Mélange de modèles de Cox avec des données hétérogènes

Eliz Peyraud, Julien Jacques, Guillaume Metzler, Alexandre Lopez

Mélange de modèles de Cox avec des données hétérogènes

Eliz Peyraud [1,2], Julien Jacques [1], Guillaume Metzler [1] et Alexandre Lopez [2]

[1] *Univ Lyon, Univ Lyon 2,*
ERIC UR 3083, Lyon, France
{eliz.peyraud,julien.jacques,guillaume.metzler}@univ-lyon2.fr.

[2] *Institut Georges Lopez (IGL)*
Lissieu, France
{epeyraud,alopez}@igl-transplantation.com

Résumé. Dans un contexte médical, l'analyse de survie en présence de covariables se fait en grande majorité à l'aide de modèles à risques proportionnels de Cox. Lorsque les données sont de nature hétérogènes, des mélanges de ces modèles peuvent être utilisés pour décrire au mieux les différents sous-groupes au sein d'une population. Le présent papier explique la construction d'un tel modèle de mélanges. En s'appuyant sur des travaux existants, l'application d'un mélange de modèles de Cox nous permettra de prédire la probabilité de survie à un temps t d'un patient selon son appartenance à un certain sous-groupe de la population, après un acte chirurgical lourd.

Mots-clés. Analyse de survie, Modèle de Cox, Modèles de mélanges, données médicales, données hétérogènes.

Abstract. In a medical context, the vast majority of survival analysis in the presence of covariates is done using Cox proportional hazards models. When the data are heterogeneous in nature, mixtures of these models can be used to best describe the different subgroups within the population. This paper explains the construction of such a mixture model. Based on existing work, the application of a mixture of Cox models will allow us to predict the probability of survival at a time t of a patient according to his belonging to a certain subgroup of the population, after a heavy surgical procedure.

Keywords. Survival Analysis, Cox Proportional Hazards Model, Mixture model, Medical Data, Heterogeneous Data.

1 Introduction

L'analyse de survie est une méthode couramment utilisée dans le domaine médical pour, entre autre, estimer la durée de vie d'un patient suite à une opération lourde ou à la prise d'un traitement conséquent [1]. Dans un contexte exclusivement médical, ces

modèles sont importants pour guider les médecins dans le choix des traitements à prescrire en estimant le taux de survie du patient avec la prise en compte de ce nouveau facteur.

A ce titre, les modèles de Cox [6] constituent les principaux outils de modélisation de ce phénomène via la prise en compte d'informations relatives aux patients. En effet, ceux-ci permettent de modéliser la probabilité de survie d'un patient après un évènement en fonction des covariables.

En revanche, l'hétérogénéité des profils de patients (hétérogénéité représentée par les différentes covariables) peut souvent mettre à mal l'efficacité de tels modèles. Les groupes d'âges auxquels appartiennent les patients, le genre, ou encore les antécédants médicaux, sont entre autres des facteurs discriminant la population en différents sous-groupes. Pour y remédier, les mélanges de modèles de Cox peuvent être utilisés d'une part pour décrire l'ensemble de la population de patients étudiés, et d'autre part pour identifier ces différents sous-groupes. En nous appuyant sur des travaux déjà réalisés [3, 4], nous allons, dans ce papier, présenter une application directe des mélanges de modèles de Cox.

La section 2 présentera le modèle à risques proportionnels de Cox ainsi que son application dans le cas de mélanges. Dans la section 3 nous mettrons en oeuvre des expériences pour tester l'efficacité de l'algorithme d'inférence associé. Enfin nous terminerons sur quelques persepectives notamment la sélection de variables dans les cas en grande dimension.

2 Modèle à risques proportionnels de Cox

On cherche à analyser la probabilité de survie du patient après un évènement ayant eu lieu à un temps t_0 (*e.g.* un acte chirurgical lourd). Cela passe par l'étude d'une variable aléatoire T qui représente la *durée de survie* du patient à partir de l'évènement qui s'est produit à t_0. Pour la suite on supposera $t_0 = 0$ (sans perte de généralité).

On définit alors la *fonction de survie* d'un individu X, notée $S(t \mid X)$, comme la probabilité que le temps de survie T du patient soit supérieur à un certain temps t :

$$S(t \mid X) = \mathbb{P}(T \geq t \mid X),$$

où X représente l'ensemble des covariables de l'individu concerné.

