Bootstrap et Processus Empiriques

Cours de Statistique Mathématique Avancée

Plan du Cours

- Introduction au Bootstrap
- Processus Empiriques et FRE
- Théorème de Donsker
- Bootstrap du Processus Empirique
- Consistance du Bootstrap
- Applications aux Tests Statistiques
- Bootstrap Lisse et Extensions
- Implémentation et Validation

Principe du Bootstrap

Idée Fondamentale (Efron, 1979)

Le **Bootstrap** est une méthode de rééchantillonnage qui consiste à générer des échantillons bootstrap à partir d'un échantillon observé $X_1, \ldots, X_n \sim F$ en utilisant **le rééchantillonnage avec remise**. Cela permet d'estimer la distribution d'une statistique sans faire d'hypothèses fortes sur la distribution sous-jacente F.
L'idée centrale est de générer des **échantillons bootstrap** $X_1^*, \ldots, X_n^*$ à partir de la **fonction de répartition empirique** F_n:

$$X_1^*, \ldots, X_n^* \sim F_n$$

où F_n est la fonction de répartition empirique de l'échantillon observé.

Applications

Les applications du bootstrap comprennent :

- Estimation de **biais** et **variance** de statistiques.
- Construction d'**intervalles de confiance**.
- Tests d'**hypothèses** (ex : test de Kolmogorov-Smirnov)

Algorithme Bootstrap de Base

Étapes Fondamentales

1. **Échantillon original**: $X_1, \ldots, X_n \sim F$
2. **Estimateur**: $\hat{\theta}_n = T(X_1, \ldots, X_n)$
3. **Rééchantillonnage**: $X_1^*, \ldots, X_n^* \sim F_n$
4. **Estimateur bootstrap**: $\hat{\theta}_n^* = T(X_1^*, \ldots, X_n^*)$
5. **Répéter** B fois et approximer la distribution.

Distribution Bootstrap

La fonction de répartition empirique bootstrap de l'estimateur $\hat{\theta}_n$ est définie par :

$$\hat{F}^*_{\hat{\theta}}(t) = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}_{\{\hat{\theta}_n^{*(b)} \leq t\}}$$

où B est le nombre de répliques bootstrap, et $\hat{\theta}_n^{*(b)}$ est l'estimateur obtenu à partir du b-ème échantillon bootstrap.

Fonction de Répartition Empirique (FRE)

Definition (FRE)

La **fonction de répartition empirique** (FRE) pour un échantillon $X_1, \ldots, X_n$ est définie par :

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq t\}}$$

Cela donne la proportion des valeurs de l'échantillon inférieures ou égales à t.

Propriétés

- **Estimateur non biaisé** : $\mathbb{E}[F_n(t)] = F(t)$, où $F(t)$ est la fonction de répartition théorique.
- **Consistance forte** : $F_n(t) \xrightarrow{p.s.} F(t)$ (convergence presque sûre).
- **Distribution** : $nF_n(t) \sim \text{Binomial}(n, F(t))$, ce qui signifie que $F_n(t)$ suit une distribution binomiale pour un échantillon de taille n.

Processus Empirique

Definition (Processus Empirique Standard)

Le **processus empirique standard** est défini comme :

$$\mathbb{G}_n(t) = \sqrt{n}(F_n(t) - F(t))$$

où $F_n(t)$ est la fonction de répartition empirique et $F(t)$ la fonction de répartition théorique.

Propriétés Ponctuelles

Pour un t fixé, $\mathbb{G}_n(t)$ converge en loi vers une **distribution normale** :

$$\mathbb{G}_n(t) \xrightarrow{(d)} \mathcal{N}(0, F(t)(1 - F(t)))$$

ce qui signifie que le processus $\mathbb{G}_n$ suit une **distribution gaussienne** pour chaque point t.

Limitation

L'approche ponctuelle ignore la structure de corrélation du processus, ce qui peut

Convergence Fonctionnelle

Espace des Fonctions

La convergence en loi des processus empiriques se fait dans l'espace des fonctions $D[0,1]$ (espace des fonctions continues à gauche et discontinues à droite), avec la **topologie de Skorokhod**.

