LISA pour BART

Radu Craiu

Département des sciences statistiques
Université de Toronto

Conjointement avec
Reihaneh Entezari (Centre Bosch sur l'intelligence artificielle)
et
Jeffrey Rosenthal (Univ. de Toronto)

SSC, Calgary
Mai, 2019

Plan

Motivation pour LISA

- À cause des développements MCMC, les statisticiens bayésiens ont été pendant plus de 30 ans *libérés au niveau des calculs* lorsqu'ils pensaient à un modèle statistique.
- De grands ensembles de données et/ou des vraisemblances insolubles ont mené le calcul bayésien à la croisée des chemins.
- L'échantillonneur Metropolis-Hastings est un des algorithmes les plus utilisés en MCMC. Il fonctionne de la façon suivante:
  - Étant donné l'état actuel de la chaîne θ, prélever $\xi \sim q(\xi|\theta)$.
  - Accepter ξ avec probabilité min $\left\{1, \frac{\pi(\xi|\mathbf{y})q(\theta|\xi)}{\pi(\theta|\mathbf{y})q(\xi|\theta)}\right\}$.
  - Si ξ est accepté, l'état suivant est ξ, sinon il est (encore) θ.
- Il nécessite le calcul de la vraisemblance à chaque itération ce qui est coûteux quand les données sont volumineuses.

Motivation pour LISA

- Solutions possibles: diviser pour régner, traitement séquentiel, pseudo-marginal, pré-calculer, etc
- D & R: Diviser en lots, $\mathbf{y}^{(1)} \cup \ldots \mathbf{y}^{(K)}$, distribuer l'échantillonnage des K distributions quasi-postérieures

$$\pi_j(\theta) \propto [L_k(\theta|\mathbf{y}^{(j)})]^a [p_j(\theta)]^b$$

  entre K unités de traitements.
- Selon les valeurs de a et b, élaborer des stratégies de recombinaisons pour les échantillons π_j afin de récupérer les caractéristiques des distributions a posteriori.
- Objectif: minimiser la perte d'information comparativement à l'analyse complète a posteriori.

Exemple: Consensus Monte Carlo (Scott et al., 2016)

- Considérer la loi a posteriori complète $\pi(\boldsymbol{\theta}|\mathbf{y}) \propto p(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta})$ où $f(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{N} f(y_i|\boldsymbol{\theta})$
- La loi a posteriori spécifique au lot est défini par

$$\pi_{h,CMC} \propto [p(\boldsymbol{\theta})]^{\frac{1}{K}} f(\mathbf{y}^{(h)}|\boldsymbol{\theta})$$

- Les échantillons MCMC sont recueillis indépendamment de chaque π_h et ils sont combinés en utilisant une moyenne pondérée puisque

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \prod_{h=1}^{K} \pi_{h,CMC}(\boldsymbol{\theta}|\mathbf{y}^{(h)}).$$

- La théorie fonctionne si les distributions a posteriori sont gaussiennes.
- Motivation: CMC ne performe pas bien pour le modèle d'arbres de régression additifs bayésiens (BART).

LISA: objectifs initiaux

- Améliorer l'utilisation de BART pour des données volumineuses.
- Rapprocher la vraisemblance spécifique au lot de la vraisemblance des données entières.
- Définir $\pi_{h,LISA} \propto p(\boldsymbol{\theta})[f(\mathbf{y}^{(h)}|\boldsymbol{\theta})]^K$.
- BF intrinsèque (Berger & Perrichi, JASA '96), clonage de données (Lele & al., JASA '10), robustesse bayésienne (Holmes & Walker, Bmka, '17), etc.