Lors du receuil des informations, il n'est pas rare que le suivi du patient s'arrête avant que l'on ait pu observer la date de décès. On parle dans ce cas de censure. Pour la suite nous définirons le temps t comme le temps de décès observé si l'individu n'est pas censuré, ou la date de dernières nouvelles si l'individu est censuré.

L'approche historique classique permettant de modéliser la fonction de survie S à partir de covariables est le modèle à risque proportionnel de Cox [5].

2.1 Modèle à risques proportionnels de Cox

Le modèle de Cox est un modèle semi-paramétrique servant à modéliser le phénomène d'apparition d'évènements (comme le décès d'un patient) au cours du temps, en fonction des covariables.

Ces phénomènes interviennent selon une loi $\lambda(t, X)$, appélée *risque instantané* ou *fonction de défaillance*, qui selon le modèle de Cox peut être modélisée comme suit :

$$\lambda(t, X) = \lambda_0(t) \exp(\beta^T X),$$

où λ_0 est appelé la fonction de risque de base (autrement dit le risque instantané de l'évènement lorsque toutes les covariables sont nulles), $\beta \in \mathbb{R}^p$ est le vecteur de paramètres du modèle et $X \in \mathbb{R}^p$ un vecteur de covariables.

La fonction de survie S, le risque instantané λ, et la fonction de densité f sont liés par la relation :

$$f(t, X) = \lambda(t, X) S(t \mid X).$$

Dans l'expression de λ le premier terme $\lambda_0(t)$ dépend uniquement du temps tandis que le second terme $\exp(\beta^T X)$ ne dépend que des covariables. Une hypothèse majeure qui en découle pour pouvoir utiliser le modèle de Cox est l'hypothèse des risques proportionnels : le rapport de risque instantané entre deux individus doit être constant en temps.

La fonction $\lambda_0(t)$ étant de forme inconnue, nous ne pouvons utiliser directement la vraisemblance du modèle pour en estimer les paramètres. C'est pourquoi à partir de l'observation d'un échantillon $(X_1, \ldots, X_n, \delta_1, \ldots, \delta_n)$, où δ_i est l'indicatrice de censure de l'individu i ($\delta_i = 0$ si l'individu est censuré, 1 sinon), l'estimateur $\hat{\beta}$ est obtenu en maximisant la log-vraisemblance partielle [5] (indépendante de λ_0) :

$$\log\left(L(\beta, X_1, \ldots, X_n, \delta_1, \ldots, \delta_n)\right) = \sum_{i=1}^{n} \delta_i \left[\beta^T X_i - \log\left(\sum_{j \in R(t_i)} \exp(\beta^T X_j) \right) \right],$$

où $R(t_i)$ est l'ensemble des individus susceptibles de subir une défaillance (ou censure) au temps t_i, i.e. l'ensemble des individus encore en vie au temps $t_i - \varepsilon,\ \varepsilon > 0$ (ou non censurés).

Cette estimation nous permet alors de retrouver la fonction de survie associée. En effet, la fonction de survie est liée à la fonction de risque instantané par la relation suivante :

$$S(t \mid X) = \exp\left(- \int_0^t \lambda(u, X)\, du \exp(\beta^T X) \right).$$

Ce modèle présente ses limites lorsque les patients ne sont pas similaires et ne peuvent être décrits avec les mêmes paramètres. Des critères restrictifs sur la population identifiés aujourd'hui par les médecins sont alors appliqués lors des études statistiques (adultes, sans antécédants médicaux, etc). Une façon de palier ces difficultés est d'utiliser un mélange

de modèle de Cox qui permettra non seulement de prendre en compte l'hétérogénéité des patients dans un modèle unique, mais également de trouver de nouveaux critères restrictifs qui n'ont pas encore été découverts par les médécins jusqu'à présent et d'identifier les variables clés pour prédire la survie spécifique à chaque sous-groupes de patients.

2.2 Mélange de modèles de Cox

Dans cette section, nous supposerons que la forme de la fonction du risque de base λ_0 est connue (pouvant par exemple suivre une loi exponentielle, ou une loi de Weibull telle que $\lambda_0(\alpha, l, t) = l\alpha t^{\alpha-1}$, $l, \alpha > 0$). Cette approximation du risque de base nous permet de basculer dans un cas paramétrique, ce qui facilitera notamment l'étude de la fonction de vraisemblance pour la suite.

