Υπολογιστική Στατιστική

Κατερίνα Ορφανογιαννάκη

Τμήμα Μαθηματικών
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
korfanog@math.uoa.gr

2020-2021

Bootstrap στη Γραμμική Παλινδρόμηση

Υπάρχουν 2 διαφορετικές προσεγγίσεις:

- Επαναδειγματοληψία με επανάθεση από τις παρατηρήσεις. Τώρα κάθε παρατήρηση είναι ένα ολόκληρο διάνυσμα που συνδέεται με την αρχική παρατήρηση.
- Εφαρμοφή *Bootstrap* στα κατάλοιπα του μοντέλου που έχουμε προσαρμόσει στα αρχικά δεδομένα.

Η δεύτερη προσέγγιση είναι προτιμότερη καθώς η πρώτη προσέγγιση καταστρατηγεί την υπόθεση για σταθερό πίνακα σχεδιασμού. Η *Bootstrap* στην γραμμική παλινδρόμηση αφαιρεί υποθέσεις αναφορικά με την κατανομή των καταλοίπων και άρα επιτρέπει συμπερασματολογία ακόμα και αν τα σφάλματα δεν ακολουθούν την κανονική κατανομή.

Bootstrap στα κατάλοιπα

Έστω το μοντέλο $Y = \beta X + \epsilon$ χρησιμοποιόντας τον σύνηθες συμβολισμό.

Ο αλγόριθμος *Bootstrap* είναι ο ακόλουθος:

- Προσαρμόστε το μοντέλο στα αρχικά δεδομένα. Εκτιμείστε τις παραμέτρους του μοντέλου $\hat{\beta}$ και τα κατάλοιπα του προσαρμοσμένου μοντέλου $\hat{\epsilon}_i$, $i = 1, \ldots, n$.
- Πάρτε *Bootstrap* δείγμα $\epsilon^* = (\epsilon_1^*, \ldots, \epsilon_n^*)$ από τα κατάλοιπα με δειγματοληψία με επανάθεση.
- Χρησιμοποιόντας τον πίνακα σχεδιασμού δημιουργείστε *Bootstrap* τιμές για τη μεταβλητή απόκρισης χρησιμοποιόντας τη σχέση:

$$Y^* = \hat{\beta} X + \epsilon^*$$

- Εφαρμόστε το μοντέλο χρησιμοποιόντας σα μεταβλητή απόκρισης την Y^* και πίνακα σχεδιασμού τον X.
- Κρατείστε όλες τις ποσότητες που σας ενδιαφέρουν από το προσαρμοσμένο μοντέλο (π.χ. *MSE*, *F-statistic*, συντελεστές κλπ)
- Επαναλάβετε τη διαδικασία B φορές.

Παράδειγμα

Ένας ορνιθολόγος κατέγραψε για 12 σπουργίτια την ηλικία τους σε μέρες και το μήκος των φτερών τους σε εκατοστά με σκοπό να ελέγξει αν υπάρχει μια γραμμική σχέση του μήκος των φτερών με την ηλικία. Τα δεδομένα φαίνονται στον παρακάτω πίνακα:

Μήκος φτερών (σε εμ)	Ηλικία (σε μέρες)
1.40	3
1.50	3
2.20	5
2.40	6
3.10	8
3.20	9
3.20	10
3.90	11
4.10	12
4.70	14
4.50	15
5.20	17

Παράδειγμα

1. Να εκτιμήσετε τα τυπικά σφάλματα των συντελεστών α και β της παλινδρόμησης.
2. Είναι η σταθερά διαφορετική της μονάδας;
3. Υπάρχει σχέση ανάμεσα στις 2 μεταβλητές;

Συγκεντρωτικά αποτελέσματα

Τυπικά σφάλματα και 95% δ.ε.:

			95% δ.ε.	
	Μέση τιμή	Τυπική απόκλιση	κλασσική μέθοδος	
$\hat{s}^2$	0.026	0.00776	0.011	0.041
$\hat{a}$	0.777	0.10961	0.563	0.991
$\hat{\beta}$	0.266	0.01052	0.245	0.286
$\hat{F}$	717.536	288.291	406.493	1566.891
$\hat{R}^2$	0.984	0.0044	0.975	0.993

Παραμετρική *Bootstrap* στην παλινδρόμηση

Αντί για τη χρήση μη παραμετρικής *Bootstrap* μπορούμε να χρησιμοποιήσουμε παραμετρική *Bootstrap* με παρόμοιο τρόπο. Αυτό υπονοεί ότι υποθέτουμε ότι τα σφάλματα ακολουθούν κάποια κατανομή (π.χ. t ή κάποια μείξη κανονικών). Τότε πλήρη συμπερασματολογία είναι διαθέσιμη με χρήση *Bootstrap* ενώ κάτι αντίστοιχο είναι πολύ δύσκολο όταν χρησιμοποιούμε κλασσιμές προσεγγίσεις.
Η μόνη διαφορά είναι ότι τα σφάλματα για κάθε δείγμα *Bootstrap* γεννιούνται από την κατανομή που έχουμε κάθε φορά υποθέσει.

Η *Bootstrap* σε Χρονολογικές σειρές

Θεωρείστε ένα μοντέλο $AR(2)$:

$$Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \epsilon_t$$

Στην κλασσική περίπτωση υποθέτουμε κανονικότητα των καταλοίπων. Αν αυτή η υπόθεση δεν ικανοποιείται μπορούμε να χρησιμοποιήσουμε *Bootstrap* προκειμένου να προχωρήσουμε σε συμπερασματολογία.

