O estymacji przedziałowej funkcji dyfuzji metodami repróbkowania

Jan Mielniczuk, Paweł Teisseyre

Instytut Podstaw Informatyki, Polska Akademia Nauk

4 grudnia 2009

Wstęp

Rozważamy ciąg stacjonarny $X_0, X_\Delta, \ldots, X_{n\Delta}$ generowany z procesu autoregresyjnego postaci

$$X_{(i+1)\Delta} = X_{i\Delta} + \mu(X_{i\Delta})\Delta + \sigma(X_{i\Delta})\sqrt{\Delta}\varepsilon_{i+1}, \quad i = 0, \ldots, n-1, \tag{1}$$

dla pewnego ustalonego $\Delta > 0$, gdzie

- $\varepsilon_i, i = 2, \ldots, n$ są niezależnymi zmiennymi losowymi o rozkładzie $N(0, 1)$,
- $X_0 = 0$,
- funkcje $\mu(\cdot)$ i $\sigma(\cdot)$ są nieznanymi funkcjami, odpowiednio: dryfu i dyfuzji.

Wstęp

Dla małych Δ proces (1) spełniający równanie

$$X_{(i+1)\Delta} - X_{i\Delta} = \mu(X_{i\Delta})\Delta + \sigma(X_{i\Delta})\sqrt{\Delta}\varepsilon_{i+1}, \qquad i = 0, \ldots, n-1,$$

jest aproksymacją **procesu Itô**

$$dY_t = \mu(Y_t)\, dt + \sigma(Y_t)\, dW_t, \quad t \geqslant 0,$$

gdzie

- W_t to proces Wienera określony na przedziale $[0, \infty)$,
- $W_0 = 0$.

Warunki na zbieżność są podane n.p. w książce *Numerical Solution of Stochastic Differential Equations. Kloaden, P., Platen, E.*.

Wstęp

W dalszej częsci piszemy X_i zamiast $X_{i\Delta}$.
Zakładamy że proces (1) X_i spełniający równanie

$$X_{i+1} - X_i = \mu(X_i)\Delta + \sigma(X_i)\sqrt{\Delta}\varepsilon_{i+1}, \qquad i = 0, \ldots, n-1,$$

ma reprezentację:

$$X_i = J(\ldots, \varepsilon_{i-1}, \varepsilon_i),$$

dla pewnej funkcji mierzalnej J, co implikuje **ergodyczność procesu** (a więc również stacjonarność).

Estymacja funkcji dyfuzji

Definicja (Estymator Stantona, 1997)

$$\hat{\sigma}^2(x) = \frac{\sum_{i=1}^{n} Z_i^2 K_h(X_i - x)}{\sum_{i=1}^{n} K_h(X_i - x)},$$

gdzie

- $Z_i := \Delta^{-1/2}(X_{i+1} - X_i)$,
- $h = h(n)$ jest współczynnikiem wygładzającym ($h \to 0$ przy $n \to \infty$),
- K jest gęstością prawdopodobieństwa,
- $K_h(u) := h^{-1}K(u/h)$,

Estymacja funkcji dyfuzji

Zauważmy, że

$$\mathbb{E}(Z_i^2|X_i = x) = \mathbb{E}(\Delta^{-1}(X_{i+1} - X_i)^2|X_i = x) =$$

$$\Delta^{-1}\mathbb{E}\{[\mu(X_i) + \sigma(X_i)\sqrt{\Delta}\varepsilon_{i+1}]^2|X_i = x\} = \Delta\mu^2(x) + \sigma^2(x).$$

Estymator Stantona przybliża więc wielkość

$$\Delta\mu^2(x) + \sigma^2(x)$$

a nie wielkość $\sigma^2(x)$.

Estymacja funkcji dyfuzji

Przyjmijmy następujące założenia:

1. μ and σ są dwukrotnie różniczkowalne w otoczeniu x i spełniają warunek Lipshitza na $\mathbb{R}$,
2. f_ε jest ograniczona i spełnia warunek Lipshitza na $\mathbb{R}$,
3. $\inf_{y\in\mathbb{R}} \sigma(y) > 0$ i gęstość stacjonarna $f(x) > 0$,
4. K jest symetryczną i ograniczoną gęstością prawdopodobieństwa i ma zwarty nośnik,
5. $nh_n^5 \to C \geqslant 0$,
6. $\|X_i - X_i'\| = \mathcal{O}(r^i)$ dla pewnego $0 < r < 1$, gdzie $X_i' = J(\ldots, \varepsilon_{-1}, \varepsilon_0', \varepsilon_i)$ i ε_0' jest niezależną kopią ε_0.

