Régression gaussienne à poids logistiques et maximum de vraisemblance pénalisé

Lucie Montuelle [1] & Erwan Le Pennec [2]

[1] *Select, Inria Saclay Idf / LMO, Université Paris-Sud*
E-mail:lucie.montuelle@math.u-psud.fr
[2] *Select, Inria Saclay Idf / LMO, Université Paris-Sud*
E-mail:erwan.le_pennec@inria.fr

Résumé. Nous considérons le problème de l'estimation de densité conditionnelle à l'aide de mélanges gaussiens à poids logistiques et moyennes dépendant de la covariable. Nous souhaitons sélectionner le nombre de composantes du mélange, ainsi que les autres paramètres par maximum de vraisemblance pénalisé. Nous donnons une borne inférieure sur la pénalité telle que notre estimateur satisfasse une inégalité d'oracle. Des expériences numériques appuient notre analyse théorique.

Mots-clés. Mélange gaussien, sélection de modèle, maximum de vraisemblance pénalisé, densité conditionnelle

Abstract. We wish to estimate conditional density using Gaussian mixture model with logistic weights and means depending on the covariate. We aim at selecting the number of components of this model as well as the other parameters by a penalized maximum likelihood approach. We provide a lower bound on penalty that ensures an oracle inequality for our estimator. We perform some numerical experiments that support our theoretical analysis.

Keywords. Gaussian mixture, model selection, penalized maximum likelihood, conditional density

1 Cadre de travail et notations

Nous considérons un problème d'estimation de la densité conditionnelle $s_0(.|x)$, par rapport à la mesure de Lebesgue, d'une variable Y connaissant la covariable X. n paires $(X_i, Y_i)_{1\leq i\leq n}$ de variables aléatoires sont observées, où les covariables X_i sont indépendantes et les Y_i sont indépendantes conditionnellement aux X_i. Motivés par des applications en classification non supervisée, nous utilisons pour cela une modélisation de la densité conditionnelle s_0 par un mélange de régressions gaussiennes à poids logistiques:

$$s_{K,\nu,\Sigma,w}(y|x) = \sum_{k=1}^{K} \pi_{w(x),k} \Phi_{\nu_k(x),\Sigma_k}(y)$$

où $K \in \mathbb{N}^*$ est le nombre de composantes du mélange, $\Phi_{\nu(x),\Sigma}$ est la densité de la gaussienne de moyenne $\nu(x)$ et de matrice de covariance Σ,

$$\Phi_{\mu,\Sigma}(y) = \frac{1}{\sqrt{(2\pi)^p|\Sigma|}} \exp\left(\frac{-1}{2}(y-\mu)'\Sigma(y-\mu)\right)$$

et les poids peuvent toujours être définis à partir d'un K-tuplet $(w_1, \ldots, w_K)$ suivant un schéma logistique:

$$\pi_{w(x),k} = \frac{e^{w_k(x)}}{\sum_{k'=1}^{K} e^{w'_k(x)}}.$$

Estimer la densité s_0 revient donc à évaluer les fonctions w_k et ν_k, les matrices Σ_k, ainsi que le nombre de classes K. Cette estimation se base sur la minimisation d'une mesure de l'erreur entre la densité candidate et la vraie densité conditionnelle.

Des bornes non asymptotiques, reposant sur l'analyse de l'entropie à crochets des modèles, sont proposées par Maugis et Michel (2011) pour les mélanges gaussiens classiques. Seuls Chamroukhi et autres (2010) considèrent un cas particulier de notre modèle et fournissent des simulations numériques basées sur l'EM et le critère BIC.

Nous adopterons le point de vue de la sélection de modèles et utiliserons l'estimateur du maximum de vraisemblance pénalisé. Nous fournirons dans cet exposé une borne inférieure sur la pénalité permettant d'obtenir une borne non asymptotique de l'erreur. Ce résultat sera illustré par des simulations numériques.

2 Procédure de maximum de vraisemblance pénalisé

Notre estimateur repose sur un principe de maximum de vraisemblance pénalisé. Pour cela, nous définissons une collection $\mathcal{S} = (S_m)_{m\in\mathcal{M}}$ d'ensembles de mélanges de régressions gaussiennes à poids logistiques. A chaque ensemble S_m correspondent un nombre de composantes K, des ensembles de fonctions W_K et Υ_K respectivement pour les poids et pour les moyennes et une structure V_K pour les variances des gaussiennes. A modèle fixé, une façon naturelle de sélectionner un estimateur est de maximiser la vraisemblance:

$$\hat{s}_m = \arg\min_{s_m\in S_m} \sum_{i=1}^{n} -\log s_m(X_i|Y_i)$$

Notre procédure vise à sélectionner le meilleur modèle possible, le modèle au sein duquel l'estimateur choisi minimisera une distance de type Kullback-Leibler à la densité cible. Pour cela, nous définissons une pénalité pen(m) choisie proportionnelle à la dimension du modèle S_m, choisissons le modèle $S_{\hat{m}}$ qui minimise la vraisemblance pénalisée

$$S_{\hat{m}} = \arg\min_{S_m} \sum_{i=1}^{n} -\log \hat{s}_m(X_i|Y_i) + \text{pen}(m).$$

et définissons notre estimateur final par $\hat{s}_{\hat{m}}$.

