Pour commencer

Sophie Manuel Stéphane Sadio Wiam CHAOUI

Contents

Motivation :

On a une arbre phylogénétique, qu'un branchement évolutif (i.e. la création d'une espèce) apparaît après une durée aléatoire d'une loi fixé μ indépendamment du passé et du futur évolutifs des espèces. On cherche cette loi μ ? sa variance ? sa moyenne ?
(*mots clés*: Arbre phylogénétique , branchement évolutif,loi fixé)

Remarques/Questions :

C'est quoi le but/la question pricipale du projet ?
Pourquoi l'estimation non-paramétrique ?(le lien avec la création d'une espèce et les arbres phylogénétique)
Pourquoi l'estimation de densité à noyau ?
En pratique toute fonction mesurable$\hat{f}$ des data est un estimateur de f,comment donc peut on juger(évaluer) les performances de ces estimateurs afin de choisir un ?(indication : Risque quadratique ($R(\hat{f}, f)$))

Statistique non-paramétrique . Pourquoi ? :

Dans notre cas on a des données observées quantitatives présenter par les successions de branchements qui composent un arbre phylogénétique .
Soit un arbre phylogénétique on cherche a éstimer la fonction f qui donne la durée qui faut pour qu'un branchement évolutif apparaisse.(qui a générer l'echantillon aléatoire)
Discréption formelle : χ=$\{X_1, ..., X_n\}$ un échantillon de variables obsérvés qui ont pour fonction de densité f $\in$ F où F est un espace fonctionnel et a partie des obsérvations (data)$X_1, ...X_n$ on veut éstimer cette fonction densité f sur la quelle on fait le moins d'hypothéses possibles .On a alors le modéle suivant $\{P = P_f, f \in F\}$.(Ce qui revient a une estimation non-paramétrque)

Estimateur de densité à noyau , Pourquoi ? Pourquoi pas d'autres méthodes ?.

Un noyau est une fonction intégrable qu'on note K : $R \to R$ et qui vérifie $\int_R K(u)du = 1$.
Pourquoi l'estimation a noyau est plus intéressante ?
Soient h>0 et K_h : u$\in$ R $\frac{K(\frac{u}{h})}{h}$ alors pour la famille $(K_h)_{h \geq 0}$ on a le résultat de convolution suivant :
$K_h * f$: x $\to$ $\int_R K_h(x - x')f(x')dx'$ tend vers f quand h tend vers 0 .

La densite f peut alors etre éstimer par le produit de convolution $K_h * f$(qui satisfait$K_h * f = E_f[K_h(x - X_1)]$?)
Donc l'estimateur a noyau de f pour un h $>$ 0 fixé est :
$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i)$ $x \in R$

Evaluer un estimateur :

Pour cela il faut définir le rique associé d'une éstimation$\hat{f}$ pour l'estimation de f .
La fonction de risque est :
$R(\hat{f}, f) = E_f[||\hat{f} - f||^2]$(afin de comparer les deux fonctions)
Faut se poser la question sur le choix de la distance (la norme)?
On prend souvent les distance L^p pour p =1,2 ou∞

On veux que le risque soit minimal , tend vers 0 pour un nombre d'observations assez grand
Dans le cas d'un estimateur a noyau :
$R(\hat{f},f) = E_f[||\hat{f}-f||^2] = ||f-K_h*f||^2 + E_f[||\hat{f}-K_h*f||^2]$ (pourquoi on a l'égalité ? reponse en P.S a la fin)
(P.10 et 11 de AN INTRODUCTION TO NONPARAMETRIC ADAPTIVE ESTIMATION)

On retrouve $R(\hat{f},f) \leq ||f-K_h*f||^2 + ||K^|||^2\frac{1}{nh}$
Pour minimiser cette derniere expression le choix de h est trés influent.(Le choix de h est plus crucial pour la qualité de l'estimateur que celui de K)
Un paramétre trop faible provoque l'apparition de détails artificiels sur le graph de l'estimateur(La variance devient trop grande) ,par contre une valeur de h trop grande on aura la majorité des caractéristiques éffacée.

Méthodes adabtives :

D'apres ce qui précede on a introduit la notion de l'estimation de densité a noyau qui dépend d'un paramétre de lissage h .C'est a dire qu'on a bien définit une famille $(\hat{f}_h)_{h\in\beta_n}$ des estimateurs de la vrai fonction densité f.
Comment peut on alors construire un estimateur a risque optimal a partir de cette famille $(\hat{f}_h)_{h\in\beta_n}$ en prenant en cosideration les obsérvations ?(On veut un estimateur adabtive qui donne le meuilleur equilibre biais-variance)

Methode de Goldenshluger-Lepski :

P.S:

L'esimation non-paramétrique ne fait aucune hypothése sur la nature/forme/type de la distribution des variables aléatoire/sur l'appartenance de la fonction densité
Plus que le risque on peux juger un estimateur selon son efficacité

On que :

$E_f||\hat{f}-f||^2 = E_f[||\hat{f}+E_f(\hat{f})-(E_f(\hat{f})-f)||]^2$

$E_f||\hat{f}-f||]^2 = E_f||\hat{f}+E_f(\hat{f})||^2 + E_f||E_f(\hat{f})-f||^2 - 2E_f(<\hat{f}-E_f(\hat{f});E_f(\hat{f})-f>)$
Comme $\hat{f}est$ déterministe :
$2E_f(<\hat{f}-E_f(\hat{f});E_f(\hat{f})-f>) = 2<0, E_f(\hat{f})-f>=0$ Ainsi $||E_f(\hat{f})-f||$ est déterministe :
On obtient :
$R(\hat{f},f) = E_f[||\hat{f}-f||^2] = ||f-K_h*f||^2 + E_f[||\hat{f}-K_h*f||^2]$