Estymacja funkcji dyfuzji

Twierdzenie (rozkład asymptotyczny estymatora Stantona)

Załóżmy że spełnione są warunki (1)-(5). Wówczas

$$\sqrt{nh}[\hat{\sigma}^2(x) - \sigma^2(x) - \Delta\mu^2(x)] \xrightarrow{d} N\left(\sqrt{C}C_w, \frac{v(x)}{f(x)}\right),$$

gdzie

- $v(x) = \mathbb{E}[2\mu(x)\sigma(x)\sqrt{\Delta}\varepsilon_{i+1} + (\varepsilon_{i+1}^2 - 1)\sigma^2(x)]^2 \int K^2(v)dv$,
- $C_w = \int v^2 K(v)dv \cdot [f'(x)w'(x) + \frac{1}{2}f(x)w''(x)]/f(x)$,
- $w(x) = \Delta\mu^2(x) + \sigma^2(x)$.

Metody repróbkowania

1. Konstrukcja punktowych asymptotycznych przedziałów ufności dla funkcji $\sigma^2(\cdot)$ w oparciu o rozkład asymptotyczny jest bardzo trudna z uwagi na występowanie nieznanych funkcji $\mu(\cdot)$ oraz $\sigma^2(\cdot)$ w asymptotycznym obciążeniu i wariancji.
2. Można skonstruować przedziały ufności przy użyciu **metod repróbkowania**.
3. Celem repróbkowania jest generacja pseudoprób o właściwościach podobnych do oryginalnej próby $X_1, \ldots, X_n$.

Metody repróbkowania- Bootstrap autoregresyjny

Autoregression bootstrap

$$X_1^* = X_1,$$

$$X_{i+1}^* = \Delta\bar{\mu}(X_i^*) + X_i^* + \bar{\sigma}(X_i^*)\varepsilon_{i+1}^*\sqrt{\Delta},\ i = 1, \ldots, n-1,$$

gdzie:

- $\bar{\mu}(\cdot)$ i $\bar{\sigma}(\cdot)$ są pewnymi wstępnymi estymatorami $\mu(\cdot)$ i $\sigma(\cdot)$.
- $\{\varepsilon_i^*, i = 2, \ldots, n\}$ są próbkowane z rozkładu $N(0,1)$ lub ze studentyzowanych rezyduów postaci

$$\left\{ \frac{(X_{i+1} - \bar{\mu}(X_i) - X_i)}{\bar{\sigma}(X_i)\sqrt{\Delta}}, i = 1, \ldots, n \right\}.$$

Metody repróbkowania- Bootstrap par (pair bootstrap)

Przypomnijmy że funkcja $\sigma^2(\cdot)$ może być traktowana jako aproksymacja funkcji regresji dla danych $\{(X_i, Z_i), i = 1, \ldots, n\}$, gdzie $Z_i = \Delta^{-1/2}(X_{i+1} - X_i)$.

Pair bootstrap

$$\{(X_{N_i}, Z_{N_i}), i = 1, \ldots, n-1\},$$

gdzie

- $N_1, \ldots, N_n$ jest ciągiem niezależnych zmiennych losowych o rozkładzie jednostajnym na $\{1, \ldots, n-1\}$.

Metody repróbkowania- Subsampling

Subsampling

Bloki danych $\mathcal{B}_{i,b} := (X_{i\Delta}, \ldots, X_{(i+b-1)\Delta})$,

dla $i = 1, \ldots, n - b + 1$ są próbami wielkości b z oryginalnego procesu.

Metody repróbkowania- Subsampling

Niech $\hat{\sigma}^2_{i,b}(x)$ oznacza est. Stantona obliczony na podstawie $\mathcal{B}_{i,b}$.