LISA: Ancrer les intuitions

- $\hat{\boldsymbol{\theta}}_{n,L}^{(j)}$ et $\hat{\boldsymbol{\theta}}_{n,C}^{(j)}$ désignent les j-ième modes quasi-à posteriori dans LISA et CMC
- $\hat{l}_{n,L}^{(j)}$ et $\hat{l}_{n,C}^{(j)}$ désignent la dérivée seconde négative pour le j-ième log quasi-postérieure pour LISA et CMC

A1: Il existe $\boldsymbol{\theta}_L, \boldsymbol{\theta}_C$ tels que si nous définissons $\epsilon_{n,L}^{(j)} = |\hat{\boldsymbol{\theta}}_{n,L}^{(j)} - \boldsymbol{\theta}_L|$ et $\epsilon_{n,C}^{(j)} = |\hat{\boldsymbol{\theta}}_{n,C}^{(j)} - \boldsymbol{\theta}_C|$, alors $\max_{1 \leq j \leq K} \epsilon_{n,L}^{(j)} \to 0$ et $\max_{1 \leq j \leq K} \epsilon_{n,C}^{(j)} \to 0$ w.p. 1 quand $n \to \infty$.

A2: $|\hat{l}_{n,L}^{(i)} - \hat{l}_{n,L}^{(j)}| \longrightarrow 0$ et $|\hat{l}_{n,C}^{(i)} - \hat{l}_{n,C}^{(j)}| \to 0$ w.p. 1 $\forall\ i \neq j$ quand $n \to \infty$.

A3: $\pi_{Complet}$, $\pi_{h,LISA}$, et $\pi_{h,CMC}$ sont des distributions unimodales qui possèdent des dérivées continues d'ordre 2.

LISA: Ancrer les intuitions

- Hypothèses:
  - **A1** à **A3** sont satisfaites
  - $\hat{I}_N^{1/2}(\boldsymbol{\theta}_{Complet} - \hat{\boldsymbol{\theta}}_N) \xrightarrow{D} N(0, I)$ quand $N \to \infty$ (K est fixe), où $\boldsymbol{\theta}_{Complet} \sim \pi_{Complet}(\boldsymbol{\theta}|\vec{Y}_N)$ alors

- Alors si K est fixe et $N \to \infty$

$$\hat{I}_N^{1/2}(\boldsymbol{\theta}_{j,L} - \hat{\boldsymbol{\theta}}_N) \xrightarrow{D} N(0, I)$$
$$\hat{I}_N^{1/2}(\boldsymbol{\theta}_{j,C} - \hat{\boldsymbol{\theta}}_N) \xrightarrow{D} N(0, KI), \quad \forall \quad j \in \{1, \ldots, K\}.$$

- La propriété asymptotique indique que les prélèvements dans chaque lot peuvent être utilisés sans pondération.

LISA: Exemple Bernoulli

- Considérons $\mathbf{y}_N = \{y_1, ..., y_N\}$, N variables aléatoires i.i.d. Bernoulli (θ).
- Loi a priori $p(\theta) = \text{Beta}(\alpha, \beta)$ pour le paramètre θ
- Fixer $S = \sum_{i=1}^{N} y_i$ et $S_j = \#$ de 1 dans le j-ième lot. Alors:

COMPLET $\pi_{Complet}(\theta|\mathbf{y}_N)$ est $\text{Beta}(S + \alpha, N - S + \beta)$

CMC: $\pi_{j,CMC}(\theta|\mathbf{y}^{(j)})$ est $\text{Beta}\left(S_j + \frac{\alpha-1}{K} + 1, \frac{N}{K} - S_j + \frac{\beta-1}{K} + 1\right)$

LISA: $\pi_{j,LISA}(\theta|\mathbf{y}^{(j)})$ est $\text{Beta}(S_j K + \alpha, (n - S_j)K + \beta)$

- Si $S_j = S/K$ et $n = N/K$ alors $\pi_{j,LISA}(\theta|\mathbf{y}^{(j)}) = \pi_{Complet}(\theta|\mathbf{y}_N)$
- Aucune pondération requise!

BART - Chipman et al. (AOAS, 2010)

