Plans d'expérience pour l'estimation de la courbe de concentration et de l'AUC

Mohamad BELOUNI et Karim BENHENNI

Journée de rentrée des Doctorants du département Probabilités-Statistique du LJK, 20 novembre 2012

Laboratoire Jean Kuntzmann, Equipe IPS

Grenoble, France

Ecole Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique

Université Joseph Fourier

Plan de l'exposé

Modèle

Pour une fonction de concentration g définie dans un intervalle d'observation $[t_0, t_n]$, on considère le modèle :

$$X(t_i) \quad = \quad g(t_i) + \varepsilon(t_i) \qquad \forall i = 0, 1, \ldots, n,$$

où $\{t_0, t_1, \ldots, t_n\}$ sont les instants d'observation de X dans l'intervalle $[t_0, t_n]$ et les erreurs $\varepsilon(t_i)$ forment un processus aléatoire autocorrélé de moyenne nulle $\mathbb{E}(\varepsilon(t)) = 0$ et de fonction d'autocovariance connue $\mathbb{E}(\varepsilon(t)\varepsilon(s)) = R(t, s)$.

On fixe les points $t_0 = 0$ et $t_n = 1$, et on s'intéresse au choix optimal des points intermédiaires $\{t_1, \ldots, t_{n-1}\}$ pour l'estimation de l'intégrale (AUC) suivante :

$$I(g) = \int_0^1 g(t)\,\mathrm{d}t.$$

On note l'estimateur linéaire de $I(g)$ par $L_n(g)$. Le but est de sélectionner les temps d'observations $\{t_1^*, \ldots, t_{n-1}^*\}$ qui minimisent l'erreur quadratique moyenne

$$\mathrm{EQM}(L_n(g)) = \mathbb{E}(L_n(g) - I(g))^2.$$

Plans d'échantillonnage réguliers de taille $n+1$:

On suppose que $h(t)$ est une densité positive sur $[0,1]$ avec une fonction de répartition strictement positive

$$H(t) = \int_0^t h(s)\,\mathrm{d}s, \qquad 0 \leq t \leq 1.$$

Le plan d'échantillonnage régulier défini par :

$$T_n = \left\{ t_{n,k} = H^{-1}\left(\frac{k}{n}\right), k = 0, \ldots, n \right\},$$

comprend les bornes de l'intervalle $[0,1]$, $t_{n,0} = 0$ et $t_{n,n} = 1$. Pour une densité uniforme sur $[0,1]$, $h \equiv 1_{[0,1]}$, le plan d'échantillonnage régulier est périodique.

1. **Définition du modèle et plans d'échantillonnages**

2. **Estimation de la fonction de concentration et de l'AUC**

   - Modèle de régression linéaire multiple
   - Estimateur BLUE
   - Estimateur simple et plan d'échantillonnage optimal

3. **Perspectives**

1. **Définition du modèle et plans d'échantillonnages**

2. **Estimation de la fonction de concentration et de l'AUC**

   - Modèle de régression linéaire multiple
   - Estimateur BLUE
   - Estimateur simple et plan d'échantillonnage optimal

3. **Perspectives**

Modèle de régression linéaire multiple

$$X(t) = \boldsymbol{\beta}' \mathbf{f}(t) + \varepsilon(t), \qquad t \in [0,1]\,,$$

où le processus d'erreur $\varepsilon(t)$ est centré et d'autocovariance $\mathbb{E}(\varepsilon(t)\varepsilon(s)) = R(t,s)$,

$\boldsymbol{\beta}' = (\beta_1, \ldots, \beta_q)$ est un vecteur de paramètres inconnus et

$\mathbf{f}'(t) = (f_1(t), \ldots, f_q(t))$ est un vecteur de fonctions de régression connues de la forme

$$f_j(t) = \int_0^1 R(t,s)\varphi_j(s)\,\mathrm{d}s,\ j = 1, \ldots, q \quad t \in [0,1] \tag{1}$$

où φ_j, $j = 1, \ldots, q$, sont des fonctions continues sur $[0,1]$.