Modèles de mélanges de Cox Pour un mélange à $g \in \mathbb{N}$ composantes, la fonction de survie s'écrit comme la somme de chaque fonctions de survie S_h du mélange multipliées par leurs proportions respectives π_h $(h = 1, .., g)$:

$$S(t \mid X) = \sum_{h=1}^{g} \pi_h S_h(t \mid X).$$

De même pour la fonction de densité f :

$$f(X, t) = \sum_{h=1}^{g} \pi_h f_h(X, t).$$

Le risque instantané étant défini comme le rapport entre la fonction de densité f et la fonction de survie S, on trouve facilement les formes des fonctions de densité du mélange par la relation : $f_h(X, t) = \lambda_h(X, t)S_h(t \mid X)$

Le risque de base λ_0 ayant à présent une forme paramétrique supposée connue, il n'est plus nécessaire d'utiliser la vraisemblance partielle dans ce cas. La vraisemblance du modèle peut alors être ré-écrite comme le produit des fonctions de densité dans le cas des individus non censurés (*i.e.* dont on connait leurs dates de décès) ou des fonctions de survie dans le cas des individus censurés :

$$\begin{aligned} L(\boldsymbol{\beta}, \mathbf{l}, \boldsymbol{\alpha}, \mathbf{x}, \mathbf{t}, \boldsymbol{\delta}) &= \prod_{j=1}^{n} f(x_j, t_j, \beta, l, \alpha)^{\delta_j} S(x_j, t_j, \beta, l, \alpha)^{1-\delta_j}, \\ &= \prod_{j=1}^{n} \left(\sum_{h=1}^{g} \pi_h f_h(x_j, t_j, \beta_h, l_h, \alpha_h) \right)^{\delta_j} \left(\sum_{h=1}^{g} \pi_h S_h(x_j, t_j, \beta_h, l_h, \alpha_h) \right)^{1-\delta_j}, \end{aligned}$$

où $(\mathbf{x}, \mathbf{t}, \boldsymbol{\delta}) = (x_1, ..., x_n, t_1, ..., t_n, \delta_1, ..., \delta_n)$ représente l'ensemble des individus avec leurs temps de vie et censures respectives et $(\boldsymbol{\beta}, \mathbf{l}, \boldsymbol{\alpha}) = (\beta_1, ..., \beta_g, l_1, ..., l_g, \alpha_1, ..., \alpha_g)$ l'ensemble des paramètres selon chaque composante du mélange.

Estimation à l'aide d'un algorithme EM Nous allons nous focaliser sur un cas particulier du modèle de Cox, où la fonction de risque de base λ_0 est modélisée selon la loi de Weibull. Le risque de base s'écrit donc ainsi :

$$\lambda_0(\alpha, l, t) = l\alpha t^{\alpha-1},$$

avec α, l deux réels positifs.

Les paramètres $(\pi_1, ..., \pi_{g-1}, \alpha, l, \beta)$ peuvent alors être estimés par maximum de vraisemblance à l'aide d'un algorithme EM (Expectation-Maximization). La première étape de cet algorithme consiste à calculer pour chaque individu la probabilité *a posteriori* d'apppartenir à chaque composante. Autrement dit, il nous faut écrire la vraisemblance complète du modèle. Pour cela, on introduit une variable binaire $Z_{jh} = 1$ si l'individu j appartient à la classe h, 0 sinon. La log-vraisemblance complétée s'écrit alors :

$$\begin{aligned}\ell_c(\boldsymbol{\beta}, \mathbf{l}, \boldsymbol{\alpha}, \mathbf{x}, \mathbf{t}, \boldsymbol{\delta}, Z) = \sum_{j=1}^{n} [&\delta_j Z_{jh} \log(\sum_{h=1}^{g} \pi_h f_h(x_j, t_j, \beta_h, l_h, \alpha_h)) \\ &+ (1-\delta_j) Z_{jh} \log(\sum_{h=1}^{g} \pi_h S_h(x_j, t_j, \beta_h, l_h, \alpha_h))].\end{aligned}$$