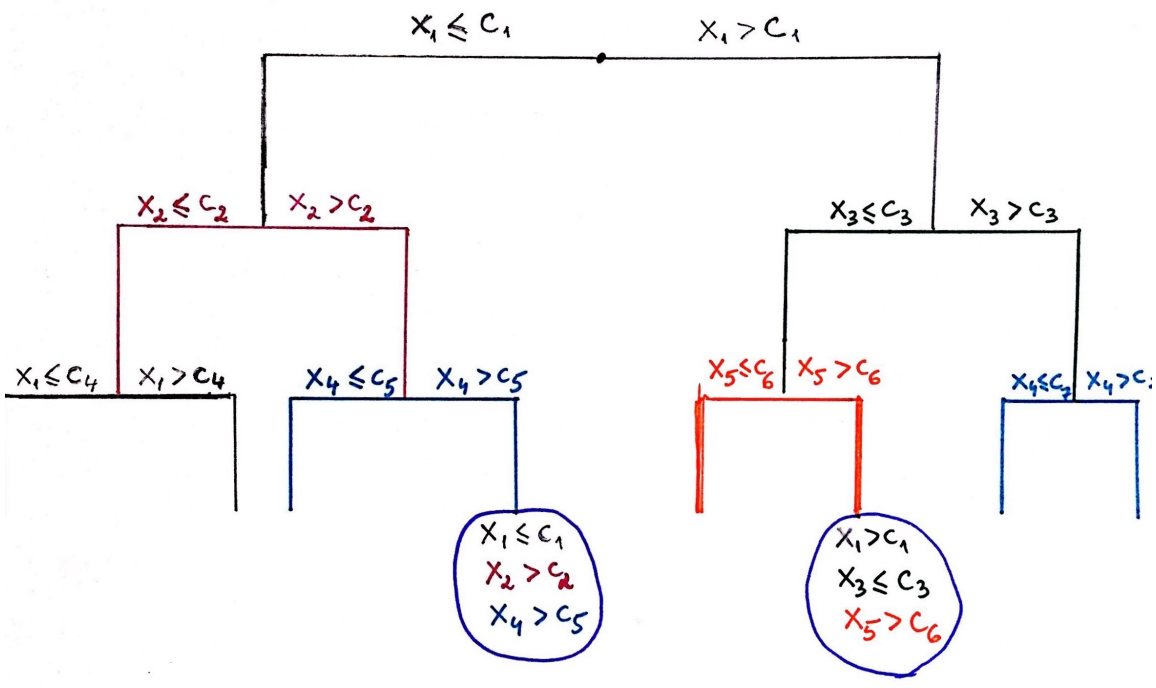
- Approche bayésienne flexible pour régression non-paramétrique
- Cadre de la régression $Y = f(X) + \epsilon$ où le prédicteur $f(X)$ est la somme de (plusieurs) modèles d'arbres de régression

$$f(X) = g_1(X, T_1, M_1) + \ldots + g_m(X, T_m, M_m),$$

avec $\epsilon \sim N(0, \sigma^2)$.

- L'accent est mis sur la prédiction.

BART - Un arbre



- Un arbre T avec b nœuds terminaux qui ont des paramètres $M = (\mu_1, \ldots, \mu_b)$.
- Les règles de séparation $\rightarrow$ partition de l'espace covarié
- BART ajuste une constante pour les données dans chaque nœud marginal résultant en une approximation constante par parties de f.

MCMC & BART

Les a prioris sont:

- $p(\sigma) = \text{Inverse-Gamma}(\frac{\nu}{2}, \frac{\nu\lambda}{2})$,
- $p(\mu_j|\mu_\mu, \sigma_\mu) = N(\mu_\mu, \sigma_\mu)$
- $p(T_j)$, se caractérise par trois éléments:
  - La probabilité qu'un nœud à profondeur $d = 0, 1, ...$ est non-terminal.
  - La distribution de la variable de séparation à chaque nœud intérieur.
  - La distribution de la règle de séparation à chaque nœud intérieur.

Le postérieur BART

$$\pi(\theta) = \pi(\theta|Y,X) \propto \underbrace{\left\{(\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{m} g(x_i;M_j,T_j))^2}\right\}}_{\text{Vraisemblance}} \times$$

$$\underbrace{\left\{\underbrace{(\sigma^2)^{-\frac{\nu}{2}-1} e^{-\frac{\nu\lambda}{2\sigma^2}}}_{\text{Priori de } \sigma^2} \left[\prod_{j=1}^{m} \sigma_\mu^{-b_j} (2\pi)^{-\frac{b_j}{2}} e^{-\frac{1}{2\sigma_\mu^2}\sum_{k=1}^{b_j}(\mu_{kj}-\mu_\mu)^2} p(T_j)\right]\right\}}_{\text{a priori}}. \quad (1)$$

