IFT 3245
Simulation et modèles

Fabian Bastin
DIRO
Université de Montréal

Automne 2016

Variables de contrôle (VC)

Idée: exploiter de l'information auxiliaire pour faire une correction à l'estimateur.

On se restreint ici aux VCs linéaires.

Soit X un estimateur sans biais de μ et $\boldsymbol{C} = (C^{(1)}, \ldots, C^{(q)})^T$ des VCs corrélées avec X, d'espérance connue $E[\boldsymbol{C}] = \boldsymbol{\nu} = (\nu^{(1)}, \ldots, \nu^{(q)})^T$.

L'estimateur avec VC est:

$$X_c = X - \boldsymbol{\beta}^T(\boldsymbol{C} - \boldsymbol{\nu}) = X - \sum_{\ell=1}^{q} \beta_\ell (C^{(\ell)} - \nu^{(\ell)}),$$

où $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^T$ (des constantes).
On a $E[X_c] = E[X] = \mu$.

Choix de β

Comment choisir β?

Soient $\mathbf{\Sigma}_{\mathrm{c}} = \mathrm{Cov}[\boldsymbol{C}]$ et
$\mathbf{\Sigma}_{\mathrm{cX}} = (\mathrm{Cov}(X, C^{(1)}), \ldots, \mathrm{Cov}(X, C^{(q)}))^T$.

Hypothèse (CV1)

$Var[X] = \sigma^2 < \infty$, $\mathbf{\Sigma}_{\mathrm{c}}$ et $\mathbf{\Sigma}_{\mathrm{cX}}$ sont finies, et $\mathbf{\Sigma}_{\mathrm{c}}$ est définie positive (et donc inversible).

On a alors

$$\mathrm{Var}[X_{\mathrm{c}}] = \mathrm{Var}[X] + \boldsymbol{\beta}^T \mathbf{\Sigma}_{\mathrm{c}} \boldsymbol{\beta} - 2\boldsymbol{\beta}^T \mathbf{\Sigma}_{\mathrm{cX}}.$$

Choix de β

Pour minimiser par rapport à β, il suffit d'annuler le gradient par rapport à β:

$$0 = \nabla_{\beta} \mathrm{Var}[X_c] = 2\mathbf{\Sigma}_c\beta - 2\mathbf{\Sigma}_{cX}.$$

Le minimum est donc atteint pour

$$\beta = \beta^* = \mathbf{\Sigma}_c^{-1}\mathbf{\Sigma}_{cX},$$

qui donne la variance minimale

$$\mathrm{Var}[X_c] \ = \ (1 - R_{cX}^2)\mathrm{Var}[X] \stackrel{\text{def}}{=} \sigma_c^2,$$

où

$$R_{cX}^2 = \frac{\mathbf{\Sigma}_{cX}^T\mathbf{\Sigma}_c^{-1}\mathbf{\Sigma}_{cX}}{\mathrm{Var}[X]}$$

(le carré du coefficient de corrélation multiple entre $\boldsymbol{C}$ et X) et la variance est réduite par le facteur $1 - R_{cX}^2 = \sigma_c^2/\sigma^2$. Mais avec $\beta \neq \beta^*$, la variance peut augmenter.

Types de VCs

(a) variables internes, basées sur des quantités déjà calculées durant la simulation;

(b) variables externes, obtenues par des simulations additionnelles;

(c) VCs implicites obtenues via une moyenne pondérée.
Soient $X^{(0)}, \ldots, X^{(q)}$ des estimateurs sans biais de μ.
Posons

$$X_c = \sum_{\ell=0}^{q} \beta_\ell X^{(\ell)} = X^{(0)} - \sum_{\ell=1}^{q} \beta_k (X^{(0)} - X^{(\ell)})$$

où $\sum_{\ell=0}^{q} \beta_\ell = 1$.
On peut interpréter $C^{(\ell)} = X^{(0)} - X^{(\ell)}$, $\ell = 1, \ldots, q$, comme VC pour $X = X^{(0)}$.

En pratique, on ne connaît pas $\beta^* = \boldsymbol{\Sigma}_{\mathrm{c}}^{-1}\boldsymbol{\Sigma}_{\mathrm{cX}}$ (parfois $\boldsymbol{\Sigma}_{\mathrm{c}}$, mais jamais $\boldsymbol{\Sigma}_{\mathrm{cX}}$).

On peut l'estimer, disons par $\hat{\beta}_n$, calculé à partir de $(X_1, \boldsymbol{C}_1), \ldots, (X_n, \boldsymbol{C}_n)$.

Posons

$$X_{\mathrm{ce,i}} = X_i - \hat{\beta}_n^T(\boldsymbol{C}_i - \boldsymbol{\nu}),$$

et

$$\bar{X}_{\mathrm{ce,n}} = \bar{X}_n - \hat{\beta}_n^T(\bar{\boldsymbol{C}}_n - \boldsymbol{\nu}).$$

Estimation de β^*: propriétés asymptotiques

Theorème

Sous l'hypothèse CV1, lorsque $n \to \infty$, si $\hat{\beta}_n \xrightarrow{D} \beta^*$, alors

$$\sqrt{n}(\bar{X}_{\mathrm{c,n}} - \bar{X}_{\mathrm{ce,n}}) \xrightarrow{D} 0,$$

$$S^2_{\mathrm{ce,n}} \stackrel{def}{=} \frac{1}{n}\sum_{i=1}^{n}(X_{\mathrm{ce,i}} - \bar{X}_{\mathrm{ce,n}})^2 \xrightarrow{D} \sigma^2_{\mathrm{c}},$$

$$\frac{\sqrt{n}(\bar{X}_{\mathrm{ce,n}} - \mu)}{S_{\mathrm{ce,n}}} \xrightarrow{D} \frac{\sqrt{n}(\bar{X}_{\mathrm{c,n}} - \mu)}{\sigma_{\mathrm{c}}} \xrightarrow{D} N(0,1).$$

On peut utiliser ce théorème pour calculer un IC pour μ, en supposant que $\sqrt{n}(\bar{X}_{\mathrm{ce,n}} - \mu)/S_{\mathrm{ce,n}} \sim N(0,1)$.

