Une Introduction au Bootstrap

Patrice Bertail

CREST, Laboratoire de Statistique

Université Paris X, Nanterre

1) Idées de bases de la méthode (Origine, algorithme, Intervalles de confiance, propriétés)
2) Extensions dans le cas dépendant

17 juin 2003

Patrice.Bertail@ensae.fr

Les origines du Bootstrap

To pull oneself up with its own Bootstrap :

Se soulever soit même en tirant sur ses lacets : Efron (1979)(1981).

Idée très ancienne : utilisation répétée des données (sous-échantillons) pour estimer la variabilité d'une statistique (Inde, 30's, Jackknife, 50's).

Méthode de calcul intensif fortement liée au développement de l'informatique

Rééchantillonnage

Deux idées : -Estimation non-paramétrique (toute l'information est dans les données) P_n

-Algorithme de Monte-Carlo

- $(x_1, x_2, x_3, \ldots, x_n)$ données i.i.d. Loi F

Ex. $(2.2, 3.2, 3.4, 2.5, 5.2, 3.3, 3.0, 3.7, 4.1, 3.1)$

Statistique d'intérêt

$$T_n(x_1, \ldots, x_n) = n^{-1} \sum_{i=1}^{n} x_i \underset{Exemple}{=} 3.37$$

Algorithme

Données :

$(2.2,\ 3.2\ ,\ 3.4\ ,\ 2.5\ ,\ 5.2\ ,\ 3.3\ ,\ 3.0\ ,\ 3.7\ ,\ 4.1\ ,\ 3.1$

Répeter B fois (B grand=999)

Tirage uniforme avec remise (poids $1/n$)

dans les données $(x_1^*, x_2^*, x_3^*, \ldots, x_n^*)$

$\underset{Exemple}{=}$ $(5.2,\ 3.1,\ 3.1\ ,\ 3.2,\ 2.5,\ 2.5,\ 3.7,\ 3.3\ ,\ 3.7, 2.$