(**E-step**) La première étape de l'algorithme EM revient à calculer l'espérance conditionnelle de Z de la vraisemblance completée :

$$\begin{aligned}\mathbb{E}(\ell_c(\boldsymbol{\beta}, \mathbf{l}, \boldsymbol{\alpha}, \mathbf{x}, \mathbf{t}, \boldsymbol{\delta}, Z)|\mathbf{x}, \mathbf{t}, \boldsymbol{\delta}) = \sum_{j=1}^{n} [&\delta_j \mathbb{E}(Z_{jh}|\mathbf{x}, \mathbf{t}, \boldsymbol{\delta}) \log(\sum_{h=1}^{g} \pi_h f_h(x_j, t_j, \beta_h, l_h, \alpha_h)) \\ &+ (1-\delta_j) \mathbb{E}(Z_{jh}|\mathbf{x}, \mathbf{t}, \boldsymbol{\delta}) \log(\sum_{h=1}^{g} \pi_h S_h(x_j, t_j, \beta_h, l_h, \alpha_h))].\end{aligned}$$

En posant alors $\mathbb{E}(Z_{jh}|\mathbf{x}, \mathbf{t}, \boldsymbol{\delta}) = \tau_{j_h}$ on trouve finalement la probabilité *a posteriori* pour un individu j d'appartenir à la composante h du mélange :

$$\tau_{j_h} = \frac{(\mathbb{1}_{\delta=1} f_h(x_j, t_j, \beta_h, l_h, \alpha_h) + \mathbb{1}_{\delta=0} S_h(x_j, t_j, \beta_h, l_h, \alpha_h)) \times \pi_h}{\sum_{l=1}^{g} (\mathbb{1}_{\delta=1} f_l(x_j, t_j, \beta_l, l_l, \alpha_l) + \mathbb{1}_{\delta=0} S_l(x_j, t_j, \beta_l, l_l, \alpha_l)) \times \pi_l}.$$

(**M-step**) La mise à jour des paramètres $(\pi_1, ..., \pi_{g-1})$ à l'itération k se fait ensuite classiquement par :

$$\pi_h^{k+1} = \sum_{j=1}^{n} \frac{\tau_{hj}^{(k)}}{n}, \qquad \forall h \in \{1, ..., g-1\},$$

et $\pi_g = 1 - \sum_{h=1}^{g-1} \pi_h$.
Pour les paramètres $(\beta_h^{k+1}, \alpha_h^{k+1}, l_h^{k+1})$, la mise à jour se fait selon la résolution du système d'équations [1] suivant [3] :

$$\begin{cases} \sum_{j=1}^{n} \tau_{jh}^{(k)} \left(\delta_j - \exp\left(x_j^T \beta_h^{(k)}\right) l_h^{(k)} t_j^{\alpha_h^{(k)}} \right) x_j = 0, \\ \sum_{j=1}^{n} \tau_{jh}^{(k)} \left(\dfrac{\delta_j}{l_h^{(k)}} - \exp\left(x_j^T \beta_h^{(k)}\right) l_h^{(k)} t_j^{\alpha_h^{(k)}} \right) = 0, \\ \sum_{j=1}^{n} \tau_{jh}^{(k)} \left(\dfrac{\delta_j}{\alpha_h^{(k)}} + \delta_j \log(t_j) - \exp\left(x_j^T \beta_h^{(k)}\right) l_h^{(k)} t_j^{\alpha_h^{(k)}} \log(t_j) \right) = 0. \end{cases}$$

Les étapes d'Expectation et de Maximization de l'algorithme sont répétées k fois de manière itérative jusqu'à convergence. On considérera par ailleurs des initialisations aléatoires multiples de l'algorithme EM. Enfin, le nombre de composantes du mélange sera sélectionné par minimisation du critère BIC :

$$BIC = -2 \log(L(\hat{\boldsymbol{\beta}}, \hat{\mathbf{l}}, \hat{\boldsymbol{\alpha}}, \mathbf{x}, \mathbf{t}, \boldsymbol{\delta})) + r \log(n),$$

avec n le nombre d'individus et r le nombre de paramètres du modèle.

3 Expériences

Les données utilisées pour les expériences suivantes sont des données simulées sous la forme (x, t, δ). Celles-ci sont utilisées en vue de faire des premières expériences sur la convergence de l'algorithme EM proposé pour les mélanges de modèles de Cox.