Méthode de base:

$$\hat{\beta}_n = \hat{\mathbf{\Sigma}}_{\mathrm{c}}^{-1}\hat{\mathbf{\Sigma}}_{\mathrm{cX}}$$

où les éléments de $\hat{\mathbf{\Sigma}}_{\mathrm{c}}$ et $\hat{\mathbf{\Sigma}}_{\mathrm{cX}}$ sont

$$\begin{aligned}
\hat{\sigma}_{\mathrm{c}}^{(\ell,k)} &= \frac{1}{n-1}\sum_{i=1}^{n}(C_i^{(\ell)} - \bar{C}_n^{(\ell)})(C_i^{(k)} - \bar{C}_n^{(k)}),\\
\hat{\sigma}_{\mathrm{cX}}^{(\ell)} &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)(C_i^{(\ell)} - \bar{C}_n^{(\ell)}).
\end{aligned}$$

Cas multinormal

Dans le cas où $\begin{pmatrix} X_i \\ \boldsymbol{C}_i \end{pmatrix} \sim$ normal, on peut utiliser la théorie de la régression linéaire (avec estimateurs moindres carrés) pour le modèle

$$X = \mu + \boldsymbol{\beta}^T(\boldsymbol{C} - \boldsymbol{\nu}) + \epsilon$$

où $\epsilon \sim N(0, \sigma_\epsilon^2)$.

Si on définit

$$\tilde{S}^2_{\mathrm{ce,n}} = \frac{n}{n-q-1}\left(\frac{1}{n} + \frac{(\bar{\boldsymbol{C}}_n - \boldsymbol{\nu})^T \hat{\boldsymbol{\Sigma}}_{\mathrm{c}}^{-1} (\bar{\boldsymbol{C}}_n - \boldsymbol{\nu})}{n-1}\right) \sum_{i=1}^{n} (X_{\mathrm{ce,i}} - \bar{X}_{\mathrm{ce,n}})^2,$$

où $\hat{\beta}_n$ utilise les covariances empiriques, on a le théorème ci-après.

Cas multinormal

Theorème

Si les $(X_i, \mathbf{C}_i^T)^T$ sont i.i.d normaux, alors

$$
\begin{aligned}
E[\bar{X}_{\mathrm{ce,n}}] &= \mu \qquad \text{(\underline{aucun biais})}, \\
E[\tilde{S}^2_{\mathrm{ce,n}}/n] &= Var[\bar{X}_{\mathrm{ce,n}}] = \frac{n-2}{n-q-2}(1 - R^2_{\mathrm{cX}})Var[\bar{X}_n], \\
\sqrt{n}(\bar{X}_{\mathrm{ce,n}} - \mu)/\tilde{S}_{\mathrm{ce,n}} &\sim Student(n-q-1) \qquad \text{(\underline{loi exacte})}.
\end{aligned}
$$

Permet de calculer un IC avec couverture exacte pour n fini. <u>Facteur d'inflation</u> de la variance $(n-2)/(n-q-2) > 1$ dû à l'estimation de β^*. Si on a déjà q VC, l'ajout d'une nouvelle VC n'est rentable que si la valeur de $(1 - R^2_{\mathrm{cX}})$ est réduite d'une fraction $\geq 1/(n-q-2)$.

Expériences pilotes pour estimer β^*?

Pour avoir un estimateur sans biais de β^*, on peut faire une expérience pilote de n_0 observations, calculer un estimateur $\hat{\beta}_0$ et l'utiliser pour les $n - n_0$ observations restantes. On obtient:

$$\begin{aligned}
\bar{X}_{\mathrm{cp,n}} &= \frac{1}{n - n_0} \sum_{i=n_0+1}^{n} (X_i - \hat{\beta}_0^T (\boldsymbol{C}_i - \boldsymbol{\nu})) \text{ et} \\
S^2_{\mathrm{cp,n}} &= \frac{1}{(n - n_0 - 1)} \sum_{i=n_0+1}^{n} (X_i - \hat{\beta}_0^T (\boldsymbol{C}_i - \boldsymbol{\nu}) - \bar{X}_{\mathrm{cp,n}})^2.
\end{aligned}$$

On a $E[\bar{X}_{\mathrm{cp,n}}] = \mu$ et $E[S^2_{\mathrm{cp,n}}/(n - n_0)] = \mathrm{Var}[\bar{X}_{\mathrm{cp,n}}]$.
Mais sous l'hypothèse de normalité,

$$\frac{\mathrm{Var}[\bar{X}_{\mathrm{cp,n}}]}{\mathrm{Var}[\bar{X}_{\mathrm{ce,n}}]} = \frac{n(n-q-2)(n_0-2)}{(n-n_0)(n-2)(n_0-q-2)} > 1.$$

C'est donc <u>inefficace</u>.