MCMC & BART

- L'échantillonneur MCMC contient les étapes suivantes

Étape S Échantillonner σ étant donné $(T_1, M_1), \ldots, (T_m, M_m)$ utilisant $\sigma^2 \mid (T_1, M_1), ..., (T_m, M_m), Y, X \propto$ Inverse-Gamma(ρ, γ) où $\rho = \frac{\nu+n}{2}$ et $\gamma = \frac{1}{2}\left[\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m} g(x_i; M_j, T_j)\right)^2 + \lambda\nu\right]$.

Étape R Pour $1 \leq j \leq m$, échantillonner (T_j, M_j) étant donné $T_{-j}, M_{-j}, X, \mathbf{y}, \sigma$.

Étape R

- Échantillonner $T_j|R_j, \sigma$, où $R_j = y - \sum_{k \neq j} g(x; M_k, T_k)$ utilise Metropolis-Hastings pour **GRANDIR**, **ÉLAGUER** et **CHANGER**

- Supposons que nous proposons T_*, alors le taux d'acceptation sera:

$$r = \underbrace{\frac{P(T_* \to T)}{P(T \to T_*)}}_{\text{ratio de transition}} \times \underbrace{\frac{P(R \mid T_*, \sigma^2)}{P(R \mid T, \sigma^2)}}_{\text{ratio de vraisemblance}} \times \underbrace{\frac{P(T_*)}{P(T)}}_{\text{ratio de la structure de l'arbre}}$$

- Échantillonner $M_j|T_j, R_j, \sigma$ utilisant

$$\mu_{ij} \mid T_j, R_j, \sigma \sim \mathcal{N}\left(\frac{\frac{\sigma^2}{\sigma_\mu^2}\ \mu_\mu + n_i \bar{R}_{j(i)}}{\frac{\sigma^2}{\sigma_\mu^2} + n_i}, \frac{\sigma^2}{\frac{\sigma^2}{\sigma_\mu^2} + n_i}\right), \text{ où } \bar{R}_{j(i)}$$

désigne la moyenne résiduelle (calculée sans arbre j) au nœud terminal i avec un nombre total d'observations n_i.

LISA & BART

- $f(x) = 10\sin(\pi x_1 x_2) + 20(x_3 - 0,5)^2 + 10x_4 + 5x_5$ avec $N = 20\,000$, $K = 30$, $\sigma = 3$.
- Comparer LISA & *Machine unique*:
  - Les arbres ont tendance à être plus gros $\rightarrow$ Moins de données dans chaque nœud terminal
  - σ est fortement sous-estimé
  - Des taux d'acceptation plus faibles pour les mouvements des arbres.

Méthode	Nœuds	Moy $\hat{\sigma}^2$	95% IC pour σ^2
LISA (*poids unif*)	55	0,001	[0,0009 ; 0,0011]
MachineUnique	7	9,04	[8,85 ; 9,21]

Intermezzo: LISA & Exemple de régression normale

- Soit $Y = X\beta + \epsilon$, $\beta \in \mathbf{R}^p$, $X \in \mathbf{R}^{N \times p}$ et $Y, \epsilon \in \mathbf{R}^N$ avec $\epsilon \sim N(0, \sigma^2 \mathbf{I}_N)$.
- Soit l'a priori de Jeffrey $p(\beta, \sigma^2) \propto 1/\sigma^2$

Intermezzo: LISA & Exemple de régression normale

COMPLET

$$\sigma^2 \sim \text{Inverse-Gamma}\left(\frac{N-p}{2}, \frac{s^2(N-p)}{2}\right)$$

$$\beta|\sigma^2 \sim N(\hat{\beta}, \sigma^2(X^TX)^{-1})$$

avec $\hat{\beta} = (X^TX)^{-1}X^TY$ et
$s^2 = \frac{(Y-X\hat{\beta})^T(Y-X\hat{\beta})}{N-p}$.

$E[\beta_{Complet}|Y,X] = (X^TX)^{-1}X^TY$ et
$\text{Var}(\beta_{Complet}|Y,X) = (X^TX)^{-1}\frac{(N-p)/2}{(N-p)/2-1}s^2 = (X^TX)^{-1}s^2 + O(N^{-1})$.