Hypothèses

- (H1) Sur la diagonale ($t = s$), on suppose que

$$R^{0,1}(t, t-) = \lim_{s \downarrow t} \partial R(t, s)/\partial s, \;\; R^{0,1}(t, t+) = \lim_{s \uparrow t} \partial R(t, s)/\partial s,$$

existent et sont continues et

$$\alpha(t) = R^{0,1}(t, t-) - R^{0,1}(t, t+),$$

est supposée être strictement positive.

- (H2) En dehors de la diagonale ($t \neq s$) du carré unité, $R(t, s)$ est supposée avoir des dérivées partielles mixtes continues jusqu'à l'ordre deux .

1. **Définition du modèle et plans d'échantillonnages**

2. **Estimation de la fonction de concentration et de l'AUC**

   - Modèle de régression linéaire multiple
   - Estimateur BLUE
   - Estimateur simple et plan d'échantillonnage optimal

3. **Perspectives**

Estimateur BLUE

On veut estimer la surface (AUC) sous la courbe de la fonction de concentration $g(t) = \beta' \boldsymbol{f}(t)$ définie par

$$\text{AUC} = I(g) = \int_0^1 g(t)\,\mathrm{d}t.$$

L'estimateur BLUE de β sur tout l'intervalle d'observation $[0, 1]$ est défini par

$$\hat{\beta} = \boldsymbol{S}^{-1} \int_0^1 X(t)\varphi(t)\,\mathrm{d}t,$$

où $\varphi'(t) = (\varphi_1(t), \ldots, \varphi_q(t))$, et la matrice $\boldsymbol{S}$ de termes généraux

$$s_{ij} = \int_0^1 \int_0^1 \varphi_i(t) R(t,s) \varphi_j(s)\,\mathrm{d}t\,\mathrm{d}s, \quad i,j = 1, \ldots, q$$

est supposée être inversible. On estime alors l'AUC ($I(g)$) par :

$$L(g) = \boldsymbol{z}' \hat{\beta},$$

où $\boldsymbol{z}' = (z_1, , \ldots, z_q); z_j = \int_0^1 f_j(t)\,\mathrm{d}t,\ j = 1, \ldots, q.$

Si $X(t)$ est observé en $n+1$ points d'un plan d'échantillonnage régulier $\boldsymbol{T_n} = \{t_{n,k}\}_0^n$ dans l'intervalle $[0,1]$, alors l'estimateur BLUE $\hat{\boldsymbol{\beta}}_{\boldsymbol{n}}^{\text{blue}}$ de $\boldsymbol{\beta}$ du paramètre $\boldsymbol{\beta}$ défini par :

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{n}}^{\text{blue}} = \boldsymbol{A}_{T_n}^{-1}\boldsymbol{\eta},$$

où

$$\boldsymbol{A}_{T_n} = (a_{rs})_{q\times q}\,,$$

est supposée être inversible et de termes généraux

$a_{rs} = \sum_{i,j=0}^{n} f_r(t_i)\boldsymbol{R}_{\boldsymbol{T_n}}^{-1}(t_i,t_j)f_s(t_j)\ r,s = 1,\ldots,q$ et

$$\boldsymbol{\eta} = (\eta_r)_{q\times 1}\,,$$

où

$$\eta_r = \sum_{i,j=0}^{n} f_r(t_i)\boldsymbol{R}_{\boldsymbol{T_n}}^{-1}(t_i,t_j)X(t_j),\ r = 1,\ldots,q$$

et $\boldsymbol{R}_{\boldsymbol{T_n}} = (\boldsymbol{R}_{\boldsymbol{T_n}}(t_{n,k},t_{n,j}))_{(n+1)\times(n+1)}$ est supposée être inversible pour tout n.

alors on introduit l'estimateur de l'AUC ($I(g)$) par :

$$L_n^{\text{blue}}(g) = \boldsymbol{z}' \hat{\beta}_{\boldsymbol{n}}^{\text{blue}} = \boldsymbol{z}' \boldsymbol{A}_{T_n}^{-1} \eta.$$

Inconvenients :

Les estimateurs BLUE, $\hat{\boldsymbol{\beta}}_{\boldsymbol{n}}^{\text{blue}}$, $L_n^{\text{blue}}(g)$ sont instables car ils utilisent l'inverse de la matrice d'autocovariance.