Theorem

Załóżmy że spełnione są warunki (1)-(5) oraz że $nh_n/bh_b \to \infty$ *przy* $n/b \to \infty$. *Wówczas*

$$P\{\sqrt{nh_n}[\hat{\sigma}^2(x) - \sigma^2(x) - \Delta\mu^2(x)] \leqslant c_{n,b}(1-\alpha)\} \to 1-\alpha,$$

gdzie

- $c_{n,b}(1-\alpha)$ oznacza kwantyl empiryczny rzędu $(1-\alpha)$ rozkładu $\sqrt{bh_b}[\hat{\sigma}^2_{i,b}(x) - \hat{\sigma}^2(x)]$

Metody repróbkowania- Subsampling

Stąd wyznaczamy przedział ufności (**uwaga**: dla wielkości $\sigma^2(x) + \Delta\mu^2(x)$).

Bootstrapowy przedział ufności

$$\left[\hat{\sigma}^2(x)(1+\eta_n) - \eta_n \hat{\sigma}^2_{b,i,1-\alpha/2}(x), \hat{\sigma}^2(x)(1+\eta_n) - \eta_n \hat{\sigma}^2_{b,i,\alpha/2}(x)\right],$$

gdzie

- $\eta_n = (bh_b/nh_n)^{1/2}$
- $\hat{\sigma}^2_{b,i,\alpha}(x)$ jest kwantylem empirycznym rzędu α est. $\hat{\sigma}^2_{i,b}(x)$.

Wyniki symulacji

W symulacjach rozważamy procesy ciągłe opisane stochastycznymi równaniami różniczkowymi.

Model Vasicka

$$dX_t = \kappa(\alpha - X_t)dt + \sigma dW_t, \quad t \geqslant 0, \quad \kappa > 0,$$

Model CIR

$$dX_t = \kappa(\alpha - X_t)dt + \sigma\sqrt{X_t}dW_t, \quad t \geqslant 0 \quad , 2\kappa\alpha \geqslant \sigma^2.$$

Dla powyższych modeli **znana jest gęstość przejścia**, co umożliwia generację dokładnej trajektorii procesów.

Wyniki symulacji- wyznaczenie optymalnej długości bloku

Zakładamy, że długość bloku $b \to \infty$ oraz $\frac{b}{n} \to 0$ przy $n \to \infty$.