3 Résultat théorique

Pour mesurer les performances de notre estimateur, nous définissons la distance de Kullback-Leibler tensorisée $KL^{\otimes n}$ par

$$KL_\rho^{\otimes n}(s,t) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\rho}KL(s(.|X_i), t(.|X_i))\right]$$

ainsi qu'une distance légérement plus petite: la distance de Jensen-Kullback-Leibler tensorisée $JKL^{\otimes n}$, correspondant à la distance de Kullback-Leibler entre la vraie densité et une combinaison convexe de celle-ci avec la densité estimée. Pour $\rho \in]0;1[$,

$$JKL_\rho^{\otimes n}(s,t) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\rho}KL(s(.|X_i), (1-\rho)s(.|X_i) + \rho t(.|X_i))\right]$$

L'intérêt de cette divergence est sa bornitude. En effet, cette distance est toujours bornée par $\frac{1}{\rho}\ln\frac{1}{1-\rho}$ et proche de la distance de Kullback-Leibler lorsque s et t sont proches. Cela permet d'obtenir une inégalité d'oracle sous des hypothèses faibles sur la collection de modèles et leur complexité.

Notre principal résultat nécessite deux hypothèses. La première, de type inégalité de Kraft, permet de contrôler la complexité de la collection.

Hypothèse (K): Il existe une famille $(x_m)_{m\in\mathcal{M}}$ de réels positifs telle que $\sum_{m\in\mathcal{M}} e^{-x_m} \leq \Xi < +\infty$.

La seconde est un contrôle entropique, qui est toujours satisfait lorsque les espaces W_K et Υ_K sont les images d'espaces paramétriques.

Hypothèse (DIM): Il existe deux constantes C_W et C_Υ telles que pour tout modèle S_m de la collection $\mathcal{S}$,

$$H_{\max_k \|\|_\infty}(\sigma, W_K) \leq \dim(W_K)\left(C_W + \ln\frac{1}{\sigma}\right)$$

et

$$H_{\max_k \sup_x \|\|_2}(\sigma, \Upsilon_K) \leq \dim(\Upsilon_K)\left(C_\Upsilon + \ln\frac{1}{\sigma}\right).$$

Notre résultat principal s'énonce alors

Théorème 1 *Pour toute collection de mélanges de régressions gaussiennes $(S_m)_{m\in\mathcal{M}}$, satisfaisant les hypothèses (K) et (DIM), il existe une constante C telle que pour tout $\rho \in (0;1)$ et pour tout $C_1 > 1$, il existe deux constantes κ_0 et C_2, dépendant seulement*

de ρ et C_1 telles que, si pour tout $m \in \mathcal{M}$, $pen(m) = \kappa((C + \ln(n))\dim(S_m) + x_m)$ avec $\kappa > \kappa_0$, l'estimateur du maximum de vraisemblance pénalisé $\hat{s}_{\hat{m}}$ vérifie:

$$\mathbb{E}\left[JKL_\rho^{\otimes n}(s_0, \widehat{s}_{\widehat{m}})\right] \leq C_1 \inf_{m\in\mathcal{M}} \left(\inf_{s_m \in S_m} KL_\lambda^{\otimes n}(s_0, s_m) + \frac{\kappa((C + \ln n)\dim(S_m) + x_m)}{n} \right) + C_1 \frac{\eta + \eta'}{n} + C_2 \frac{\Xi}{n}.$$

Cette inégalité montre que le risque de l'estimateur mesuré en distance JKL est majoré par le meilleur compromis entre un biais de modèle mesuré en norme KL et un terme jouant le rôle d'une variance. Ce terme est précisement, au terme x_m près, le meilleur majorant non asymptotique connu de la variance de l'estimateur par maximum de vraisemblance dans le modèle S_m. Pour minimiser l'arbitraire, x_m est choisi de façon à être négligeable face au terme $(C + \ln(n))\dim(S_m)$. Cela permet de prendre la pénalité proportionnelle à la dimension du modèle, quitte à légèrement augmenter κ.

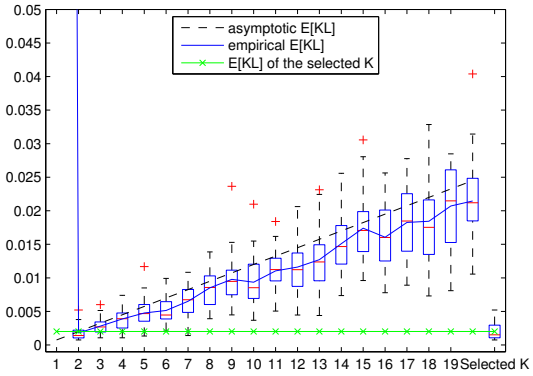
Ce cadre permet de sélectionner le nombre de classes, la structure de covariance (voir Celeux et Govaert (1995)) ou encore le degré maximal d'espaces de polynômes définissant W_K et Υ_K.