Remède :

On construit deux estimateurs $\hat{\boldsymbol{\beta}}_{\boldsymbol{n}}^{\text{trap}}(h)$ et $L_n^{\text{trap}}(g)$ qui sont simples et plus stables car ils ne dépendent pas de la matrice d'autocovariance et qui utilisent seulement les observations aux points du plan d'échantillonnage régulier.

1. **Définition du modèle et plans d'échantillonnages**

2. **Estimation de la fonction de concentration et de l'AUC**

   - Modèle de régression linéaire multiple
   - Estimateur BLUE
   - Estimateur simple et plan d'échantillonnage optimal

3. **Perspectives**

Estimateur simple et plan d'échantillonnage optimal

Si $X(t)$ est observé en $n+1$ points d'un plan d'échantillonnage régulier $\boldsymbol{T_n} = \{t_{n,k}\}_0^n$ dans l'intervalle $[0,1]$, alors on construit l'estimateur simple de β par :

$$\hat{\beta}_{\boldsymbol{n}}^{\text{trap}}(h) = \boldsymbol{S}_n^{*-1} \boldsymbol{L_{n,q}}(X),$$

où

$$\boldsymbol{L_{n,q}}(X) = \frac{1}{2n} \sum_{k=0}^{n-1} \left((\frac{\varphi X}{h})(t_k) + (\frac{\varphi X}{h})(t_{k+1}) \right),$$

est un vecteur colonne

$$\boldsymbol{S_n^*} = \left(s_{ij}^*\right)_{q\times q},$$

est supposée être inversible,

$$s_{ij}^* = \frac{1}{2n} \sum_{k=0}^{n-1} \left((\frac{\varphi_i}{h} f_j)(t_k) + (\frac{\varphi_i}{h} f_j)(t_{k+1}) \right), \quad i,j = 1,\ldots,q$$

où $f_j(t), j = 1,\ldots,q$ est définie par (1)

alors on introduit l'estimateur de l'AUC ($I(g)$) par :

$$L_n^{\text{trap}}(g) = \boldsymbol{z}' \hat{\beta}_{\boldsymbol{n}}^{\text{trap}}(h) = \boldsymbol{z}' \boldsymbol{S}_n^{*-1} \boldsymbol{L_{n,q}}(X),$$

où $\boldsymbol{z}' = (z_1, , \ldots, z_q); z_j = \int_0^1 f_j(t)\,\mathrm{d}t,\ j = 1, \ldots, q.$

Théorème

Si les hypothèses $(H1)$ *et* $(H2)$ *sont vérifiées et si on suppose que* $\varphi_i/h, i = 1, \ldots, q$ *est deux fois continûment différentiable, alors l'estimateur* $\hat{\beta}_{\boldsymbol{n}}^{trap}(h)$ *vérifie*

$$\lim_{n \longrightarrow \infty} n^2 \left(EQM(\hat{\beta}_{\boldsymbol{n}}^{trap}(h)) - trace(\boldsymbol{S}^{-1}) \right) = \frac{1}{12} \int_0^1 \alpha(t) \frac{\varphi'(t) \boldsymbol{S}^{-2} \varphi(t)}{h^2(t)} \, \mathrm{d}t.$$

Par ailleurs, l'estimateur $\hat{\beta}_{\boldsymbol{n}}^{trap}(h)$ *avec un plan d'échantillonnage* $\boldsymbol{T}_{\boldsymbol{n}}^{*}$ *engendré par la densité*

$$h^*(t) = \left\{ \alpha(t) \varphi'(t) \boldsymbol{S}^{-2} \varphi(t) \right\}^{1/3} / \int_0^1 \left\{ \alpha(u) \varphi'(u) \boldsymbol{S}^{-2} \varphi(u) \right\}^{1/3} \mathrm{d}u,$$

est asymptotiquement optimal.