Calcul de la valeur de la statistique

$$T_n^{*(1)}(x_1^*, x_2^*, x_3^*, \ldots, x_n^*) \underset{Exemple}{=} 3.28$$

Fin

Variance et distribution Bootstrap

Résultats : Collections de B valeurs $T_n^{*(i)},\ i = 1, ..., B$

$$3.17,\ 3.40,\ 3.28,\ 3.27,\ 3.37, ...,$$

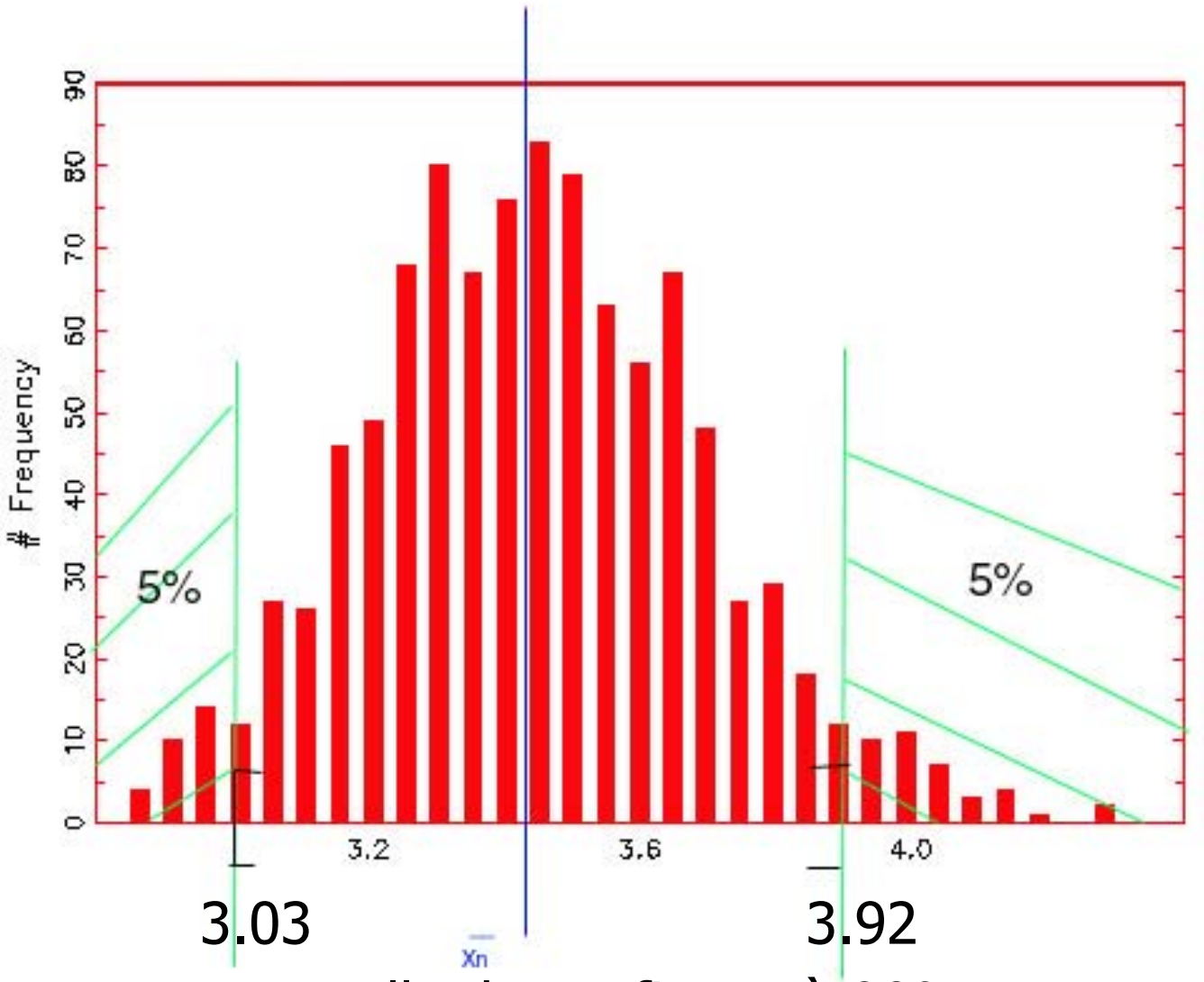
1er type d'utilisation : La variance empirique de ces valeurs Est un estimateur de la volatilité de la statistique de départ

Exemple : 0.073 (B=999), Calcul direct 0.070

2éme type d'utilisation : La distribution empirique de ces valeurs est un estimateur de la vraie distribution

→ Construction d'intervalles de confiance

Intervalles de confiance Bootstrap



Intervalle de confiance à 90%

Propriétés

- Facilité d'utilisation (calcul parallèlisable)
- Mêmes propriétés que les intervalles de confiance asymptotique usuels.
- Possibilité d'avoir des améliorations spectaculaires en travaillant avec des statistiques standardisées (t-percentile).
- Se généralise aux données multidimensionnelles, censurées, et à toutes statistiques « lisses » (fonctionnelles Hadamard différentiables).

Variantes du Bootstrap

Bootstrap paramétrique

Si un modèle pour les données est disponible, F_θ
Ex : modèle gaussien $N(m, \sigma^2),\ \theta = (m, \sigma^2)$

1ère étape : Estimation préalable du modèle : $\widehat{\theta_n}$
2ème étape : Rééchantillonnage dans $F_{\widehat{\theta_n}}$

Ex: génération de données dans N(3.7,0.07)

Bootstrap lissé ou perturbé

Tirage dans les données auxquelles on rajoute une perturbation
Gaussienne N(O,n^{-2}) -> Meilleur pour des fractiles (VaR)