Definition (Convergence en Loi Fonctionnelle)

On dit que $\mathbb{G}_n \xrightarrow{(d)} \mathbb{G}$ dans $D[0,1]$ si pour toute fonction continue bornée Φ :

$$\mathbb{E}[\Phi(\mathbb{G}_n)] \to \mathbb{E}[\Phi(\mathbb{G})]$$

Cela signifie que la convergence des processus empiriques est mesurée sur **toutes les trajectoires** du processus.

Théorème de Donsker

Theorem (Donsker, 1952)

Soit $X_1, \ldots, X_n$ des variables aléatoires indépendantes et identiquement distribuées (i.i.d.) de loi uniforme sur $[0,1]$. Alors :

$$\boxed{\mathbb{G}_n \xrightarrow{(d)} \mathbb{B}^0}$$

*dans $D[0,1]$, où $\mathbb{B}^0$ est un **pont brownien**.*

Pont Brownien

- $\mathbb{B}^0(0) = \mathbb{B}^0(1) = 0$
- $\mathrm{Cov}(\mathbb{B}^0(s), \mathbb{B}^0(t)) = s \wedge t - st$
- Trajectoires continues (presque sûrement)

Propriété d'Invariance

Universalité de la Limite

Le **pont brownien** $\mathbb{B}^0$ est une **loi limite universelle** pour les processus empiriques, indépendamment de la loi F des données.

Theorem (Cas Général)

Pour toute fonction de répartition continue F, la transformation du processus empirique donne :

$$\mathbb{G}_n \xrightarrow{(d)} \mathbb{B}^0 \circ F$$

après transformation probante $U_i = F(X_i) \sim U[0,1]$.

Processus Empirique Bootstrap

Definition (Processus Empirique Bootstrap)

Le **processus empirique bootstrap** est défini comme :

$$\mathbb{G}_n^*(t) = \sqrt{n}(F_n^*(t) - F_n(t))$$

où $F_n^*(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i^* \leq t\}}$, et X_i^* est un échantillon bootstrap tiré de F_n.

Interprétation

- $\mathbb{G}_n^*$ est la version bootstrap de $\mathbb{G}_n$.
- F_n est utilisé comme approximation de la fonction de répartition F inconnue.
- La normalisation reste similaire à celle de $\mathbb{G}_n$.

Consistance du Bootstrap

Theorem (Giné-Zinn, 1990)

Soit $X_1, \ldots, X_n$ i.i.d. de loi F. Alors :

$$\sup_{h \in \mathcal{BL}_1} |\mathbb{E}^*[h(\mathbb{G}_n^*)] - \mathbb{E}[h(\mathbb{B}^0 \circ F)]| \xrightarrow{P} 0$$

où $\mathcal{BL}_1$ est la classe des fonctions 1-Lipschitz bornées.

Interprétation

La loi bootstrap de $\mathbb{G}_n^*$ approxime la loi limite de $\mathbb{G}_n$:

$$\mathcal{L}(\mathbb{G}_n^* | X_1, \ldots, X_n) \approx \mathcal{L}(\mathbb{B}^0 \circ F)$$

Conditions de Consistance

Theorem (Conditions Suffisantes)

Le bootstrap est consistant sous les conditions suivantes :

- *F est continue.*
- *La classe $\{\mathbf{1}_{\{x \leq t\}}, t \in \mathbb{R}\}$ est Donsker.*
- *Les observations sont i.i.d.*

Cas Pathologiques

Le bootstrap peut échouer pour :

- Distributions discrètes.
- Statistiques non lisses.
- Très petits échantillons.

Test de Kolmogorov-Smirnov Bootstrap

Problème

Tester $H_0 : F = F_0$ vs $H_1 : F \neq F_0$

Algorithme Bootstrap

1. Calculer $D_n = \sup_t |F_n(t) - F_0(t)|$
2. Générer B échantillons bootstrap sous H_0
3. Calculer $D_n^* = \sup_t |F_n^*(t) - F_0(t)|$ pour chaque réplique
4. p-value $= \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}_{\{D_n^{*(b)} > D_n\}}$

Theorem (Consistance)

Sous H_0: $\sqrt{n} D_n^* \xrightarrow{(d)} \sup_t |\mathbb{B}^0(t)|$