Corollaire

L'estimateur $\hat{\beta}_{\boldsymbol{n}}^{trap}(h^*)$, *construit à partir de* $\boldsymbol{T_n^*}$ *et engendré par la densité* $h^*(t)$, *vérifie :*

$$\lim_{n\longrightarrow\infty} n^2 \left(EQM(\hat{\beta}_{\boldsymbol{n}}^{trap}(h^*)) - trace(\boldsymbol{S^{-1}}) \right) =$$

$$\frac{1}{12}\left(\int_0^1 \left\{ \alpha(t)\varphi'(t)\boldsymbol{S^{-2}}\varphi(t) \right\}^{1/3} \mathrm{d}t \right)^3$$

Théorème

Sous les hypothèses $(H1)$ *et* $(H2)$ *et si on suppose que* φ_i/h, $i = 1, \ldots, q$ *est deux fois continûment différentiable, alors l'estimateur* $L_n^{trap}(g)$ *de l'AUC, vérifie*

$$\lim_{n \longrightarrow \infty} n^2 \left(EQM(L_n^{trap}(g)) - \boldsymbol{z}' \boldsymbol{S}^{-1} \boldsymbol{z} \right) = \frac{1}{12} \int_0^1 \alpha(t) \frac{\varphi'(t) \left(\boldsymbol{S}^{-1} \boldsymbol{A} \boldsymbol{S}^{-1} \right) \varphi(t)}{h^2(t)} \, \mathrm{d}t,$$

où $A = \boldsymbol{z}\boldsymbol{z}'$.

Par ailleurs, l'estimateur $L_n^{trap}(g)$ *avec un plan d'échantillonnage* $\boldsymbol{T}_n^*$ *engendré par la densité* $h^*(t)$

$$h^*(t) = \left\{ \alpha(t) \varphi'(t) \left(\boldsymbol{S}^{-1} \boldsymbol{A} \boldsymbol{S}^{-1} \right) \varphi(t) \right\}^{1/3} / \int_0^1 \left\{ \alpha(u) \varphi'(u) \left(\boldsymbol{S}^{-1} \boldsymbol{A} \boldsymbol{S}^{-1} \right) \varphi(u) \right\}^{1/3} \mathrm{d}u.$$

est asymptotiquement optimal.

Corollaire

L'estimateur $L_n^{trap}(g)$, *construit à partir de* $\mathbf{T_n^*}$ *et engendré par la densité* $h^*(t)$, *vérifie :*

$$\lim_{n\longrightarrow\infty} n^2 \left(EQM(L_n^{trap}(g)) - EQM(L(g))\right) =$$

$$\frac{1}{12}\left(\int_0^1 \left\{\alpha(t)\varphi'(t)\left(\mathbf{S^{-1}AS^{-1}}\right)\varphi(t)\right\}^{1/3} \mathrm{d}t\right)^3$$

où $A = \mathbf{zz'}$.

Cas particulier : $q = 1$

On considère d'abord le modèle linéaire simple ($q = 1$)

$$X(t) = \beta f(t) + \varepsilon(t), \qquad t \in [0, 1]$$

où $f(t)$ est de la la forme générale suivante :

$$f(t) = \int_0^1 R(t, s)\varphi(s)\,\mathrm{d}s + \sum_{l=1}^{L} b_l R(t, a_l),$$

où φ est une fonction continue connue sur $[0, 1]$, les constantes $b_l, l = 1, \ldots, L$ sont connues et non nulles et les $a_l, l = 1, \ldots, L$ sont des points connus dans $[0, 1]$.

