) + \log \bar{G}(y_i^-) \Big]$$

$$+ (1-\delta_i) \Big[\sum_{k=0}^{K-1} \mathbb{1}_{\{z_i=k\}} \big(\log \pi_{\beta_k}(x_i) + \log \bar{F}_k(y_i^-; \alpha_k) \big) + \log g(y_i) \Big] \Big\}$$

L'algorithme QNEM

- Objectif pénalisé

$$\ell_n^{\mathrm{pen}}(\theta) = -\ell_n(\theta) + \sum_{k=0}^{K-1} \gamma_k \big((1-\eta)\|\beta_k\|_1 + \frac{\eta}{2}\|\beta_k\|_2^2\big) \qquad (1)$$

- $\ell_n^{\mathrm{comp}}(\theta)$ log-vraissemblance complétée (négative)

$$-n^{-1}\sum_{i=1}^{n}\Big\{\delta_i\Big[\sum_{k=0}^{K-1}\mathbb{1}_{\{z_i=k\}}\Big(\log \pi_{\beta_k}(x_i) + \log f_k(y_i;\alpha_k)\Big) + \log \bar{G}(y_i^-)\Big]$$
$$+(1-\delta_i)\Big[\sum_{k=0}^{K-1}\mathbb{1}_{\{z_i=k\}}\Big(\log \pi_{\beta_k}(x_i) + \log \bar{F}_k(y_i^-;\alpha_k)\Big) + \log g(y_i)\Big]\Big\}$$

- Étape E: $Q_n(\theta, \theta^{(l)}) = \mathbb{E}_{\theta^{(l)}}[\ell_n^{\mathrm{comp}}(\theta)|\boldsymbol{y}, \boldsymbol{\delta}]$

$$q_{i,k}^{(l)} = \mathbb{E}_{\theta^{(l)}}[\mathbb{1}_{\{z_i=k\}}|y_i,\delta_i] = \mathbb{P}_{\theta^{(l)}}[z_i = k|y_i,\delta_i] = \frac{\Lambda_{k,i}^{(l)}}{\sum_{r=0}^{K-1}\Lambda_{r,i}^{(l)}}$$

$$\Lambda_{k,i}^{(l)} = \Big[f_k(y_i;\alpha_k^{(l)})\bar{G}(y_i^-)\Big]^{\delta_i}\Big[g(y_i)\bar{F}_k(y_i^-;\alpha_k^{(l)})\Big]^{1-\delta_i}\pi_{\beta_k^{(l)}}(x_i)$$

L'algorithme QNEM

- Étape M: problème d'optimisation convexe...

$$\text{minimiser} \quad R_{n,k}^{(l)}(\beta_k) + \gamma_k\big((1-\eta)\|\beta_k\|_1 + \frac{\eta}{2}\|\beta_k\|_2^2\big), \qquad (2)$$

avec $R_{n,k}^{(l)}(\beta_k) = -n^{-1}\sum_{i=1}^n q_{i,k}^{(l)} \log \pi_{\beta_k}(x_i)$

L'algorithme QNEM

- Étape M: problème d'optimisation convexe...

$$\text{minimiser} \quad R_{n,k}^{(l)}(\beta_k) + \gamma_k\big((1-\eta)\|\beta_k\|_1 + \frac{\eta}{2}\|\beta_k\|_2^2\big), \qquad (2)$$

avec $R_{n,k}^{(l)}(\beta_k) = -n^{-1}\sum_{i=1}^n q_{i,k}^{(l)} \log \pi_{\beta_k}(x_i)$

- ...mais non différentiable! On réécrit alors (2) comme :

$$\text{minimiser } R_{n,k}^{(l)}(\beta_k^+ - \beta_k^-) + \gamma_k(1-\eta)\sum_{j=1}^d(\beta_{k,j}^+ + \beta_{k,j}^-) + \gamma_k\frac{\eta}{2}\|\beta_k^+ - \beta_k^-\|_2^2$$

$$\text{tel que } \beta_{k,j}^+ \geq 0 \text{ et } \beta_{k,j}^- \geq 0 \text{ pour tout } j \in \{1,\dots,d\}$$