4 Illustration numérique

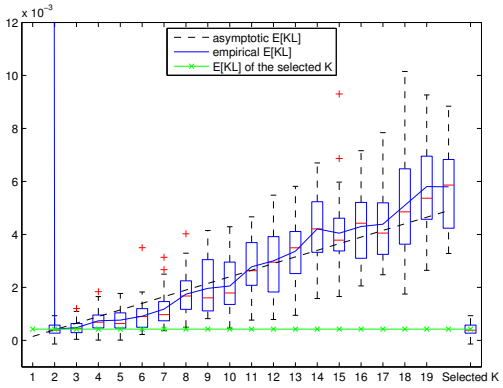
Pour illustrer ce résultat, nous nous plaçons dans le cas où X_i suit la loi uniforme sur [0;1] et $Y_i \in \mathbb{R}$. Les fonctions de poids w_k et les moyennes υ_k des estimateurs sont supposées linéaires. Aucune structure n'est imposée sur les matrices de de covariance.

Nous avons combiné les algorithmes EM et de Newton pour estimer les paramètres des modèles, puis utilisé notre critère pénalisé pour sélectionner le nombre de classes. Deux exemples ont été traités. Dans le premier, noté P, les données ont été simulées selon une densité appartenant à un modèle de la collection. Dans le second, désigné par NP, la vraie densité s_0 a des moyennes polynomiales de degré 2 et n'est donc pas dans la collection.

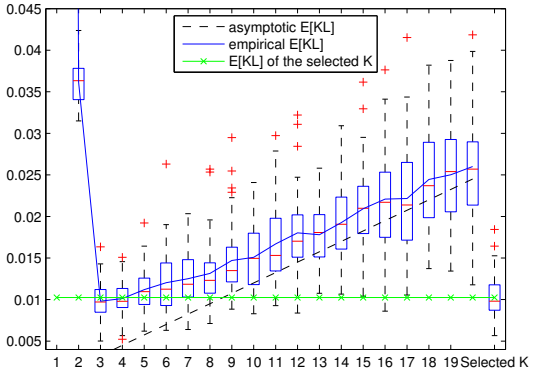
La figure représente les box-plots et la moyenne de la distance de Kullback-Leibler tensorisée évaluée par la méthode de Monte Carlo sur 55 lancers. Comme prévu par le théorème, nous constatons que la distance de Kullback-Leibler tensorisée moyenne entre la vraie densité s_0 et l'estimateur pénalisé $\hat{s}_{\hat{K}}$ est inférieure à celle entre s_0 et $\hat{s}_K$ pour $K \in \{1, \ldots, 20\}$. La moyenne de la distance de Kullback-Leibler semble avoir une pente de l'ordre de $\frac{\dim(S_m)}{2n}$ (tracé en pointillés). C'est le comportement attendu par l'heuristique AIC, lorsque la vraie densité appartient à une collection de modèles emboîtés. C'est le cas de l'exemple P. Le même phénomène est observé dans l'exemple NP mais aucune garantie théorique n'est fournie.



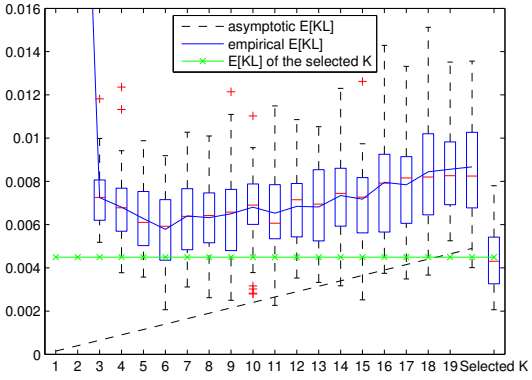
(a) Exemple P avec 2000 données



(b) Exemple P avec 10 000 données



(c) Exemple NP avec 2000 données



(d) Exemple NP avec 10 000 données

Figure 1: Box-plot de la distance de Kullback-Leibler en fonction du nombre de composants du mélange. Sur chaque graphique, le tracé le plus à droite représente la distance de Kullback-Leibler de l'estimateur pénalisé $\hat{s}_{\widehat{K}}$

Bibliographie

[1] Chamroukhi, F., Samé, A., Govaert, G. et Aknin, P. (2010), *A hidden process regression model for functional data description. Application to curve discrimination*, Neurocomputing, 73, 1210-1221.
[2] Celeux,G. et Govaert,G. (1995), *Gaussian parcimonious clustering models*, Pattern Recognition.
[3] Maugis, C. et Michel, B. (2011), *A non asymptotic penalized criterion for gaussian mixture model selection*, ESAIM P&S.