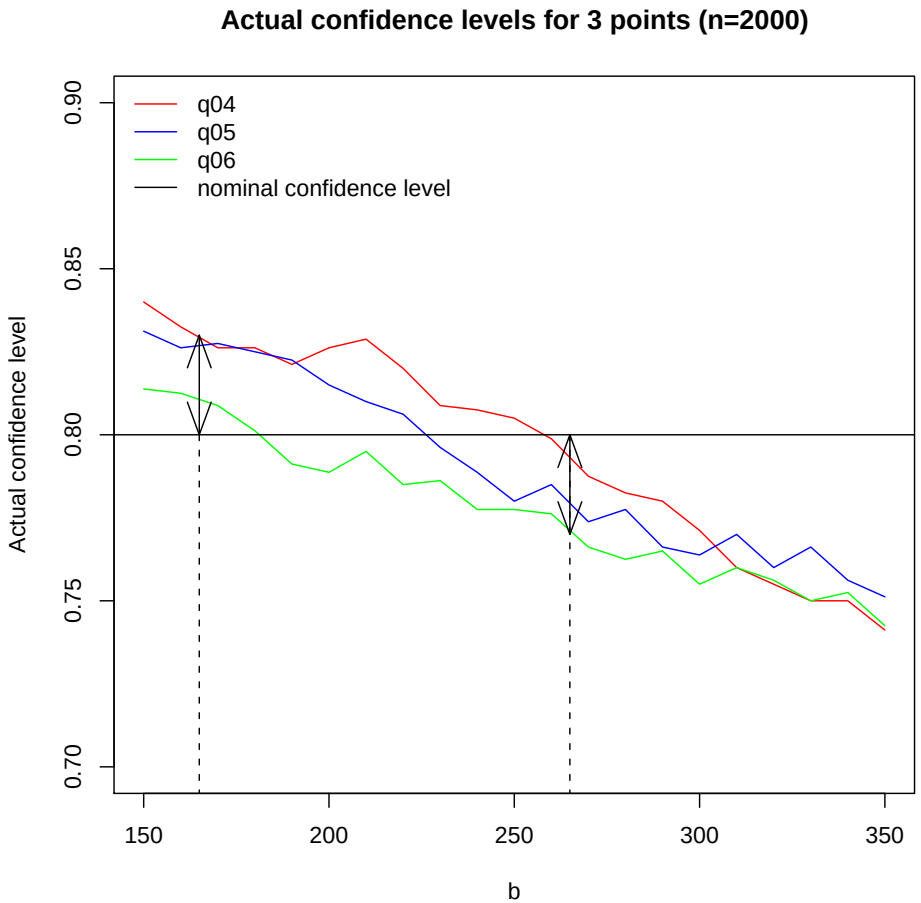
Problem

Jak wyznaczyć $b_{opt} = b_{opt}(n)$?

Chcemy znaleźć **zależność funkcyjną między b_{opt} i n** dla modeli Vasicka i CIR.

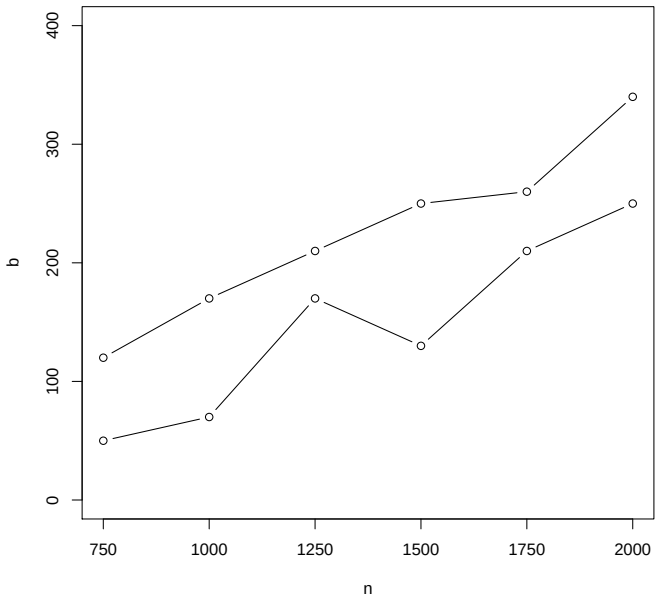
Wyniki symulacji- wyznaczenie optymalnej długości bloku

Dla danego n wyznaczamy $(b_{opt,min}, b_{opt,max})$, tak aby $|(1-\alpha_{nom}) - (1-\alpha_{act})| \leqslant 0.03$.

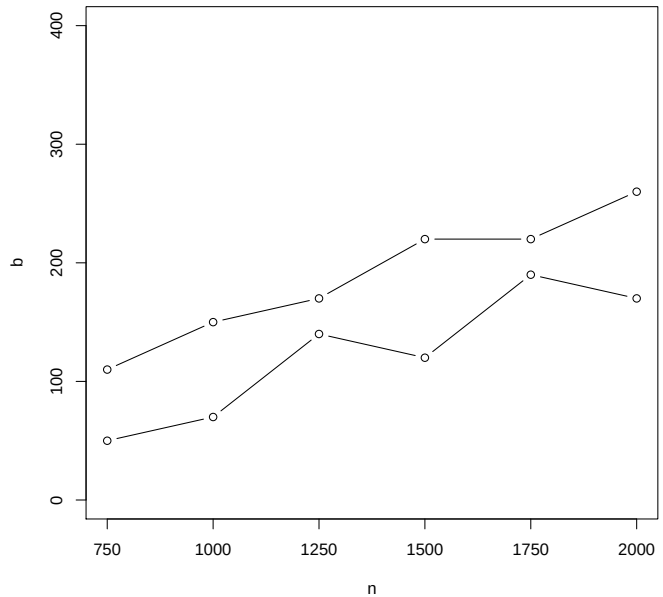


Wyniki symulacji- wyznaczenie optymalnej długości bloku

Dla $n = 500, 750, \ldots, 2000$ wyznaczamy $(b_{opt,min}, b_{opt,max})$, tworząc obszary dla modelu Vasicka oraz modelu CIR.