L'algorithme QNEM

- Étape M: problème d'optimisation convexe...

$$\text{minimiser} \quad R^{(l)}_{n,k}(\beta_k) + \gamma_k\big((1-\eta)\|\beta_k\|_1 + \frac{\eta}{2}\|\beta_k\|_2^2\big), \qquad (2)$$

avec $R^{(l)}_{n,k}(\beta_k) = -n^{-1}\sum_{i=1}^n q^{(l)}_{i,k} \log \pi_{\beta_k}(x_i)$

- ...mais non différentiable! On réécrit alors (2) comme :

$$\text{minimiser } R^{(l)}_{n,k}(\beta_k^+ - \beta_k^-) + \gamma_k(1-\eta)\sum_{j=1}^d(\beta^+_{k,j} + \beta^-_{k,j}) + \gamma_k\frac{\eta}{2}\|\beta_k^+ - \beta_k^-\|_2^2$$

tel que $\beta^+_{k,j} \geq 0$ et $\beta^-_{k,j} \geq 0$ pour tout $j \in \{1, \dots, d\}$

- Solveur L-BFGS-B, requiert le gradient qui s'écrit

$$\frac{\partial R^{(l)}_{n,k}(\beta_k)}{\partial \beta_k} = -n^{-1}\sum_{i=1}^n q^{(l)}_{i,k}\big(1 - \pi_{\beta_k}(x_i)\big)x_i$$

L'algorithme QNEM

- Étape M: problème d'optimisation convexe...

$$\text{minimiser} \quad R_{n,k}^{(l)}(\beta_k) + \gamma_k\big((1-\eta)\|\beta_k\|_1 + \frac{\eta}{2}\|\beta_k\|_2^2\big), \qquad (2)$$

avec $R_{n,k}^{(l)}(\beta_k) = -n^{-1}\sum_{i=1}^n q_{i,k}^{(l)} \log \pi_{\beta_k}(x_i)$

- ...mais non différentiable! On réécrit alors (2) comme :

$$\text{minimiser } R_{n,k}^{(l)}(\beta_k^+ - \beta_k^-) + \gamma_k(1-\eta)\sum_{j=1}^{d}(\beta_{k,j}^+ + \beta_{k,j}^-) + \gamma_k\frac{\eta}{2}\|\beta_k^+ - \beta_k^-\|_2^2$$

$$\text{tel que } \beta_{k,j}^+ \geq 0 \text{ et } \beta_{k,j}^- \geq 0 \text{ pour tout } j \in \{1,\dots,d\}$$

- Solveur L-BFGS-B, requiert le gradient qui s'écrit

$$\frac{\partial R_{n,k}^{(l)}(\beta_k)}{\partial \beta_k} = -n^{-1}\sum_{i=1}^n q_{i,k}^{(l)}\big(1 - \pi_{\beta_k}(x_i)\big)x_i$$

- Convergence vers un minimum local prouvée

Évaluation des performances

- Paramétrisation: $K = 2$, $Z = 1$ pour un risque de décès élevé

Évaluation des performances

- Paramétrisation: $K = 2$, $Z = 1$ pour un risque de décès élevé
- Lois

$$\begin{aligned} \text{Weibull} \quad & f_k(t;\alpha_k) = (1-\phi_k)^{t^{\mu_k}} - (1-\phi_k)^{(t+1)^{\mu_k}} \\ \text{Géometrique} \quad & f_k(t;\alpha_k) = \alpha_k(1-\alpha_k)^{t-1} \end{aligned}$$

