Quantile de régression : application à l'analyse de l'écotoxicité de molécules chimiques

Jonathan VILLAIN [1,2] & Ronan BUREAU [2] & Gilles DURRIEU [1]

[1] *Laboratoire de Mathématiques de Bretagne Atlantique, Université de Bretagne Sud et UMR CNRS 6205, Campus de Tohannic, 56017 Vannes*

[2] *Centre d'Études et de Recherche sur le Médicament de Normandie, UNICAEN, Caen*

ronan.bureau@unicaen.fr, {gilles.durrieu,jonathan.villain}@univ-ubs.fr

Résumé. L'estimation des propriétés écotoxicologiques de produits chimiques est une préoccupation environnementale majeure. Les modèles QSAR (Quantitative Structure-Activity Relationship) sont des modèles statistiques de régression linéaire et de classification souvent utilisés pour prédire l'écotoxicité de molécules chimiques. Nous considérons dans ce papier des régressions quantiles qui sont plus robustes à la présence de valeurs aberrantes tout en offrant l'avantage de s'intéresser à l'ensemble de la distribution conditionnelle de la variable d'intérêt et pas seulement à sa moyenne comme en régression linéaire. Nous proposons ici, dans ce souci de prédiction, des modèles quantiles en régression et Support Vector Machines (SVM) dans le domaine de la chemoinformatique.

Mots-clés. écotoxicologie, SVM, régression, classification, noyau, quantile, robustesse.

Abstract. The estimation of ecotoxicological properties of chemicals is a major environmental concern. The QSAR models (Quantitative Structure-Activity Relationship) are linear regression and classification models often used to predict the ecotoxicity of chemical molecules. We consider in this paper quantile regression estimators which are more robust to outliers providing a more detailed focus on the entire conditional distribution of the dependent variable and not only on its mean as in linear regression. We propose here, in this concern of prediction, quantile models in regression and Support Vector Machines (SVM) in the field of chemoinformatics.

Keywords. ecotoxicology, SVM, regression, classification, kernel, quantile, robustness.

1 Introduction

On parle aujourd'hui de plus en plus des problèmes liés à la santé et à l'environnement. L'estimation des propriétés toxicologiques et écotoxicologiques des produits chimiques est devenue une préoccupation environnementale majeure. Cette préoccupation a été à la base de la mise en place au niveau européen du programme REACH. Au sein de la règlementation REACH, les industriels de la chimie doivent fournir des informations sur

un certain nombre de critères concernant les propriétés physico-chimiques et (éco) toxicologiques. L'objectif général est d'estimer le risque chimique pour chaque composé chimique et d'écarter ceux classés cancérigènes, polluants ou persistants et bio-accumulatifs, avec dans ce cas une obligation de substitution. Trois types de méthodes sont autorisés pour la détermination de ces risques : les méthodes *in vivo* qui sont des méthodes précises mais qui posent un problème éthique et un coût important, les méthodes *in vitro* qui sont des méthodes associées à des tests au niveau cellulaire, méthodes intéressantes mais non développées pour l'ensemble des critères et enfin les méthodes *in silico* qui sont des méthodes basées sur une estimation à partir de bases de données. Dans ce travail, nous proposons de développer des modèles statistiques pour les méthodes *in silico* afin de décrire les molécules chimiques et relier cette description aux propriétés biologiques.

2 Modèles et estimateurs

Dans beaucoup d'applications (astronomie, biologie, chimie, médecine, physique, etc), les données sont contaminées par des valeurs aberrantes qui proviennent d'erreurs dues à l'environnement expérimental ou de tout autre cause, tout aussi triviale qu'une erreur d'enregistrement ou de lecture. Nous considérons dans ce papier des approches quantiles en régression et SVM.

2.1 Quantiles de régression

L'estimation au sens des moindres carrés (estimation L^2) est souvent utilisée du fait des facilités de calcul et de ses bonnes propriétés pour le modèle linéaire gaussien ; toutefois, ces estimateurs sont très sensibles à la présence de valeurs aberrantes. En revanche, la robustesse de la médiane (archétype d'estimation L^1) est connue de longue date. En 1964, Huber [**?**] a publié un article de référence sur l'estimation robuste du paramètre de location. Ces dernières années, un effort théorique considérable a été déployé pour construire des méthodes statistiques robustes. Mentionnons simplement ici que le travail de Huber a été étendu aux modéles linéaires par [1], [2], [9], [8], [10] et [11]. Nous considérons ici le modèle de régression linéaire suivant :

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

où $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ est le vecteur des observations, $\boldsymbol{X}$ est une matrice connue de dimension $n \times p$ ayant pour lignes $\boldsymbol{x}_i' \in \mathbb{R}^p$, $i = 1, \ldots, n$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ est un vecteur d'erreurs indépendantes de fonction de répartition F inconnue et de médiane nulle ($F^{-1}(1/2) = 0$) et $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ désigne le vecteur inconnu des paramètres de régression à estimer.