3.1 Protocole expérimental

Génération des données La génération des temps de survie $\tilde{t}$ est faite à partir du calcul de l'inverse de la fonction de survie :

$$S^{-1}(U|\tilde{t}) = \Lambda_0^{-1} \left(-\frac{\log(U)}{\exp(x^T \beta)} \right),$$

avec U qui suit une loi uniforme entre 0 et 1, et $\Lambda_0 = \int \lambda_0(u) du$. Dans le cas où $\lambda_0(t) = l\alpha t^{\alpha-1}$ (Weibull) on obtient $\Lambda_0(t) = lt^\alpha$ et donc :

$$\tilde{t} = \left(-\frac{\log(u)}{l \exp(x^T \beta)} \right)^{\frac{1}{\alpha}}.$$

1. La résolution du système est donnée par des outils de résolutions numériques de problèmes non-linéaires (par exemple scipy.optimize sous python).

Parallèlement, des temps de censure C sont simulés selon une loi exponentielle $\mathcal{E}(1/50)$. On prendra finalement comme temps $t = \min(\tilde{t}, c)$ et $\delta = 0$ ou 1 pour l'indicatrice de censure correspondante.

On se place ici dans le cas où l'on a 3 classes de proportions égales et une covariable de paramètre β. On simule les covariables x selon une loi uniforme $\mathcal{U}[2.5, 2.6]$ pour la classe $h = 1$, $\mathcal{U}[0, 0.001]$ pour la classe $h = 2$ et $\mathcal{U}[-1, -1.1]$ pour la classe $h = 3$.

Les paramètres choisis pour la simulation des données sont données dans la Table 1. Ce choix a été fait de sorte à ce que nous puissions observer des courbes de survie distinctes pour chacune des composantes du mélange (Figure 1).

$h = 1$	$h = 2$	$h = 3$
$\pi_1 = 1/3$	$\pi_2 = 1/3$	$\pi_3 = 1/3$
$\beta_1 = 0.1$	$\beta_2 = 1$	$\beta_3 = 3$
$l_1 = 0.5$	$l_2 = 0.1$	$l_3 = 2$
$\alpha_1 = 4$	$\alpha_2 = 4$	$\alpha_3 = 4$

TABLE 1 – Paramètres de simulation

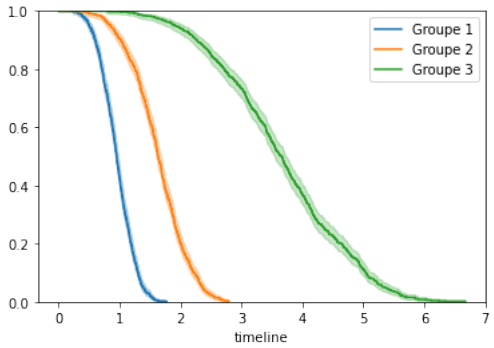

FIGURE 1 – Courbes de survie selon chaque composantes obtenues par l'estimateur de Kaplan-Meier avec un intervalle de confiance à 5%

Expériences L'algorithme sera testé 20 fois avec des simulations de données différentes. Pour chaque simulation, on tire aléatoirement les paramètres initiaux selon une loi uniforme et on répète l'opératoire également 20 fois. On observe finalement la convergence des paramètres et on conserve les résultats qui minimisent le critère BIC, avec $r = 15$ paramètres dans le cas d'un mélange à trois composantes.

On refait la même opération en augmentant le nombre d'individus dans chaque classe (ici $n = 300$ puis $n = 3000$) pour en comparer les résultats.

3.2 Résultats

Les résultats obtenus sont présentés sous forme de boxplot dans la Figure 2 . En particulier pour le paramètre β_h (qui est le paramètre qui a un impact sur les covariables, α et l décrivant uniquement le risque de base), on constate comme attendu que l'algorithme converge vers les valeurs choisies pour la simulation lorsque le nombre d'individus augmente. Les résultats observés pour les autres paramètres sont similaires.