Évaluation des performances

- Paramétrisation: $K = 2$, $Z = 1$ pour un risque de décès élevé
- Lois

$$\text{Weibull} \quad f_k(t; \alpha_k) = (1 - \phi_k)^{t^{\mu_k}} - (1 - \phi_k)^{(t+1)^{\mu_k}}$$

$$\text{Géometrique} \quad f_k(t; \alpha_k) = \alpha_k (1 - \alpha_k)^{t-1}$$

- Modèles concurrents considérés
  - Cox PH pénalisé par l'Elastic-Net

Évaluation des performances

- Paramétrisation: $K = 2$, $Z = 1$ pour un risque de décès élevé
- Lois

$$\text{Weibull} \quad f_k(t;\alpha_k) = (1-\phi_k)^{t^{\mu_k}} - (1-\phi_k)^{(t+1)^{\mu_k}}$$

$$\text{Géometrique} \quad f_k(t;\alpha_k) = \alpha_k(1-\alpha_k)^{t-1}$$

- Modèles concurrents considérés
  - Cox PH pénalisé par l'Elastic-Net
  - CURE pénalisé par l'Elastic-Net

Évaluation des performances

- Paramétrisation: $K = 2$, $Z = 1$ pour un risque de décès élevé
- Lois

$$\text{Weibull} \quad f_k(t; \alpha_k) = (1-\phi_k)^{t^{\mu_k}} - (1-\phi_k)^{(t+1)^{\mu_k}}$$

$$\text{Géometrique} \quad f_k(t; \alpha_k) = \alpha_k (1-\alpha_k)^{t-1}$$

- Modèles concurrents considérés
  - Cox PH pénalisé par l'Elastic-Net
  - CURE pénalisé par l'Elastic-Net
- Marqueur $M = \pi_{\hat{\beta}}(X)$ ou $\exp(X^\top \hat{\beta}^{\text{cox}})$

Évaluation des performances

- Paramétrisation: $K = 2$, $Z = 1$ pour un risque de décès élevé
- Lois

$$\text{Weibull} \quad f_k(t;\alpha_k) = (1-\phi_k)^{t^{\mu_k}} - (1-\phi_k)^{(t+1)^{\mu_k}}$$

$$\text{Géometrique} \quad f_k(t;\alpha_k) = \alpha_k(1-\alpha_k)^{t-1}$$

- Modèles concurrents considérés
  - Cox PH pénalisé par l'Elastic-Net
  - CURE pénalisé par l'Elastic-Net
- Marqueur $M = \pi_{\hat{\beta}}(X)$ ou $\exp(X^\top \hat{\beta}^{\text{cox}})$
- Métriques :
  - $\text{AUC}^{\mathbb{C},\mathbb{D}}(t) = \int_{-\infty}^{\infty} \text{TPR}^{\mathbb{C}}(\xi,t) \left| \frac{\partial \text{FPR}^{\mathbb{D}}(\xi,t)}{\partial \xi} \right| \mathrm{d}\xi$ [1]

Évaluation des performances

- Paramétrisation: $K = 2$, $Z = 1$ pour un risque de décès élevé
- Lois

$$\text{Weibull} \quad f_k(t;\alpha_k) = (1-\phi_k)^{t^{\mu_k}} - (1-\phi_k)^{(t+1)^{\mu_k}}$$

$$\text{Géometrique} \quad f_k(t;\alpha_k) = \alpha_k(1-\alpha_k)^{t-1}$$

- Modèles concurrents considérés
  - Cox PH pénalisé par l'Elastic-Net
  - CURE pénalisé par l'Elastic-Net
- Marqueur $M = \pi_{\hat\beta}(X)$ ou $\exp(X^\top \hat\beta^{\text{cox}})$
- Métriques :
  - $\text{AUC}^{\mathbb{C},\mathbb{D}}(t) = \int_{-\infty}^{\infty} \text{TPR}^{\mathbb{C}}(\xi,t) \left| \frac{\partial \text{FPR}^{\mathbb{D}}(\xi,t)}{\partial \xi} \right| \mathrm{d}\xi$ [1]
  - C-index $\mathcal{C}_\tau = \mathbb{P}[M_i > M_j | Y_i < Y_j, Y_i < \tau]$