En 1978 Koenker et Basset ont proposé le concept de "quantile de régression". On

appelle θ-quantile de régression toute solution du problème de minimisation

$$\widehat{\boldsymbol{\beta}}(\theta) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\theta \left(Y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right) \tag{2}$$

où $\rho_\theta(x) = x\left(\theta - \mathbb{I}(x < 0)\right)$ et $\mathbb{I}(\mathcal{P})$ prend la valeur 1 ou 0 selon que la condition $\mathcal{P}$ est vérifiée ou non. Un cas particulier de cette classe d'estimateurs (obtenu pour $\theta = 1/2$) est l'estimateur L^1 ou la régression médiane qui s'obtient par résolution du problème de minimisation (2). La normalité asymptotique de l'estimateur $\widehat{\boldsymbol{\beta}}(\theta)$ a été donnée par [13] sous l'hypothèse d'erreurs i.i.d. et pour des erreurs indépendantes, mais pas nécessairement identiquement distributées [7] dans le modèle (1). La variance asymptotique de $\widehat{\boldsymbol{\beta}}(\theta)$ s'écrit dans sa forme générale

$$\boldsymbol{\Sigma_\theta} = \left(\theta\left(1-\theta\right)\right)\left(\boldsymbol{X'FX}\right)^{-1}\left(\boldsymbol{X'X}\right)\left(\boldsymbol{X'FX}\right)^{-1} \tag{3}$$

où $\boldsymbol{F} = \mathrm{diag}\{f_1(Q(\theta)), \ldots, f_n(Q(\theta))\}$ avec $Q(\theta)$ correspondant à la fonction quantile. Dans le cas d'une erreur i.i.d. dans le modèle (1), l'équation (3) est réduit à

$$\boldsymbol{\Sigma_\theta} = \left(\theta\left(1-\theta\right)/f^2(Q(\theta))\right)\left(\boldsymbol{X'X}\right)^{-1}, \tag{4}$$

où $1/f(Q(\theta))$ est la densité du quantile. Les variances asymptotiques (4) et (3) dépendant de la densité de probabilité des erreurs (inconnue), nous avons besoin de "bons" estimateurs de la variance asymptotique. Il est possible de procéder par une estimation directe en utilisant un estimateur non paramétrique à noyau de la densité du quantile ([4, 5]). Quand les observations sont indépendantes mais non identiquement distribuées, comme souvent dans le domaine de la chemoinformatique, il est possible d'étendre la théorie i.i.d. pour obtenir une version de l'estimateur sandwich de Huber-Eicker-White de la matrice de variance-covariance de $\widehat{\boldsymbol{\beta}}(\theta)$. D'autres estimateurs ont aussi été proposés pour ce problème, incluant le test des rangs comme décrit dans [6, 14, 15] et des méthodes de bootstrap ([16], [3], [12]).

2.2 Quantiles de régression SVM

Les Support Vector Machines (SVM) ont été développés dans les années 1990 à partir de travaux sur l'apprentissage statistique initiés par Valdimir Vapnik [18]. Le principe de base des SVM consiste à définir un hyperplan, dit de marge optimale, pour la séparation de classes comme solution d'un problème d'optimisation sous contraintes dont la fonction objectif ne s'exprime qu'à l'aide de produits scalaires entre vecteurs et dans lequel le nombre de contraintes "actives" ou vecteurs supports contrôle la complexité du modèle. Le passage à la recherche de surfaces séparatrices non linéaires est obtenu par l'introduction d'une fonction noyau dans le produit scalaire induisant implicitement une transformation non linéaire des données vers un espace de plus grande dimension. Soit $\boldsymbol{X}$ les variables

explicatives ou prédictives à valeurs dans un ensemble $\mathcal{F}$ et Y la variable à prédire. On note par $\{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$, un échantillon statistique de taille n et de loi F inconnue. Les SVM peuvent également être mis en œuvre en régression. Dans le cas non linéaire, le principe consiste à rechercher une estimation de $\hat{f}(x)$ d'un modèle $f(x)$ pour Y. Les observations faites dans l'ensemble $\mathcal{F}$ (en général $\mathbb{R}^p$) sont considérées comme étant transformées par une application non linéaire $\boldsymbol{x} \to \phi(\boldsymbol{x})$ qui va de $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{F}$ dans un espace muni d'un produit scalaire de plus grande dimension. Nous présentons maintenant la régression non linéaire quantile SVM notée QSMR.