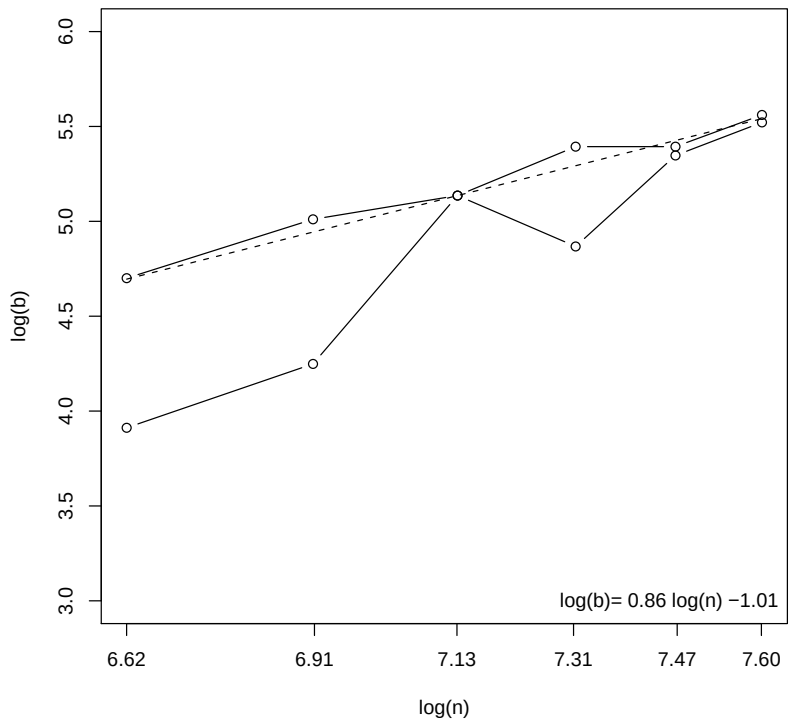
Rysunek: Model Vasicka



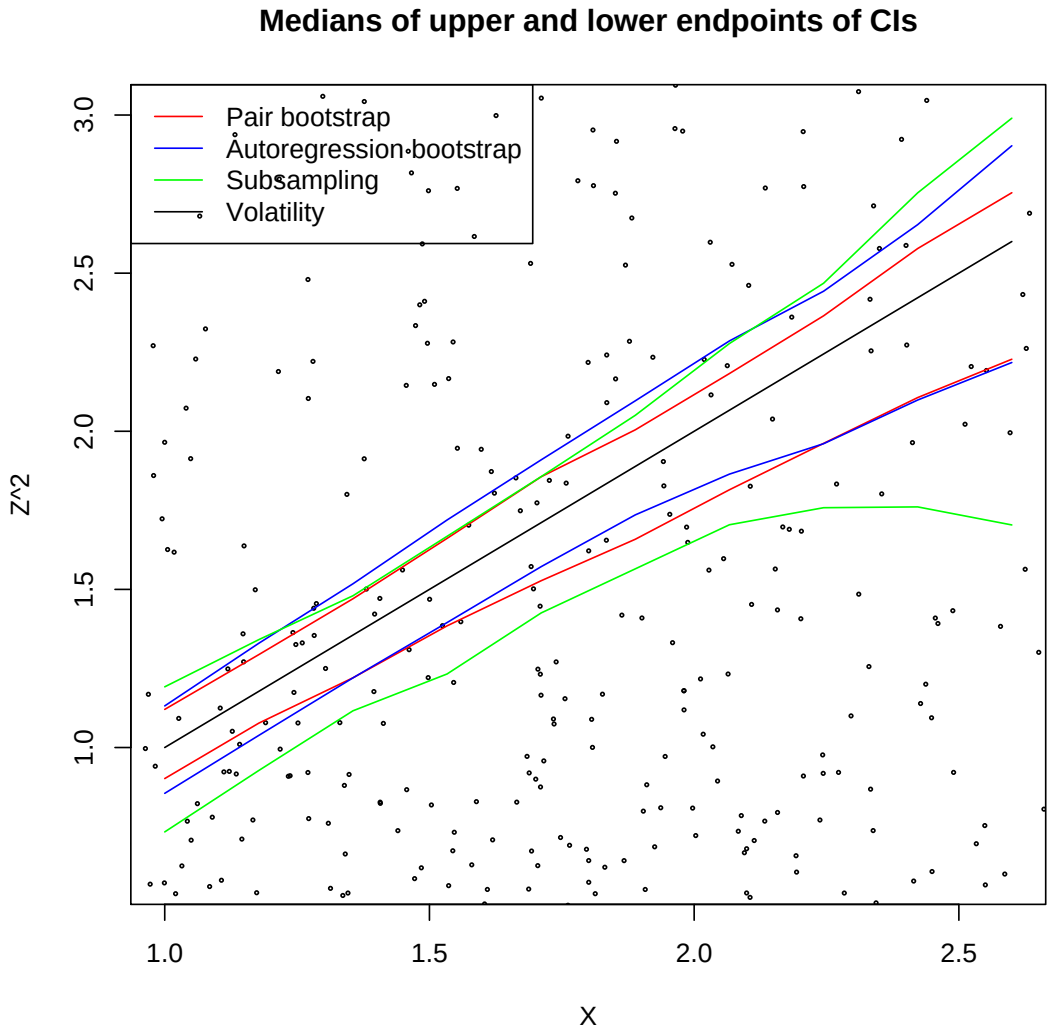
Rysunek: Model CIR

Wyniki symulacji- wyznaczenie optymalnej długości bloku

Powstaje wspólny obszar dla dwóch modeli, wraz z dopasowaną krzywą $\log(b_{opt}) = 0.86 \log(n) - 1.01$.



Wyniki symulacji- przedziały ufności dla modelu CIR



Wyniki symulacji- przedziały ufności dla modelu Vasicka

Tabela: Prawdopodobieństwa pokrycia oraz średnie długości przedziałów dla 3 metod repróbkowania dla modelu Vasicka($1 - \alpha = 0.8$).