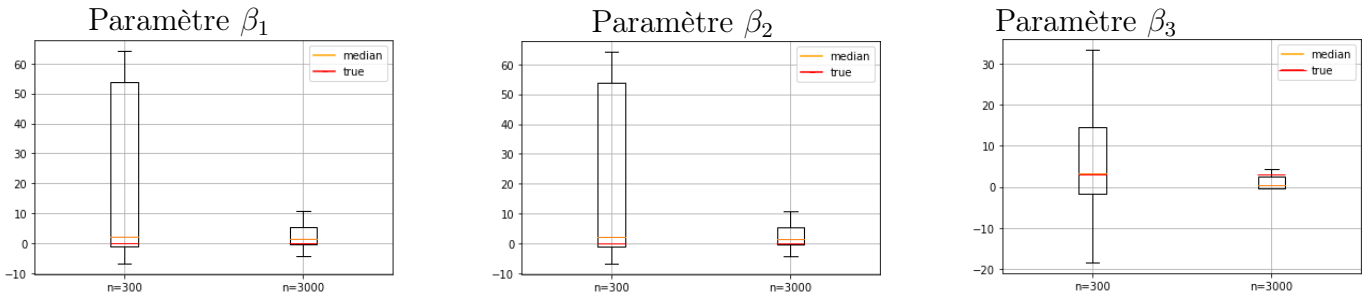

FIGURE 2 – Résultats d'expériences sur l'algorithme EM. Boxplot des estimations pour le paramètre β_1 pour la classe $h = 1$ lorsque $n = 300$ puis $n = 3000$, de même pour β_2 pour la classe $h = 2$ et β_3 pour la classe $h = 3$.

Les résultats présentés ci-dessus sur données simulées nous encouragent à faire des tests sur des données réelles. Les données réelles pour ces futurs tests proviennent d'un organisme américain retraçant le suivi post-opératoire d'environ 500 000 patients.

4 Perspectives

Les mélanges de modèle de Cox ont l'avantage de nous permettre de prendre en considération un grand nombre de facteurs au sein de notre population. Cependant, lorsque l'on se retrouve en grande dimension (grand nombre de covariables), nous aurons besoin d'effectuer une sélection de variables. Bien que la littérature soit riche concernant les modèles de Cox pénalisés [2], il n'en est pas de même pour les mélanges de modèles de Cox avec pénalisation. En s'inspirant des mélanges de modèles gaussiens pénalisés [7], nous souhaitons pour la suite de ce travail proposer un mélange de modèles de Cox avec une pénalité L_1 permettant non seulement de prendre en compte l'hétérogénéité des données mais également de faire une sélection de variables. Cependant contrairement aux formes que l'on peut retrouver pour les modèles gaussiens [7], nous proposons d'écrire le terme de

régularisation comme dépendant de chaque composante du mélange de modèles de Cox :

$$\ell_{pen}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{l}, \mathbf{x}, \mathbf{t}, \boldsymbol{\delta})) = \sum_{j=1}^{n} [\delta_j \log(\sum_{h=1}^{g} \pi_h f_h(\beta_h, \alpha_h, l_h, x_j, t_j) \\ + (1-\delta_j) \log(\sum_{h=1}^{g} \pi_h S_h(\beta_h, \alpha_h, l_h, x_j, t_j))] - \sum_{h=1}^{g} s_h \sum_{i=1}^{p} |\beta_i^h| \quad s_h > 0.$$

En écrivant la pénalité sous cette forme générale, l'hyperparamètre s serait aussi dépendant de la composante du mélange étudié. Cela permetttrait alors d'avoir une régularisation ré-affiner pour chaque cluster et donc de sélectionner les variables en fonction des informations pertinentes relativement aux différents sous-groupes de patients.

Références

[1] David Collett. *Modelling survival data in medical research*. CRC press, 2015.

[2] Jelle J Goeman. L1 penalized estimation in the cox proportional hazards model. *Biometrical journal*, 52(1) :70–84, 2010.

[3] Shu Kay Ng, Liming Xiang, and Kelvin Kai Wing Yau. *Mixture modelling for medical and health sciences*. Chapman and Hall/CRC, 2019.

[4] Rosen O. and Tanner M. Mixtures of proportional hazards regression models. *Statistics in Medicine*, 18(9) :1119–1131, 1999.

[5] Cox D. R. Regression models and life-tables. *Journal of the Royal Statistical Society : Series B (Methodological)*, 34(2) :187–202, 1972.

[6] Cox D. R. Partial likelihood. *Biometrika*, 62(2) :269–276, 1975.

[7] Nicolas Städler, Peter Bühlmann, and Sara Van De Geer. 1-penalization for mixture regression models. *Test*, 19 :209–256, 2010.