La fonction quantile y_i conditionnellement à x_i peut s'écrire pour $i = 1, \ldots, n$:

$$Q(\theta/\boldsymbol{x}_i) = \boldsymbol{w}'_\theta \, \phi(\boldsymbol{x}_i) \qquad \text{pour} \qquad \theta \in (0,1), \tag{5}$$

où $\boldsymbol{w}_\theta$ désigne le θ-quantile de régression. QSVMR peut se définir comme dans (2) en minimisant pour $\theta \in (0,1)$

$$\frac{1}{2} \, || \, \boldsymbol{w}_\theta \, ||^2 + C \sum_{i=1}^{n} \rho_\theta \left(y_i - \boldsymbol{w}'_\theta \phi(\boldsymbol{x}_i) \right), \tag{6}$$

où C désigne le degré de pénalisation.

Une solution de (6) pour $\theta \in (0,1)$ s'obtient en optimisant sa version duale quadratique (voir [17]). Le θ-quantile de régression pour $\boldsymbol{x}^\star$ s'écrit alors :

$$Q(\theta/\boldsymbol{x}^\star) = \sum_{i=1}^{n} \left(\boldsymbol{\lambda}_i^- - \boldsymbol{\lambda}_i^+ \right) K(\boldsymbol{x}_i, \boldsymbol{x}^\star) \qquad \text{et} \qquad \boldsymbol{w}_\theta = \sum_{i=1}^{n} \left(\boldsymbol{\lambda}_i^- - \boldsymbol{\lambda}_i^+ \right) \phi(\boldsymbol{x}_i), \tag{7}$$

où λ_i^-, λ_i^+ sont les multiplicateurs de Lagrange et $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ désigne une fonction noyau. Nous considérons ici la fonction noyau de type radial gaussian (RBF) donnée par :

$$k \left(\boldsymbol{x}_i, \boldsymbol{x}_j \right) = \exp \left(- \frac{|| \, \boldsymbol{x}_i - \boldsymbol{x}_j \, ||^2}{2 \, \sigma^2} \right), \tag{8}$$

où le paramètre σ désigne la taille de la fenêtre. Le paramètre σ peut être déterminé par validation croisée.

3 Application en chémoinformatique

Le rapport de toxicité (Toxic Ratio), noté TR, a été développé par Neuwoehner et al. [19] pour déterminer le mode d'action d'une molécule. Il se détermine à partir d'un modèle de régression basé sur la relation entre $\log(1/CE_{50})$ (CE_{50} est la Concentration Effective médiane) et le coefficient de séparation octanol-eau notée $\log(P)$ définie par :

$$\log \left(\frac{1}{\mathrm{CE}_{50}} \right) = a \log(P) + b, \tag{9}$$

où a et b désignent les paramètres de régression inconnues à estimer. Le Toxic Ratio (TR) est alors obtenu par le rapport entre les valeurs accessibles dans les bases de données utilisées en chemoinformatique et les valeurs prédites par le modèle de régression. En pratique, les biochimistes considèrent que pour un TR supérieur à 10, la molécule possède un mode d'action spécifique sinon la molécule est considérée comme ayant un mode d'action non-spécifique (toxicité basale). Nous proposons de calculer le rapport de toxicité à partir d'un modèle de régression quantile. Nous sélectionnons différents quantiles de régression pour le calcul du TR mais nous donnons ici les résultats pour $\theta = 0.5$.

Nous considérons 401 produits chimiques pour lesquels nous avons les valeurs de CE_{50} (concentration aboutissant à une inhibition de 50 % de la croissance d'une algue (P. subcapitata)) ainsi que des informations sur la structure des produits chimiques (descripteurs topologiques). On veut à partir de ces descripteurs déterminer un modèle afin de pouvoir prédire la valeur de CE_{50}. Nous commençons par déterminer le mode d'action des produits chimiques afin de prédire et écarter les produits ayant un mode d'action spécifique. En considérant une régression médiane pour estimer les TR, un total de 336 produits chimiques est considéré comme n'ayant pas de modes d'action spécifique. On utilise ensuite une classification SVM afin d'obtenir une prédiction des modes d'action des molécules sur l'ensemble des descripteurs qui donne respectivement une erreur de classification en apprentissage de 1.75% et en validation croisée $(2/3, 1/3)$ de 14.95%. En considérant les 368 produits chimiques prédits comme n'ayant pas de modes d'action spécifique en validation croisée, nous effectuons une régression médiane SVM. Afin de choisir le nombre de variables à considérer, nous présentons dans la Figure 1 les critères SCE_R et R^2 obtenus en validation croisée. Une régression par segments sur les valeurs de SCE_R nous conduit à considérer 73 variables. Le modèle quantile SVM avec 73 variables donne un $R^2 = 0.68$ et une $SCE_R = 141.26$ en validation croisée (2/3,1/3).