Évaluation des performances

- Paramétrisation: $K = 2$, $Z = 1$ pour un risque de décès élevé
- Lois

$$\text{Weibull} \quad f_k(t; \alpha_k) = (1 - \phi_k)^{t^{\mu_k}} - (1 - \phi_k)^{(t+1)^{\mu_k}}$$

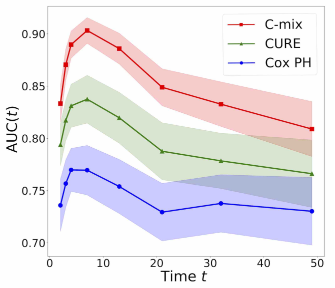
$$\text{Géometrique} \quad f_k(t; \alpha_k) = \alpha_k (1 - \alpha_k)^{t-1}$$

- Modèles concurrents considérés
  - Cox PH pénalisé par l'Elastic-Net
  - CURE pénalisé par l'Elastic-Net
- Marqueur $M = \pi_{\hat{\beta}}(X)$ ou $\exp(X^\top \hat{\beta}^{\text{cox}})$
- Métriques :
  - $\text{AUC}^{\mathbb{C},\mathbb{D}}(t) = \int_{-\infty}^{\infty} \text{TPR}^{\mathbb{C}}(\xi, t) \left| \frac{\partial \text{FPR}^{\mathbb{D}}(\xi, t)}{\partial \xi} \right| \mathrm{d}\xi$ [1]
  - C-index $\mathcal{C}_\tau = \mathbb{P}[M_i > M_j | Y_i < Y_j, Y_i < \tau]$
  - Prédiction de $T_i > \epsilon$ avec $\hat{S}_i(\epsilon | X_i = x_i)$ pour différents ϵ, où

$$\hat{S}_i(t | X_i = x_i) = \pi_{\hat{\beta}}(x_i) \hat{S}_1(t) + \big(1 - \pi_{\hat{\beta}}(x_i)\big) \hat{S}_0(t) \quad \text{(C-mix \& CURE)}$$

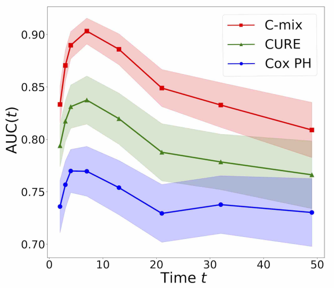
$$\hat{S}_i(t | X_i = x_i) = [\hat{S}_0^{\text{cox}}(t)]^{\exp(x_i^\top \hat{\beta}^{\text{cox}})} \quad \text{(Cox PH)}$$

Application sur données synthètiques

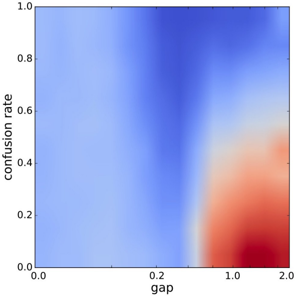


(a) AUC(t) sur des données simulées suivant le C-mix

Application sur données synthètiques



(a) AUC(t) sur des données simulées suivant le C-mix



(b) Sélection de variables du C-mix sur des données simulées suivant le modèle de Cox

Application sur données génétiques

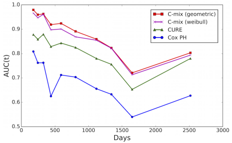
Table 1: Comparison du C-index sur les données TCGA ($d = 20531$)

Cancer		BRCA			GBM			KIRC		
Modèle		C-mix	CURE	Cox PH	C-mix	CURE	Cox PH	C-mix	CURE	Cox PH
	100	**0.792**	0.764	0.705	**0.826**	0.695	0.571	**0.768**	0.732	0.716
d	300	**0.782**	0.753	0.723	**0.849**	0.697	0.571	**0.755**	0.691	0.698
	1000	**0.817**	0.613	0.577	**0.775**	0.699	0.592	**0.743**	0.690	0.685

Application sur données génétiques

Table 1: Comparison du C-index sur les données TCGA ($d = 20531$)