Si $X(t)$ est observé en $n+1$ points d'un plan d'échantillonnage régulier $\boldsymbol{T_n} = \{t_{n,k}\}_0^n$ engendré par une densité positive $h(t)$, alors on construit l'estimateur simple de β est défini par :

$$\hat{\beta}_n^{\text{trap}} = \frac{L_n(X)}{L_n(f)},$$

où

$$\begin{aligned} L_n(X) &= \frac{1}{2n}\sum_{k=0}^{n-1}\Big((\frac{\varphi}{h}X)(t_{n,k}) + (\frac{\varphi}{h}X)(t_{n,k+1})\Big) + \sum_{l=1}^{L} b_l X(a_l), \\ L_n(f) &= \frac{1}{2n}\sum_{k=0}^{n-1}\Big((\frac{\varphi}{h}f)(t_{n,k}) + (\frac{\varphi}{h}f)(t_{n,k+1})\Big) + \sum_{j=1}^{L} b_j f(a_j). \end{aligned}$$

$\text{Var}(\hat{\beta}_n^{\text{trap}}) \longrightarrow \text{Var}(\hat{\beta}) = s_L^{-2}$ quand $n \to \infty$, où
$s_L^2 \stackrel{\Delta}{=} \int_0^1\int_0^1 \varphi(t)R(t,s)\varphi(s)\,\mathrm{d}t\,\mathrm{d}s + 2\sum_{l=1}^{L} b_l \int_0^1 \varphi(s)R(s,a_l)\,\mathrm{d}s +$
$\sum_{l=1}^{L}\sum_{j=1}^{L} b_l R(a_l,a_j)b_j.$

Théorème

Sous les hypothèses $(H1)$ *et* $(H2)$ *et si on suppose que* φ/h , $i = 1, \ldots, q$ *est deux fois continûment différentiable, alors l'estimateur* $\hat{\beta}_n^{trap}$, *vérifie*

$$\lim_{n \longrightarrow \infty} n^2 \left\{ Var(\hat{\beta}_n^{trap}) - s_L^{-2} \right\} = \frac{s_L^{-4}}{12} \int_0^1 \alpha(t) \frac{\varphi^2(t)}{h^2(t)} \, \mathrm{d}t,$$

Par ailleurs, l'estimateur $\hat{\beta}_n^{trap}$ *avec un plan d'échantillonnage* $\boldsymbol{T_n^*}$ *engendré par la densité* $h^*(t)$

$$h^*(t) = \left\{ \alpha(t)\varphi^2(t) \right\}^{1/3} / \int_0^1 \left\{ \alpha(u)\varphi^2(u) \right\}^{1/3} \mathrm{d}u,$$

est asymptotiquement optimal.

Références :

[1] SACKS, J. AND YLVISAKER, D (1966). Designs for regression problems with correlated errors, *Ann. Math, 37, 66-89.*

[2] CAMBANIS, S. AND SU, Y (1993). Sampling designs for regression coefficient estimation with correlated errors, *Annals of the Institute of Statistical Mathematics, 46,707-722.*

Exemple

On considère le modèle de régression simple

$$X(t) = \beta t + \varepsilon(t), \qquad t \in [0,1]$$

où $\varepsilon(t)$ est un processus d'erreur de type "Ornstein - Uhlenbeck"(Gauss-Markov) et sa fonction d'autocovariance $R(t,s) = \sigma^2 e^{-\lambda|t-s|}$. On montre que $f(t) = t$ peut s'écrire sous la forme suivante

$$f(t) = \int_0^1 R(t,s)\varphi(s)\,\mathrm{d}s + \sum_{l=1}^{L} b_l R(t,a_l),$$

avec

$$\varphi(s) = \lambda s/\left(2\sigma^2\right), L = 2, a_1 = 0, a_2 = 1, b_1 = -1/\left(2\lambda\sigma^2\right),$$

$$b_2 = (\lambda+1)/\left(2\lambda\sigma^2\right).$$

La fonction de sauts (H1) est :