Cas non indépendant

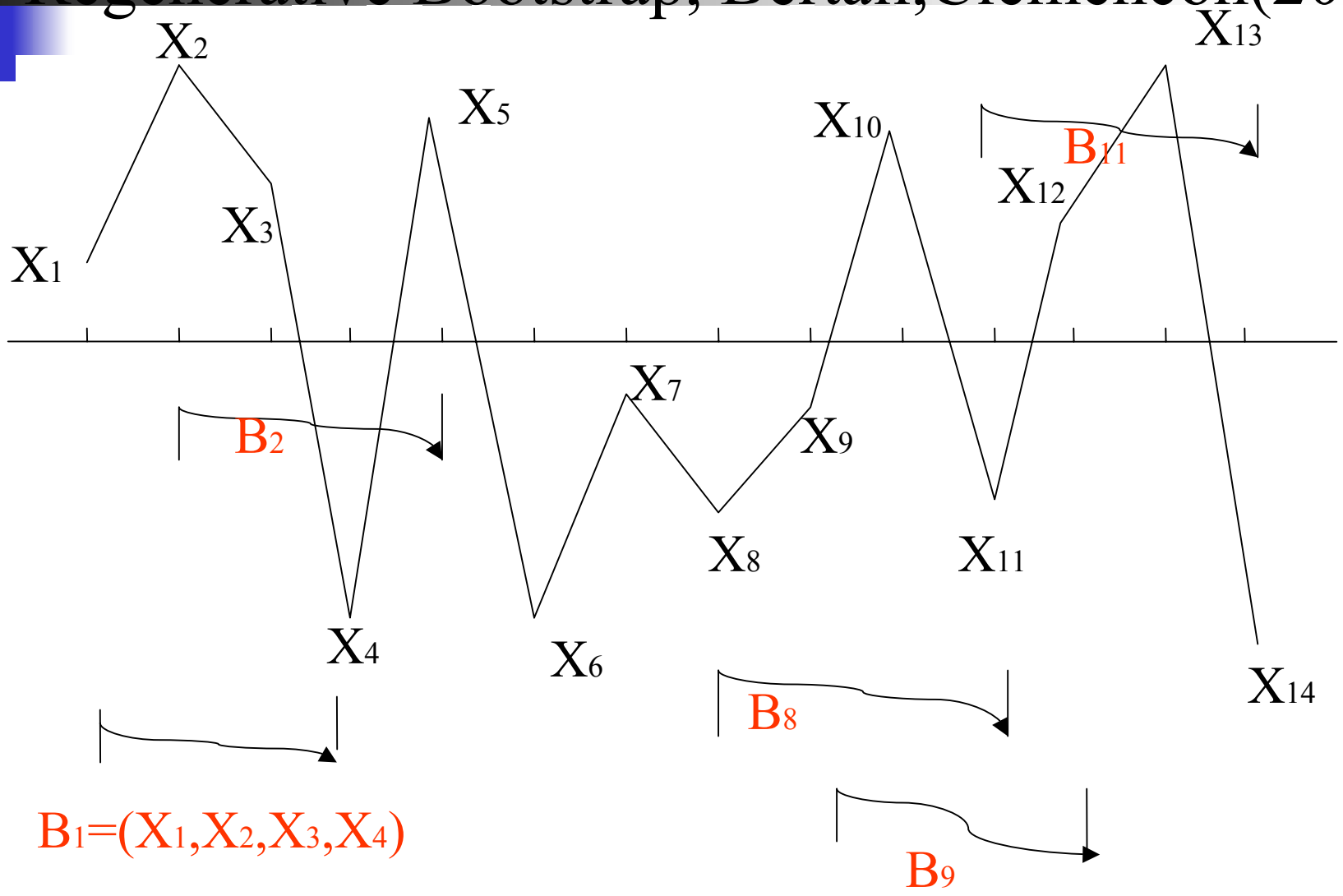
Essayer de se ramener au cas indépendant -

- soit en rééchantillonnant des blocs d'observations (gardant la dépendance)

- soit en utilisant la structure du modèle (résidus i.i.d) : Ex cas de la régression linéaire, cas des modèles AR, ARMA en séries temporelles

Moving-Block Bootstrap, Kunsch(1989)