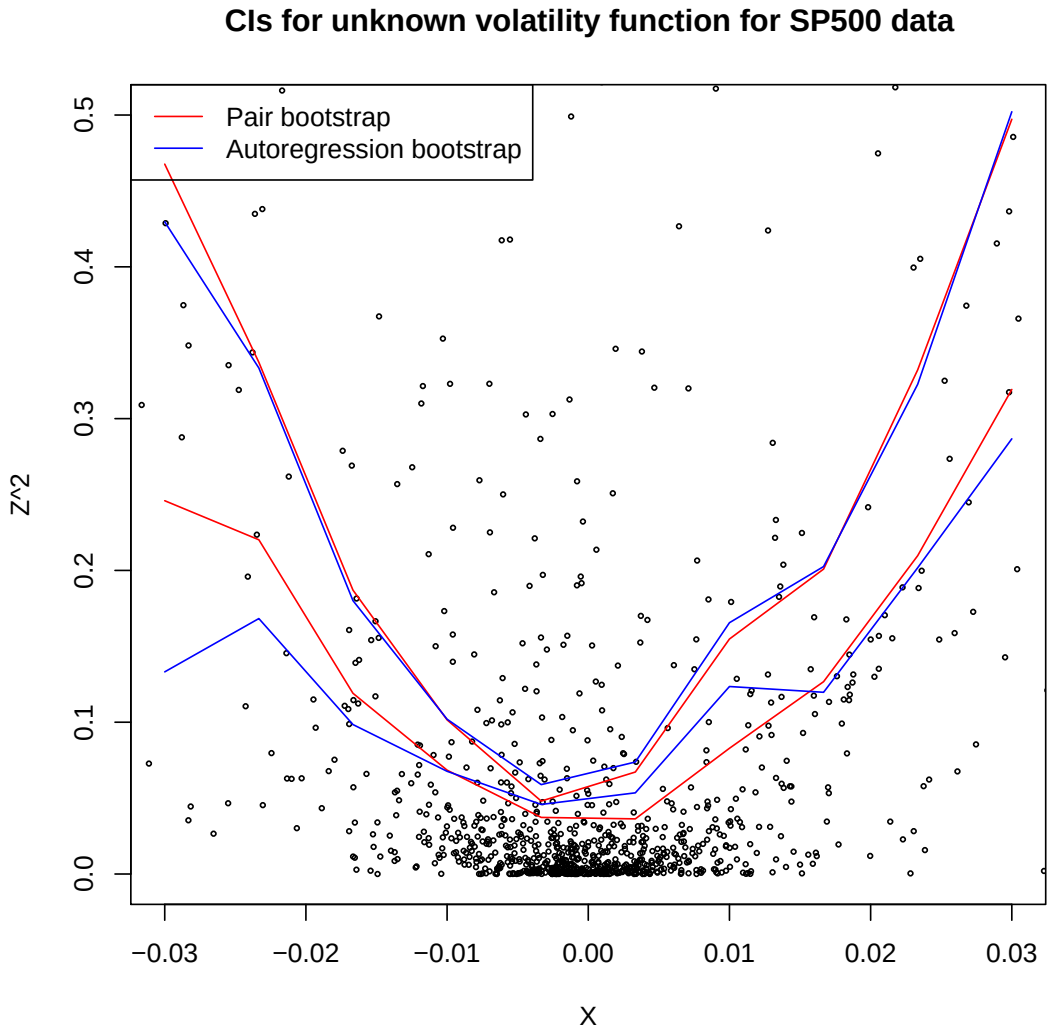
	Autoregression bootstrap	Pair bootstrap	Subsampling
$n = 1000$			
$q_{0.4}$	0.78 (0.2)	0.78 (0.18)	0.81 (0.26)
$q_{0.5}$	0.76 (0.19)	0.81 (0.17)	0.8 (0.25)
$q_{0.6}$	0.77 (0.2)	0.79 (0.17)	0.8 (0.26)
$n = 1500$			
$q_{0.4}$	0.77 (0.15)	0.73 (0.13)	0.8 (0.19)
$q_{0.5}$	0.77 (0.15)	0.73 (0.13)	0.8 (0.19)
$q_{0.6}$	0.78 (0.15)	0.74 (0.13)	0.79 (0.19)
$n = 2000$			
$q_{0.4}$	0.79 (0.13)	0.71 (0.12)	0.81 (0.16)
$q_{0.5}$	0.77 (0.12)	0.74 (0.11)	0.82 (0.15)
$q_{0.6}$	0.78 (0.13)	0.74 (0.12)	0.81 (0.16)