Cancer		BRCA			GBM			KIRC		
Modèle		C-mix	CURE	Cox PH	C-mix	CURE	Cox PH	C-mix	CURE	Cox PH
	100	**0.792**	0.764	0.705	**0.826**	0.695	0.571	**0.768**	0.732	0.716
d	300	**0.782**	0.753	0.723	**0.849**	0.697	0.571	**0.755**	0.691	0.698
	1000	**0.817**	0.613	0.577	**0.775**	0.699	0.592	**0.743**	0.690	0.685



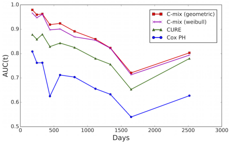
(a) AUC(t)

Figure 2: Résultats sur le cancer BRCA

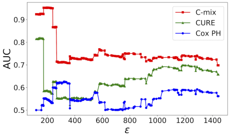
Application sur données génétiques

Table 1: Comparison du C-index sur les données TCGA ($d = 20531$)

Cancer		BRCA			GBM			KIRC		
Modèle		C-mix	CURE	Cox PH	C-mix	CURE	Cox PH	C-mix	CURE	Cox PH
d	100	**0.792**	0.764	0.705	**0.826**	0.695	0.571	**0.768**	0.732	0.716
	300	**0.782**	0.753	0.723	**0.849**	0.697	0.571	**0.755**	0.691	0.698
	1000	**0.817**	0.613	0.577	**0.775**	0.699	0.592	**0.743**	0.690	0.685



(a) AUC(t)



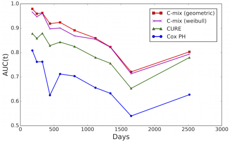
(b) AUC

Figure 2: Résultats sur le cancer BRCA

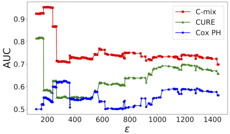
Application sur données génétiques

Table 1: Comparison du C-index sur les données TCGA ($d = 20531$)

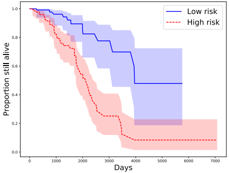
Cancer		BRCA			GBM			KIRC		
Modèle		C-mix	CURE	Cox PH	C-mix	CURE	Cox PH	C-mix	CURE	Cox PH
d	100	**0.792**	0.764	0.705	**0.826**	0.695	0.571	**0.768**	0.732	0.716
	300	**0.782**	0.753	0.723	**0.849**	0.697	0.571	**0.755**	0.691	0.698
	1000	**0.817**	0.613	0.577	**0.775**	0.699	0.592	**0.743**	0.690	0.685



(a) AUC(t)



(b) AUC



(c) Courbes de survie

Figure 2: Résultats sur le cancer BRCA

Temps de calcul

Table 2: Temps de calcul en seconde sur le BRCA $n = 1211$, (C-index)

Modèle		C-mix	CURE	Cox PH
	100	**0.025** (**0.792**)	1.992 (0.764)	0.446 (0.705)
d	300	**0.027** (**0.782**)	2.343 (0.753)	0.810 (0.723)
	1000	**0.139** (**0.817**)	12.067 (0.613)	2.145 (0.577)

Temps de calcul

Table 2: Temps de calcul en seconde sur le BRCA $n = 1211$, (C-index)

Modèle		C-mix	CURE	Cox PH
	100	**0.025** (**0.792**)	1.992 (0.764)	0.446 (0.705)
d	300	**0.027** (**0.782**)	2.343 (0.753)	0.810 (0.723)
	1000	**0.139** (**0.817**)	12.067 (0.613)	2.145 (0.577)

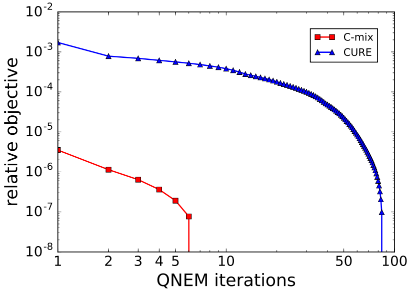


Figure 3: Comparaison de la convergence entre C-mix et CURE à travers l'objectif relatif défini par $\left(\ell_n^{\text{pen}}(\theta^{(l)}) - \ell_n^{\text{pen}}(\hat{\theta})\right)/\ell_n^{\text{pen}}(\hat{\theta})$ à la l-ième itération