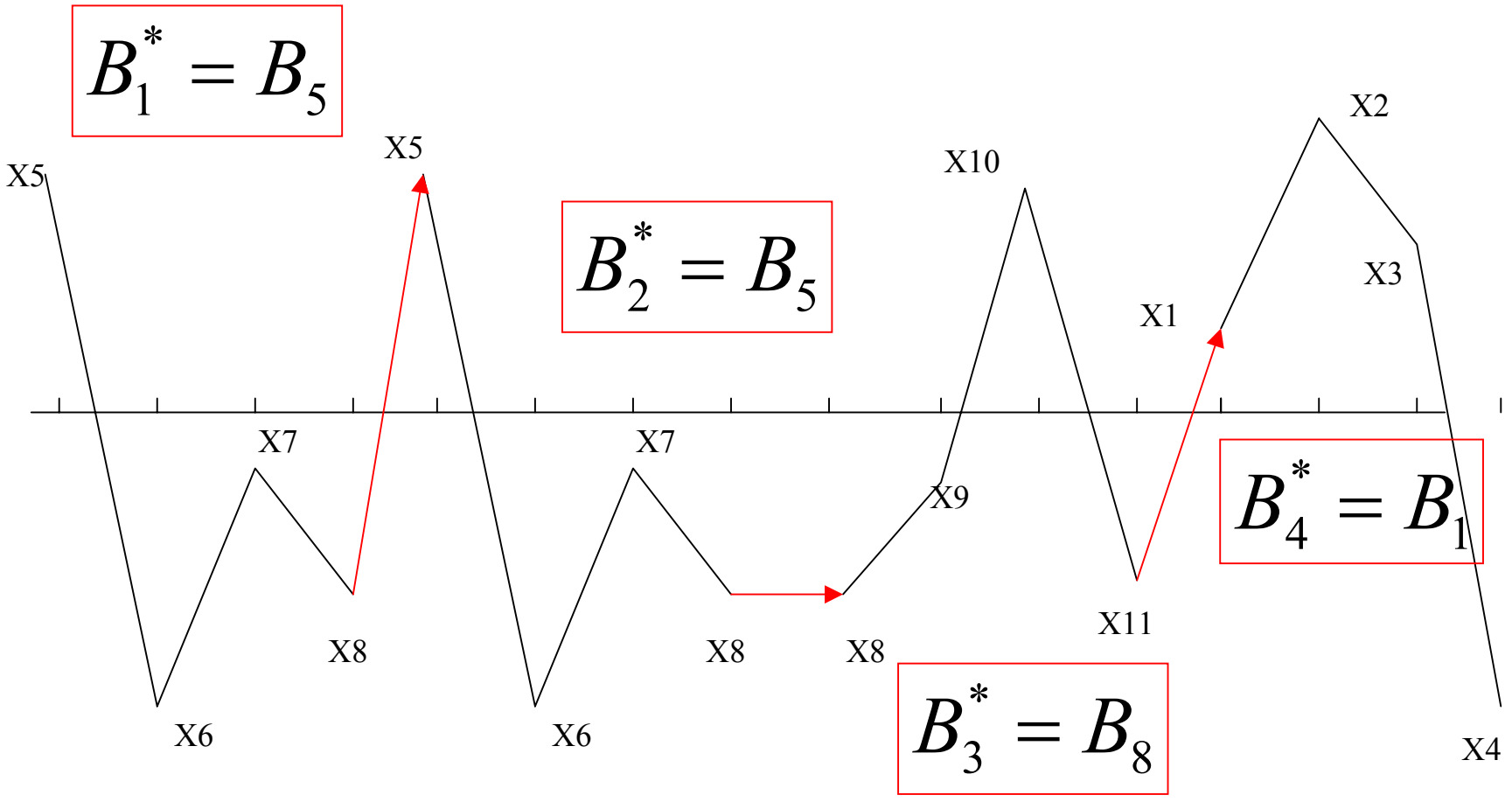
Regenerative Bootstrap, Bertail,Clémencon(2003



Longueur des blocs : o(sqrt(n))

$$(B_1^*, B_2^*, B_3^*, \ldots, B_{[n/b_n]}^*) \quad i.i.d \quad F_B = \frac{1}{(n-b_n+1)} \sum_{i=1}^{n-b_n+1} \delta$$