Wyniki symulacji- przedziały ufności dla modelu CIR

Tabela: Prawdopodobieństwa pokrycia oraz średnie długości przedziałów dla 3 metod repróbkowania dla modelu CIR($1 - \alpha = 0.8$).

	Autoregression bootstrap	Pair bootstrap	Subsampling
$n = 1000$			
$q_{0.4}$	0.77 (0.32)	0.76 (0.28)	0.81 (0.40)
$q_{0.5}$	0.76 (0.36)	0.76 (0.33)	0.81 (0.47)
$q_{0.6}$	0.76 (0.34)	0.76 (0.4)	0.78 (0.56)
$n = 1500$			
$q_{0.4}$	0.78 (0.24)	0.77 (0.22)	0.8 (0.30)
$q_{0.5}$	0.77 (0.28)	0.76 (0.26)	0.79 (0.34)
$q_{0.6}$	0.76 (0.33)	0.75 (0.31)	0.77 (0.41)
$n = 2000$			
$q_{0.4}$	0.79 (0.21)	0.77 (0.19)	0.81 (0.25)
$q_{0.5}$	0.76 (0.24)	0.77 (0.22)	0.78 (0.29)
$q_{0.6}$	0.77 (0.29)	0.75 (0.27)	0.78 (0.35)

Wyniki symulacji- przedziały ufności dla danych SP500



Wybrana literatura

1. J. Fan, *A Selective Overview of Nonparametric Methods in Financial Econometrics*, Statistical Science, Vol. 20, No. 4, pages 317–337, 2005.
2. D. N. Politis, J. P. Romano, M. Wolf, *Subsampling*, Springer New York, 1999.
3. J. Franke, J. P. Kreiss, E. Mammen, *Bootstrap of kernel smoothing in nonlinear time series*, Bernoulli, Vol. 8, No. 1, pages 1–37, 2002.
4. D. A. Freedman, *Bootstraping regression models*, The Annals of Statistics, Vol. 9, pages 1218–1298, 1981.