Estimation de quantiles conditionnels basée sur la quantification optimale sous R

Isabelle Charlier[a,b,c],**Davy Paindaveine**[a,b]and **Jérôme Saracco**[c]

[a]Département de Mathématique
Université Libre de Bruxelles
Boulevard du Triomphe, Campus Plaine, CP210, B-1050 Bruxelles, Belgique
ischarli@ulb.ac.be, dpaindav@ulb.ac.be

[b]ECARES
Avenue F.D. Roosevelt, CP114/04, B-1050 Bruxelles, Belgique

[c]Équipe CQFD et Institut de Mathématiques de Bordeaux
INRIA et Université de Bordeaux
351 Cours de la Libération, 33405 Talence
Jerome.Saracco@math.u-bordeaux1.fr

Mots clefs : Estimation non-paramétrique, Quantile conditionnel, Quantification optimale.

1 Introduction

L'intérêt principal des quantiles conditionnels est de fournir une alternative à la moyenne conditionnelle en représentant de manière plus claire et plus complète l'impact de la covariable X sur la variable dépendante Y.

Soient Y une variable aléatoire réelle et X un vecteur aléatoire de dimension d. Notre procédure d'estimation de ces quantiles conditonnels se base sur la définition suivante du quantile conditionnel d'ordre α de Y sachant $X = x$, noté $q_\alpha(x)$:

$$q_\alpha(x) = \arg\min_{a\in\mathbb{R}} \mathrm{E}[\rho_\alpha(Y-a)|X=x],$$

où $\rho_\alpha(z) = z(\alpha - \mathbb{I}_{[z<0]})$, $\mathbb{I}_A$ désignant l'indicatrice sur l'ensemble A. En pratique, la distribution conditionnelle de Y sachant $X = x$ est inconnue et nous voulons l'estimer à partir d'un échantillon de taille n, $Z^{(n)} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}_{i=1,\ldots,n}$, au moyen des fonctions de quantiles conditionnels. En effet, l'estimation de ces quantiles nous permet de construire des courbes de référence à l'intérieur desquelles se trouvera une certaine proportion d'observations.

Notre procédure d'estimation fonctionne en deux étapes. Tout d'abord, nous remplaçons dans la définition des quantiles conditionnels la covariable X par une version discrétisée dont le support est de taille N. Cette discrétisation est réalisée à l'aide de la quantification optimale en norme L_p et consiste en la projection de X sur un ensemble de N points de $\mathbb{R}^d$ appelé *grille optimale* (voir [1] pour plus de détails). Nous construisons alors un estimateur en projetant la partie en X de l'échantillon $Z^{(n)}$ sur cette grille optimale et en prenant une version empirique de cette approximation :

$$\widehat{q}_\alpha^{N,n}(x) = \arg\min_{a\in\mathbb{R}} \sum_{i=1}^{n} \rho_\alpha(Y_i - a)\mathbb{I}_{[\widehat{X}_i=\hat{x}]},$$

où $\widehat{X}_i$ et $\hat{x}$ sont respectivement la projection de X_i et x sur la grille optimale. Nous avons également défini une version bootstrap de cet estimateur, notée $\bar{q}_{\alpha,B}^{N,n}(x)$, obtenue en réalisant B estimations de ces quantiles sur la base d'échantillons générés avec remise à partir de $Z^{(n)}$ et en moyennant ces B estimations. Le nombre N de quantifieurs peut être choisi optimalement à partir de $Z^{(n)}$ en utilisant une version bootstrap d'un critère de type moindres carrés des écarts entre le vrai quantile et son estimation.

2 Implémentation en R

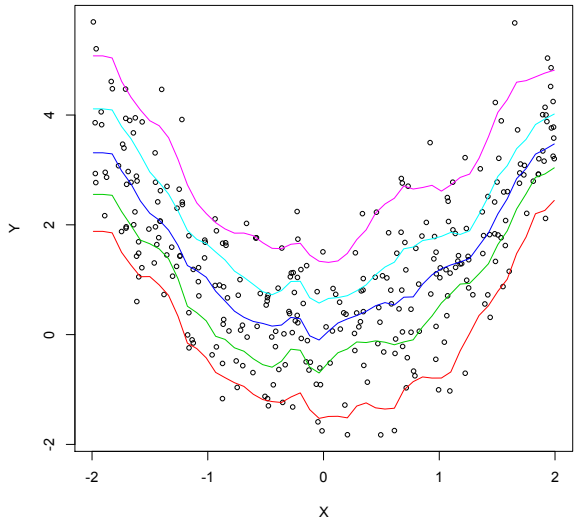
Ces deux estimateurs $\widehat{q}_{\alpha}^{N,n}(x)$ et $\bar{q}_{\alpha,B}^{N,n}(x)$ ont été implémentés dans R au sein d'une même fonction qui se décline en trois versions.

- `QuantifQuantile` : cas $d = 1$, avec sorties graphiques associées en 2D;
- `QuantifQuantile_d2` : cas $d = 2$, avec sorties graphiques associées en 3D;
- `QuantifQuantile_d` : cas général où la covariable X est de dimension d.

Pour chacune des versions, les arguments `X` et `Y` doivent être obligatoirement spécifiés, correspondant respectivement à la partie en X et en Y de l'échantillon. La dimension `d` doit être précisée uniquement dans le cas général. Les arguments suivants peuvent être modifiés par l'utilisateur :

- `x` : grille de valeurs x pour lesquelles on estime $q_\alpha(x)$ (obligatoire pour `QuantifQuantile_d`);
- `alpha` : vecteur d'ordre des quantiles;
- `testN` : une grille de valeurs de N qui seront testées pour en déduire la valeur N^* optimale;
- `p` : l'indice de la norme L_p;
- `B` : le nombre de réplications bootstrap (`B` $= 1$ est équivalent à ne pas faire de bootstrap);
- `tildeB` : le nombre de réplications intervenant dans le choix optimal de N.

Ces fonctions renvoient une liste contenant notamment la valeur optimale N^* sélectionnée par notre procédure, et $\bar{q}_{\alpha,B}^{N^*,n}(x)$ pour chaque valeur de α et x renseignée. Un package contenant ces fonctions est en cours de développement. Pour terminer, l'exemple suivant illustre l'utilisation de la fonction `QuantifQuantile`.




```
n <- 300
X <- runif(n,-2,2)
Y <- X^2 + rnorm(n)
a <- c(0.05,0.25,0.5,0.75,0.95)
t <- seq(10,50,by=5)
x <- seq(min(X),max(X),length=50)
res <- QuantifQuantile(X,Y,alpha=a,testN=t,
x=x,p=2,B=50,tildeB=30)
plot(X,Y,cex=0.7)
for(j in 1 : length(res$alpha))
{lines(res$x,res$hatq_opt[,j],col=(j+1))}
```

Référence

[1] Pagès, G., Printems, J. (2003). Optimal quandratic quantization for numerics : the Gaussian case. *Monte Carlo Methods*, **9**(2), 135-165.