IFT 3245
Simulation et modèles
Espérance conditionnelle (EC)
ou Monte Carlo conditionnel (CMC)

Fabian Bastin
DIRO
Université de Montréal

Automne 2016

Un exemple simple

Supposons que l'on veut estimer $\mu = P[Y_1 + \cdots + Y_t > x]$.
Estimateur évident: $X = I[Y_1 + \cdots + Y_t > x]$.
Estimateur CMC: $X_{\mathrm{e},s} = P[Y_1 + \cdots + Y_t > x \mid Y_1, \ldots, Y_s]$, pour $s \leq t$.

Si $s = t$: aucun changement.
Si $s = t - 1$ et les Y_j sont indépendants:
$X_{\mathrm{e},t-1} = 1 - F_t[Y_1 + \cdots + Y_{t-1}]$.
Si $s = 0$: $X_{\mathrm{e},0} = P[Y_1 + \cdots + Y_t > x] = \mu$ (variance réduite à zéro).

En fait plus s est petit, plus la variance est réduite.

Cadre général

Idée: remplacer l'estimateur X par $E[X \mid Z]$ où Z est une autre v.a., ou plus généralement par $E[X \mid \mathcal{G}]$, où $\mathcal{G}$ est une σ-algèbre donnant une information partielle sur X.

L'estimateur CMC s'écrit

$$X_{\mathrm{e}} \stackrel{\text{déf}}{=} E[X \mid \mathcal{G}]$$

et on a

$$E[X_{\mathrm{e}}] = E[E[X \mid \mathcal{G}]] = E[X].$$

Cadre général

De plus,

$$\mathrm{Var}[X] = E[\underbrace{\mathrm{Var}[X \mid \mathcal{G}]}_{\substack{\text{Var. résiduelle} \\ \text{lorsque } \mathcal{G} \text{ est connu} \\ \text{(éliminé par CMC)}}}] + \underbrace{\mathrm{Var}[E[X \mid \mathcal{G}]]}_{\substack{\text{Var. due à la} \\ \text{variation de } \mathcal{G}}}$$

et donc

$$\underline{\mathrm{Var}[X_{\mathrm{e}}]} = \mathrm{Var}[X] - E[\mathrm{Var}[X \mid \mathcal{G}]] \leq \mathrm{Var}[X].$$

(Cas particulier du théorème de Rao-Blackwell.)

Cadre général

Pour minimiser la variance, on doit maximiser $E[\mathrm{Var}[X \mid \mathcal{G}]]$, i.e., $\mathcal{G}$ doit contenir le moins d'information possible.

On peut en effet montrer que $\mathcal{G}_1 \subset \mathcal{G}_2$ implique que $E[\mathrm{Var}[X \mid \mathcal{G}_1]] \geq E[\mathrm{Var}[X \mid \mathcal{G}_2]]$.

Mais moins $\mathcal{G}_1$ contient d'information, plus il est difficile de calculer X_{e}.

On doit donc faire un compromis.

Dans certains cas, X_{e} peut être moins coûteux à calculer que X.

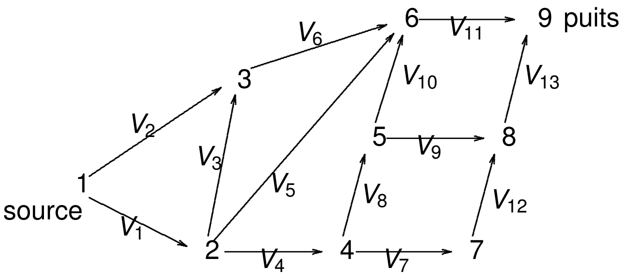
Cas limites:

Si $\mathcal{G}$ ne contient aucune information pertinente à X: $X_{\mathrm{e}} = E[X \mid \mathcal{G}] = E[X] = \mu$.

Si $\mathcal{G}$ permet de calculer X (i.e., X est $\mathcal{G}$-mesurable): $X_{\mathrm{e}} = X$.

Exemple: réseau d'activités stochastique

On veut estimer $\mu = P[T > x]$. Estimateur naïf: $X = I[T > x]$.
Soit $\mathcal{L} \subseteq \mathcal{A}$ un ensemble d'activités (arcs) tel que chaque chemin de la source au puits contient exactement un arc de $\mathcal{L}$.
($\mathcal{L}$ est une coupe orientée.) Exemple: $\mathcal{L} = \{V_4, V_5, V_6\}$.



Soit $\mathcal{B} = \mathcal{A} \setminus \mathcal{L}$ et $\mathcal{G} = \{V_j,\, j \in \mathcal{B}\}$.

$$X_e = P[T > x \mid \{V_j,\, j \in \mathcal{B}\}].$$

On le calcule comme suit.
Pour chaque $l \in \mathcal{L}$, allant disons de a_l à b_l, calculer la longueur α_l du plus long chemin de la source à a_l, puis la longueur β_l du plus long chemin de b_l au puits.
Aucun chemin passant par l est plus long que x ssi $\alpha_l + V_l + \beta_l \leq x$.
Conditionnellement à $\{V_j,\, j \in \mathcal{B}\}$, on a cette condition avec probabilité $P[V_l \leq x - \alpha_l - \beta_l] = F_l[x - \alpha_l - \beta_l]$.
Puisque les V_l sont indépendants, on obtient

$$X_e = 1 - \prod_{l \in \mathcal{L}} F_l[x - \alpha_l - \beta_l].$$

Remarque: cet estimateur peut être moins coûteux à calculer que X.

Estimateur CMC

Avec l'exemple tiré de Avramidis et Wilson (1998), cette méthode réduit la variance environ par un facteur 4.

Dans cet exemple, on peut aussi utiliser les V_j, et peut-être les V_j^2, comme variables de contrôle. On sait que ces variables sont corrélées avec X_{e}.

Exemple: banque

Soit B le nombre d'abandons dans une journée. On peut estimer $E[B]$ par B.
Supposons maintenant que l'on rend les abandons invisibles.
$\mathcal{G}$ represente tout le reste de l'information.

L'estimateur CMC est $B_{\mathrm{e}} = E[B \mid \mathcal{G}]$.
Soit $\lambda(t)$ le taux d'arrivée au temps t;
$N(t)$ le nombre de clients dans le système au temps t;
$P(t) = p_{N(t)}$.
Le processus d'arrivée conditionnel (invisible) des clients qui abandonnent est un processus de Poisson de taux $\{\lambda(t)P(t),\ t \geq 0\}$.