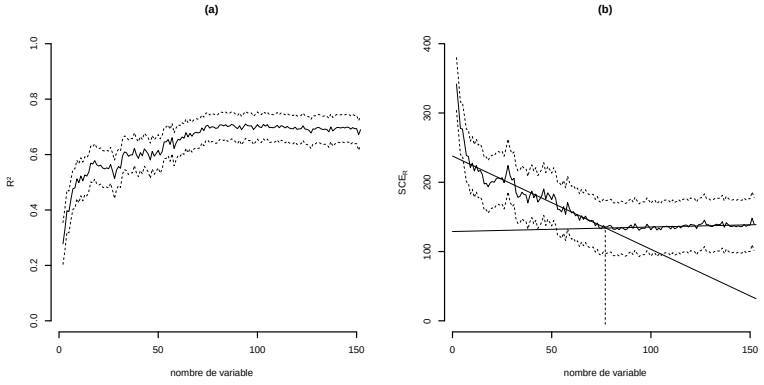

FIGURE 1 – Représentations en (a) du R^2 et en (b) du SCE_R en fonction du nombre de variable dans le modèle. Les bandes de confiance à 95% sont représentées en traits pointillés.

Références

[1] Andrews, D. F. (1974), A robust method for multiple linear regression, *Technometrics*, 16, 523–531.

[2] Bickel, P. J. (1975), One-step Huber estimates in the linear model, *J. Amer. Statist. Assoc.*, 70, 428–434.

[3] Bilias, Y., Chen, S. and Ying, Z. (2000), Simple resampling methods for censored regression quantiles, *Journal of Econometrics*, 99, 373–386.

[4] Dodge, Y. and Jurečková, J. (1995), Estimation of quantile density function based on regression quantiles, *Statistics and Probability Letters*, 23, 73–78.

[5] Durrieu, G. and Briollais, L. (2009), Sequential design for microarray experiments, *Journal of the American Statistical Association*, 104, 650–660.

[6] Gutenbrunner, C. J., Jurečková, J., Koenker, R. and Portnoy, S. (1993), Tests of linear hypotheses based on regression rank scores, *Journal of non parametric statistics*, 2, 307–333.

[7] He, X. and Shao, Q. (1996), A general Bahadur representation of M-estimators and its application to linear regression with non stochastic designs, *Ann. Statist.*, 24, 2608–2630.

[8] Huber, P. J. (1973), Robust regression : Asymptotics, conjectures and Monte Carlo, *Ann. Statist.*, 1, 799–821.

[9] Huber, P. J. and Ronchetto, E. M. (2009), *Robust Statistics*, J. Wiley, New York.

[10] Hampel, F. R., Ronchetti, E. M., Rousseeuw, J. and Stahel, W. A. (1986), *Robust Statistics*, J. Wiley, New York.

[11] Jurečková, J. and Sen, P. K. (1996), *Robust statistical procedures : Asymptotics and interrelations*, J. Wiley, New York.

[12] Kocherginsky, M., He, X. and Mu, Y. (2005), Practical Confidence Intervals for Regression Quantiles, *Journal of Computational and Graphical Statistics*, 14, 41–55.

[13] Koenker, R. W. and Bassett, G. (1978), Regression Quantiles, *Econometrica*, 46, 33–50.

[14] Koenker, R. (1994), *Confidence Intervals for regression quantiles*, Springer-Verlag, New-York, 349–359

[15] Koenker, R. (1996), *Rank Tests for Linear Models*, Springer-Verlag, New-York.

[16] Parzen, M. I., Wei, L. and Ying, Z. (1994), A resampling method based on pivotal estimating functions, *Biometrika*, 81, 341–350.

[17] Sohn, I., Kim, S., Hwang, C. and Lee, J. W. (2008), New normalization methods using support vector machine quantile regression approach in microarray analysis, *Computational Statistics and Data Analysis*, 52, 4104–4115.

[18] Vapnik, V.N. (1998), *Statistical Learning Theory*, New-York.

[19] Neuwoehner, J., Fenner, K., Escher, B. I. (2009), Physiologiccal Modes of Action of Fluoxetine and its human Metabolites in Algae, *Environmental Science & Technology*, 43, 6830–6837