Rééchantillonnage de bloc : Reconstruction de séries artificielles



Bootstrap semi-paramétrique

Modèle autorégressif

$$X_0 = x_0, \quad V(\varepsilon_t) = \sigma^2, \quad c = x_0/\sigma$$
$$X_t = X_{t-1}\rho + \varepsilon_t \; t = 1, \ldots T$$

Estimation du modèle

$\widehat{\rho_T}$: estimateur des m.c.o

$$\widehat{\varepsilon}_t = X_t - \widehat{\rho_T}\, X_{t-1} \quad \tilde{\varepsilon}_t = \widehat{\varepsilon}_t - \frac{1}{T}\sum_{t=1}^{T} \widehat{\varepsilon}_t$$

Rééchantillonnage des résidus

Répéter B fois

$(\varepsilon_1^*, \varepsilon_2^*, ..., \varepsilon_T^*)$ tirés avec remise dans $(\tilde{\varepsilon}_1, \tilde{\varepsilon}_2, ..., \tilde{\varepsilon}_T)$

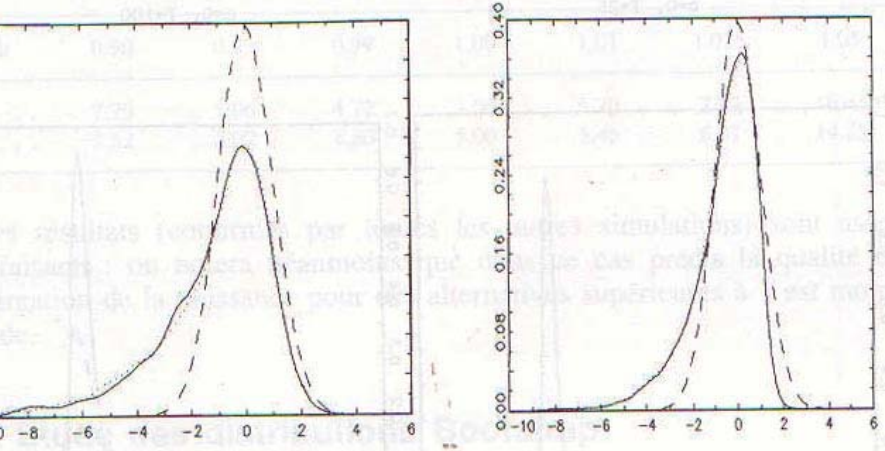
Recontruction des X_t^* :