LISA

$$\sigma^2 \sim \text{Inverse-Gamma}\left(\frac{N-p}{2}, \frac{Ks_j^2(n-p)}{2}\right)$$

$$\beta|\sigma^2 \sim N\left(\hat{\beta}_j, \frac{\sigma^2}{K}(X^{(j)\ T}X^{(j)})^{-1}\right),$$

$E[\beta|Y^{(j)},X^{(j)}] = \hat{\beta}_j$ et
$\text{Var}(\beta|Y^{(j)},X^{(j)}) = (X^{(j)\ T}X^{(j)})^{-1}\frac{s_j^2(n-p)/2}{(N-p)/2-1} = (X^{(j)\ T}X^{(j)})^{-1}\frac{s_j^2(n-p)}{(N-p)} + O(N^{-1})$.

LISA: Exemple de régression normale

- Afin de combiner les échantillons quasi-postérieurs nous proposons l'utilisation de la moyenne pondérée

$$\beta_{LISA} = (\sum_{j=1}^{K} W_j)^{-1} \sum_{j=1}^{K} W_j \beta_j,$$

où $\beta_j \sim \pi_j(\beta|Y^{(j)}, X^{(j)})$ et $W_j = \frac{X^{(j)\,T} X^{(j)}}{\sigma^2}$

$$\text{Alors } E[\beta_{LISA}|Y, X] = \hat{\beta} = (X^T X)^{-1} X^T Y, \text{ et}$$

$$\begin{aligned} \mathrm{Var}(\beta_{LISA}|Y, X) &= (X^T X)^{-1} \frac{n-p}{N-p} \left[\sum_{j=1}^{K} s_j^2 (X^{(j)\,T} X^{(j)}) \right] (X^T X)^{-1} \\ &\approx (X^T X)^{-1} \frac{n-p}{N-p} s^2 \end{aligned}$$

- Modification requise!

LISA: Exemple de régression normale

LISA

$$\begin{aligned} \sigma^{\mathbf{2}} &\sim \text{Inverse-Gamma}\left(\frac{N-p}{2}, \frac{K s_j^{\mathbf{2}}(n-p)}{2}\right) \\ \beta|\sigma^{\mathbf{2}} &\sim N\left(\hat{\beta}_j, \frac{\sigma^{\mathbf{2}}}{K}(X^{(j)\ T}X^{(j)})^{-\mathbf{1}}\right) \\ w_j &\propto \mathbf{1} \end{aligned}$$

LISA Mod

$$\begin{aligned} \sigma^{\mathbf{2}} &\sim \text{Inverse-Gamma}\left(\frac{N-p}{2}, \frac{K s_j^{\mathbf{2}}(n-p)}{2}\right) \\ \tilde{\sigma} &= \sqrt{K}\sigma \\ \beta|\tilde{\sigma}^{\mathbf{2}} &\sim N\left(\hat{\beta}_j, \frac{\tilde{\sigma}^{\mathbf{2}}}{K}(X^{(j)\ T}X^{(j)})^{-\mathbf{1}}\right) \\ w_j &\propto (X^{(j)\ T}X^{(j)})[s_j^{\mathbf{2}}]^{-\mathbf{1}} = \widehat{Var}(\hat{\beta}_j)^{-\mathbf{1}} \end{aligned}$$

LISA modifié & BART

- Introduire une étape intermédiaire entre **Étape S** et **Étape R** dans l'algorithme MCMC pour Lisa.
- Ajuster les prélèvements de σ , c'est-à-dire fixer $\tilde{\sigma}_j = \sqrt{K}\sigma_j$
- Les échantillons des lots j ont des pondérations $\propto \hat{\sigma}_j^{-2}$

LISA modifié & BART

Table: Comparer RMSE formé & testé, tailles des arbres, et $\hat{\sigma}^2$ moyen après rodage avec 95% IC dans chaque méthode pour $K = 30$ avec la MachineUnique BART.