Conclusion

- Meilleures performances que les modèles CURE et Cox PH : en prédiction, sélection de variable, robustesse, temps de calcul, interprétabilité

Conclusion

- Meilleures performances que les modèles CURE et Cox PH : en prédiction, sélection de variable, robustesse, temps de calcul, interprétabilité
- Détection de sous-groupes de patients relativement à leurs risques

Conclusion

- Meilleures performances que les modèles CURE et Cox PH : en prédiction, sélection de variable, robustesse, temps de calcul, interprétabilité
- Détection de sous-groupes de patients relativement à leurs risques

Article associé

S. Bussy, A. Guilloux, S. Gaïffas, A.S. Jannot
C-mix: a high dimensional mixture model for censored durations, with applications to genetic data
Publié dans *Statistical Methods in Medical Research*, 2017.

Conclusion

- Meilleures performances que les modèles CURE et Cox PH : en prédiction, sélection de variable, robustesse, temps de calcul, interprétabilité
- Détection de sous-groupes de patients relativement à leurs risques

Article associé

S. Bussy, A. Guilloux, S. Gaïffas, A.S. Jannot
C-mix: a high dimensional mixture model for censored durations, with applications to genetic data
Publié dans *Statistical Methods in Medical Research*, 2017.

Code Python

- Disponible à `https://github.com/SimonBussy/C-mix`

Conclusion

- Meilleures performances que les modèles CURE et Cox PH : en prédiction, sélection de variable, robustesse, temps de calcul, interprétabilité
- Détection de sous-groupes de patients relativement à leurs risques

Article associé

S. Bussy, A. Guilloux, S. Gaïffas, A.S. Jannot
C-mix: a high dimensional mixture model for censored durations, with applications to genetic data
Publié dans *Statistical Methods in Medical Research*, 2017.

Code `Python`

- Disponible à `https://github.com/SimonBussy/C-mix`
- Programmes annotés, notebooks et tutoriels

Conclusion

Code Python

- Programmes annotés, notebooks et tutoriels

```
from QNEM.inference import QNEM
from QNEM.simulation import CensoredGeomMixtureRegression

# Generate data
simu = CensoredGeomMixtureRegression(n_samples=1000, n_features=100,
                                     n_active_features=30)
X, Y, delta = simu.simulate()

# Choose between C-mix or CURE model
model = 'C-mix'

# Fit the model with a penalty strength equal to 10
learner = QNEM(model=model, l_elastic_net=10, eta=.1, max_iter=100,
               tol=1e-6, warm_start=True, fit_intercept=True)
learner.fit(X, Y, delta)

# Obtain the estimated marker (proba for been in the high risk group)
coeffs = learner.coeffs
marker = QNEM.predict_proba(X, fit_intercept, coeffs)
```

Conclusion

- Meilleures performances que les modèles CURE et Cox PH : en prédiction, sélection de variable, robustesse, temps de calcul, interprétabilité
- Détection de sous-groupes de patients relativement à leurs risques

Article associé

S. Bussy, A. Guilloux, S. Gaïffas, A.S. Jannot
C-mix: a high dimensional mixture model for censored durations, with applications to genetic data
Publié dans *Statistical Methods in Medical Research*, 2017.

Code `Python`

- Disponible à `https://github.com/SimonBussy/C-mix`
- Programmes annotés, notebooks et tutoriels

Références

[1] Patrick J Heagerty, Thomas Lumley, and Margaret S Pepe. Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2): 337–344, 2000.

[2] J. P. Klein and M. L. Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2005.

[3] Lynn Kuo and Fengchun Peng. A mixture-model approach to the analysis of survival data. *Biostatistics-Basel-*, 5:255–272, 2000.