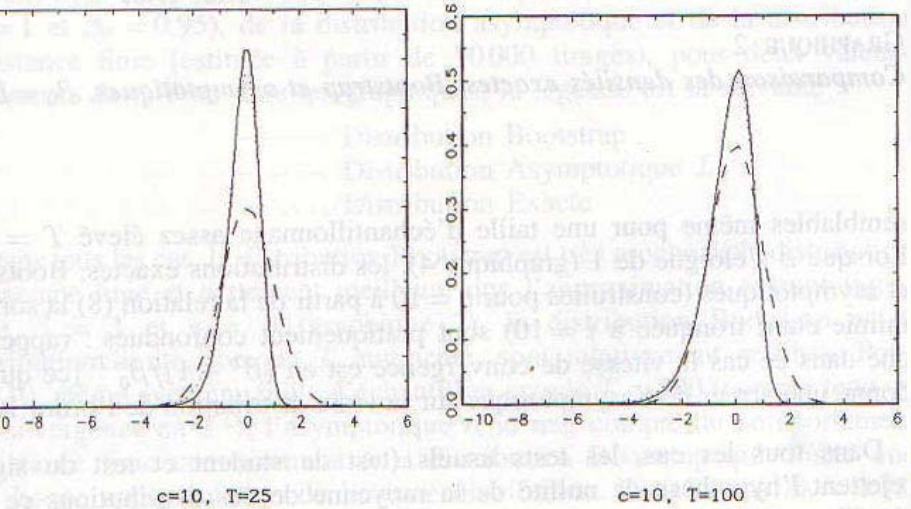
$X_0^* = x_0,$

$X_t^* = X_{t-1}^* \widehat{\rho_T} + \varepsilon_t^* \; t = 1, ...T$

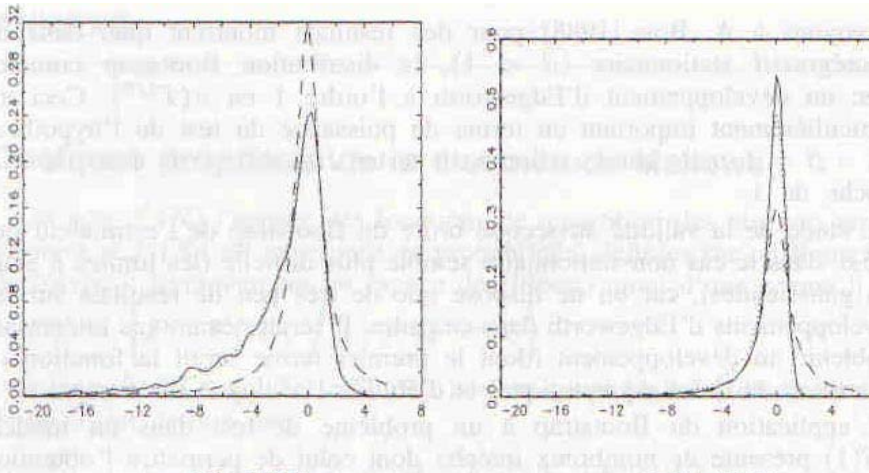
Recalculer $\widehat{\rho_T}^*$: estimateur des m.c.o

Fin

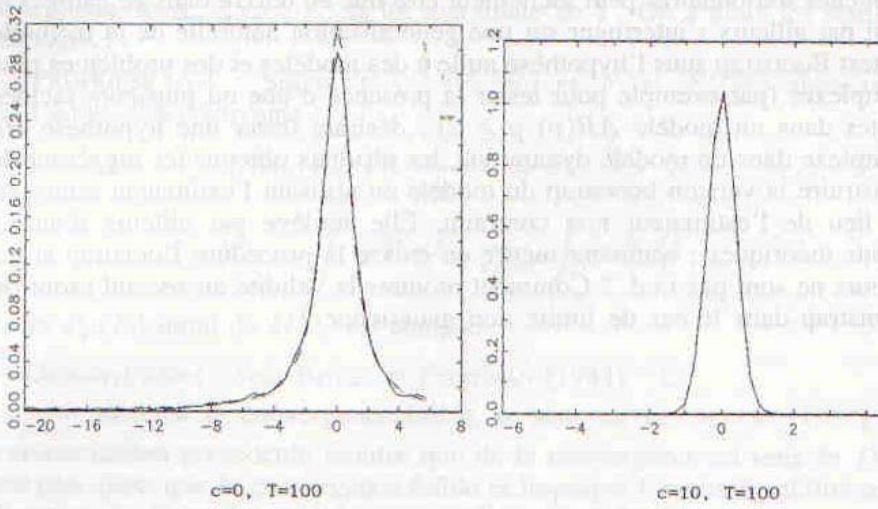




APHIQUE
mparaison des densités exactes, Bootstrap et asymptotiques, $\beta_0 = 0.95$.



GRAPHIQUE 4
Comparaison des densités exactes, Bootstrap et asymptotiques, $\beta_0 = 1.$



GRAPHIQUE
Comparaison des densités exactes, Bootstrap et asymptotiques, $\beta_0 = 1.$

Source : Bertail (1995), Annales d'Economie et de Statistique

Bibliographie succincte

- **Livres introductifs**

. Efron (1981), SIAM

.Efron, Tibshirani (1997), Chapman, Hall

. Davison, Hinkley (2000),

- **Livres théoriques**

.Hall (1992), Springer

.Shao, Tu (1995),Springer (biblio)

. Barbe, Bertail (1995), Springer

. Giné (1996), in Ecole Proba St Flour, Springer

Sous-échantillonnage

Politis, Romano, Wolf (2000), Springer