Méthode	RMSEformé	RMSEtesté	Nœuds d'arbres	Moy $\hat{\sigma}^2$	95% CI pour σ^2
CMC	2,73	2,94	602	1,91	[1,45 ; 2,88]
LISA	1,18	1,19	55	0,001	[0,0009 ; 0,0011]
LISAmod	0,57	0,59	7	7,97	[7,87 ; 8,08]
MachineUnique	0,55	0,56	7	9,04	[8,85 ; 9,21]

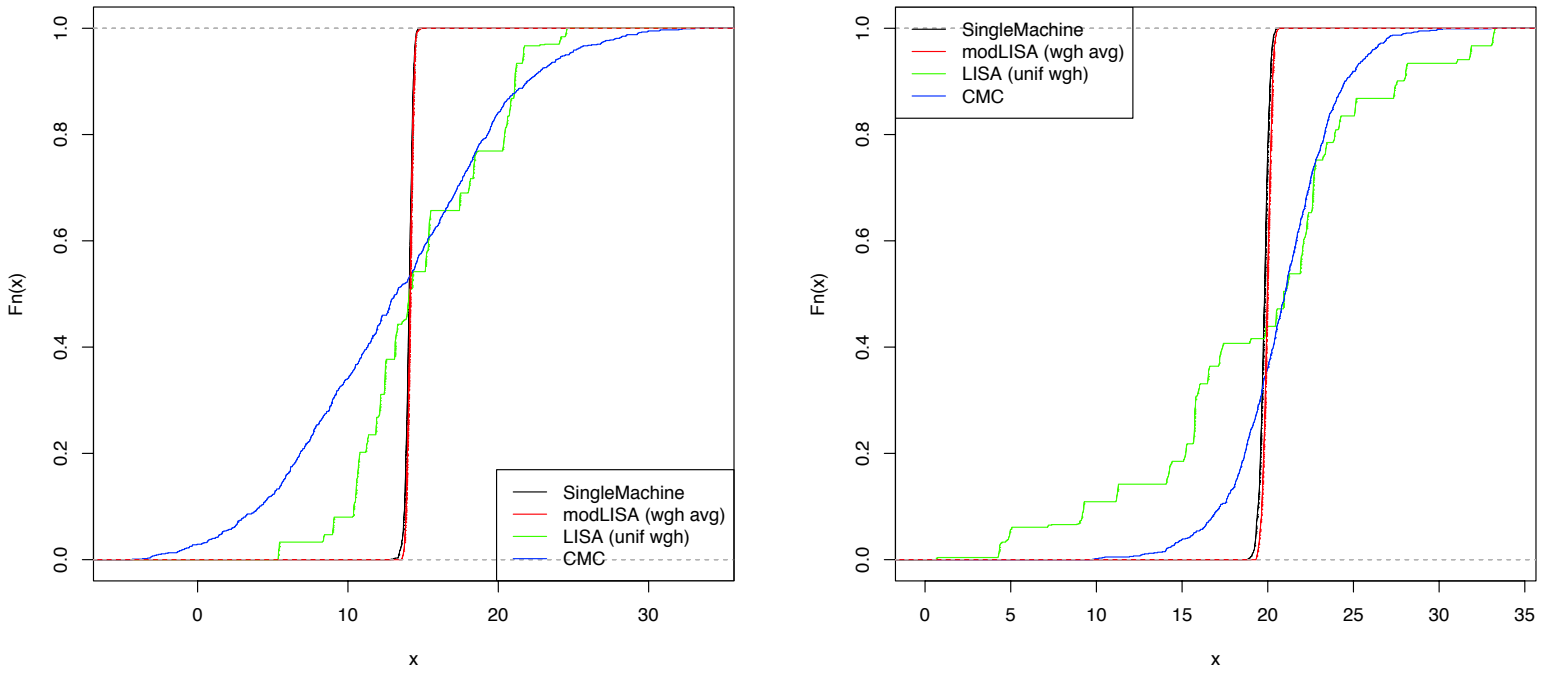
LISA modifié & BART

Taux d'acceptation moyens des mouvements d'arbres proposés.