$$\alpha(t) = R^{0,1}(t, t-) - R^{0,1}(t, t+) = 2\lambda\sigma^2 > 0$$

La densité du plan d'échantillonnage (asymptotiquement) optimal $h^*(t)$ est

$$h^*(t) = (5/3)t^{2/3}$$

et les points optimaux sont

$$t^*_{n,k} = \{k/n\}^{3/5}, k = 0, \ldots, n.$$

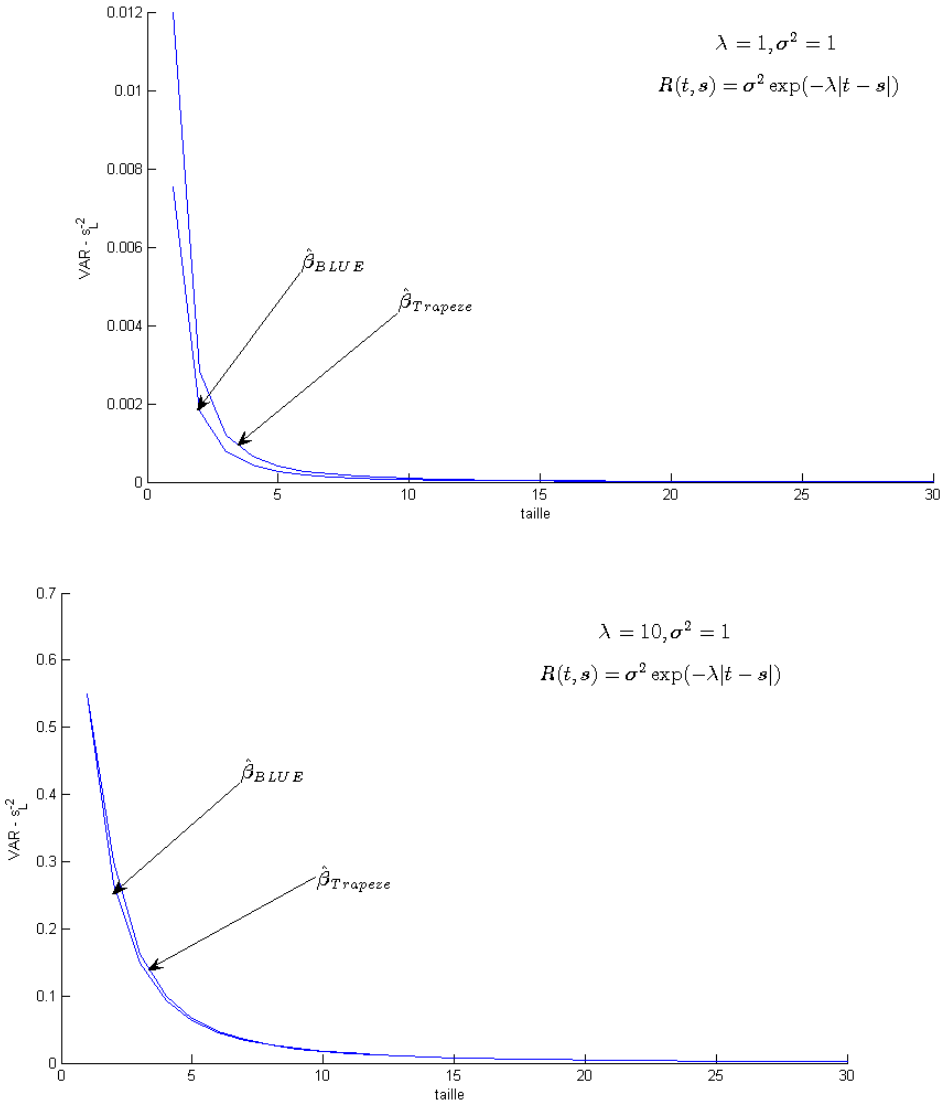

FIG.: $\left(\mathrm{Var}(\hat{\beta}_n) - \mathrm{Var}(\hat{\beta})\right)$ contre la taille de l'échantillon

1. **Définition du modèle et plans d'échantillonnages**

2. **Estimation de la fonction de concentration et de l'AUC**

   - Modèle de régression linéaire multiple
   - Estimateur BLUE
   - Estimateur simple et plan d'échantillonnage optimal

3. **Perspectives**

Perspectives

1. Plans d'échantillonnage optimaux pour l'estimation non-paramétrique de la fonction de concentration et de l'aire sous la courbe de concentration (AUC)
   En cours.
2. Taille fixée et petite : application à la pharmacocinétique (estimation + algorithmes d'optimalité)
   En cours.