Méthode	GRANDIR	ÉLAGUER	CHANGER
CMC	21%	0,03%	34%
LISA	1,8%	0,5%	1,6%
LISAmod	20%	26%	19%
MachineUnique	9%	10%	6%

LISA modifié & BART

CDF empirique pour $\hat{f}(x)$



Gauche: Test $f(x^*) = 14,4$; Droite: Formation $f(x) = 19,8$

LISA modifié & BART

Méthode	Temps moy par itération (secs)	Accéléré
CMC	11,99	31%
LISA	5,04	71%
LISAmod	1,81	90%
MachineUnique	17,28	—

Durée pour CMC, LISA, LISAmod et MachineUnique quand $K = 30$.

Couverture de l'intervalle

- Considérons deux types d'intervalles:
  - Soit $\hat{J}_{y_i}$ l'intervalle de prédiction (PI) $1-\alpha$ pour y_i
  - La couverture pour $\hat{J}_{y_i}$ est donnée par la moyenne des données formées/testées

$$\frac{\#\{\tilde{y}_j \in \hat{J}_{y_i} : \ \tilde{y}_j \overset{iid}{\sim} N(f(x_i), \sigma^2), \ 1 \leq j \leq 1000\}}{1000}.$$

  - Couverture de l'intervalle de Bayes (IC)

$$\frac{\#\{f(x_i) \in \hat{I}_{f(x_i)} : \ 1 \leq i \leq N\}}{N}$$

  où $\hat{I}_f(x_i)$ est le IC pour $f(x_i)$.
- Les deux sont considérés pour les données testées et formées.

Couverture de l'intervalle

Méthode	Int de prédiction		Int de Bayes	
	Former	Tester	Former	Tester
CMC	45,71 %	47,83 %	81,95 %	99,99 %
LISA	1,54 %	1,54 %	100 %	100 %
modLISA	92,93 %	92,91 %	60,88 %	58,45 %
MachineUnique	94,67 %	94,65 %	71,58 %	71,54 %

- IP sont influencés par $\hat{\sigma}^2$ et $\widehat{\mathrm{Var}}(\hat{f}(x))$.
- IC sont influencés par $\widehat{\mathrm{Var}}(\hat{f}(x))$.

Modèle alternatif

$$f(x) = \mathbf{1}_{[0;0,2)}(x_1)+2\cdot\mathbf{1}_{[0,2;0,4)}(x_1)+3\cdot\mathbf{1}_{[0,4;0,6)}(x_1)+4\cdot\mathbf{1}_{[0,6;0,8)}(x_1)+5\cdot\mathbf{1}_{[0,8;1)}(x_1)$$

Méthode	Test RMSE	Test Bayes
CMC	1,35	100 %
LISA (*poids unif*)	0,94	100 %
modLISA (*moy pond*)	0,24	90,16 %
MachineUnique	0,15	98,76 %

Données sur le logement

- Les données sont composées de variables sur la population et les logements.
- Prédire le revenu total d'une personne en se basant sur des variables telles que le sexe, l'âge, l'éducation (au moins un diplôme de baccalauréat), la catégorie de travailleur, l'état vivant et le statut de citoyenneté.
- $N = 437\,297$, $K = 100$, la taille d'échantillon Monte Carlo est $M = 1500$, temps > 1 jour pour *Machine unique*.

Méthode	TestRMSE	Moy $\hat{\sigma}^2$	Nœuds d'arbre	Accéléré
LISAmod (moy pond)	0,71	0,488	7	90%
MachineUnique	0,70	0,485	23	–

Conclusions

- LISAmod combiné avec des chaînes de mélanges améliorées pour BART (Pratola, BA 2016) présente des gains comparables.
- Malgré des propriétés asymptotiques intéressantes, un ajustement des échantillonneurs de type LISA est quand même nécessaire.
- La validation théorique peut s'appuyer sur un MCMC approximatif & bruyant et des erreurs dûes aux perturbations (p. ex., Mithrophanov 2005, Pillai et Smith 2015, Johndrow et al. 2017, Negrea et Rosenthal 2017).
- Des questions importantes concernant le plan d'échantillonnage des lots $\rightarrow$ Extension aux cas non-iid est une orientation future importante.
- Parmi les alternatives prometteuses on retrouve l'utilisation de chaînes de Markov non-réversibles ou d'ensembles de base.