Ο αλγόριθμος

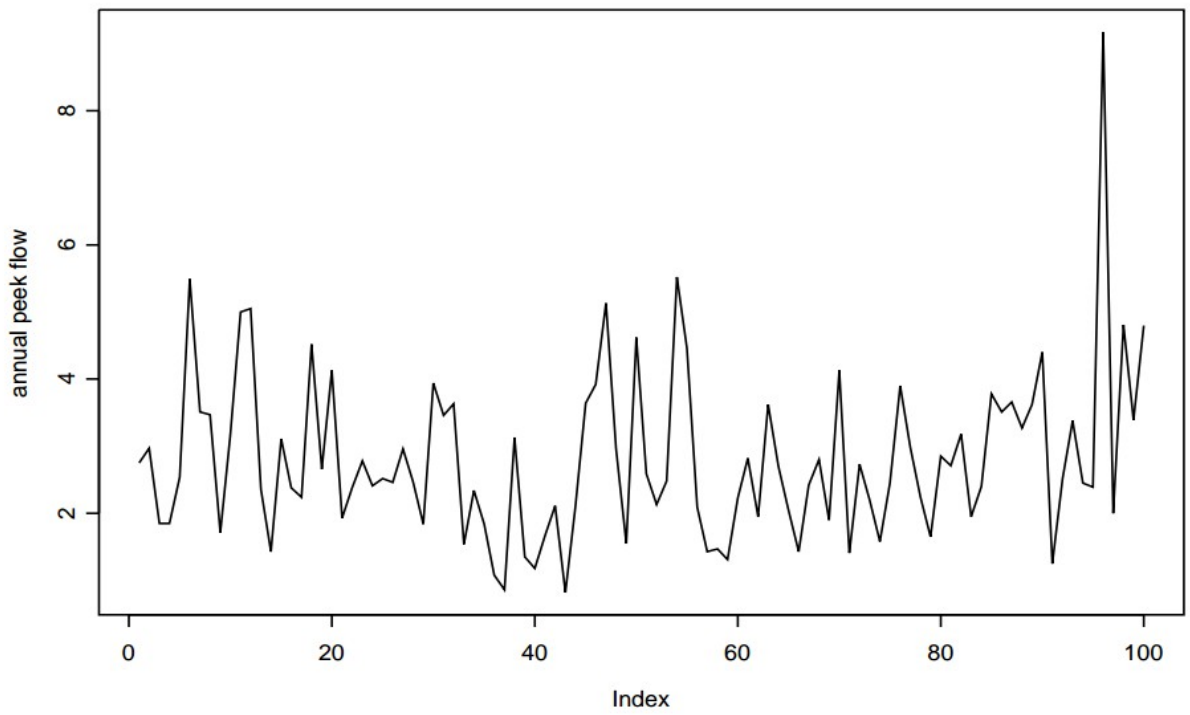
- *Βήμα 1:* Προσάρμοσε το μοντέλο $AR(p)$ στα δεδομένα. Βρες τις εκτιμήσεις $(\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p)$
- *Βήμα 2:* Βρες τα κατάλοιπα του εκτιμηθέντος μοντέλου. Έστω ότι αυτά είναι $(\hat{\varepsilon}_1, \hat{\varepsilon}_2, ..., \hat{\varepsilon}_t)$.
- *Βήμα 3:* Ξεκίνησε την *Bootstrap*: Πάρε δείγμα από τα εκτιμηθέντα κατάλοιπα με επανάθεση. Έστω ότι το *Bootstrap* δείγμα είναι το $(\hat{\varepsilon}^*_1, \hat{\varepsilon}^*_2, ..., \hat{\varepsilon}^*_t)$.
  Κατασκεύασε την σειρά ως εξής:
  1. Θέσε $y^*_i = y_i, i = 1, ..., p$
  2. Θέσε $y^*_t = \hat{\beta}_1 y^*_{t-1} + \hat{\beta}_2 y^*_{t-2} + ... + \hat{\beta}_p y^*_{t-p} + \varepsilon^*_t$, για $t = p + 1, ..., T$
  3. Προσάρμοσε το μοντέλο $AR(p)$ στη νέα σειρά $y^*_t, t = 1, ..., T$ και εκτίμησε τις παραμέτρους
  4. Επανέλαβε τη διαδικασία B φορές.
- *Βήμα 4:* Χρησιμοποίησε τις B τιμές των παραμέτρων από τα *Bootstrap* δείγματα για συμπερασματολογία (πχ ελέγχους υποθέσεων, διαστήματα εμπιστοσύνης κλπ)

Παραλλαγή παραμετρικής *Bootstrap*

Στην περίπτωση που μας ενδιαφέρει να κάνουμε παραμετρική *Bootstrap* αυτό που αλλάζει είναι πως αντί να γεννήσουμε τα κατάλοιπα από την εμπειρική τους κατανομή τα προσομοιώνουμε από την παραμετρική υπόθεση που έχουμε κάνει. Πχ αν έχουμε υποθέσει ότι τα κατάλοιπα προέρχονται από μια κατανομή t τότε απλά προσομοιώνουμε από αυτή την κατανομή.

Παράδειγμα: Ετήσια μέγιστη ροή

Τα δεδομένα αφορούν $n = 100$ από την ετήσια μέγιστη ροή σε ένα συγκεκριμένο σημείο του ποταμού Μισσουρι για τα έτη 1898-1997 σε cm^3/sec.

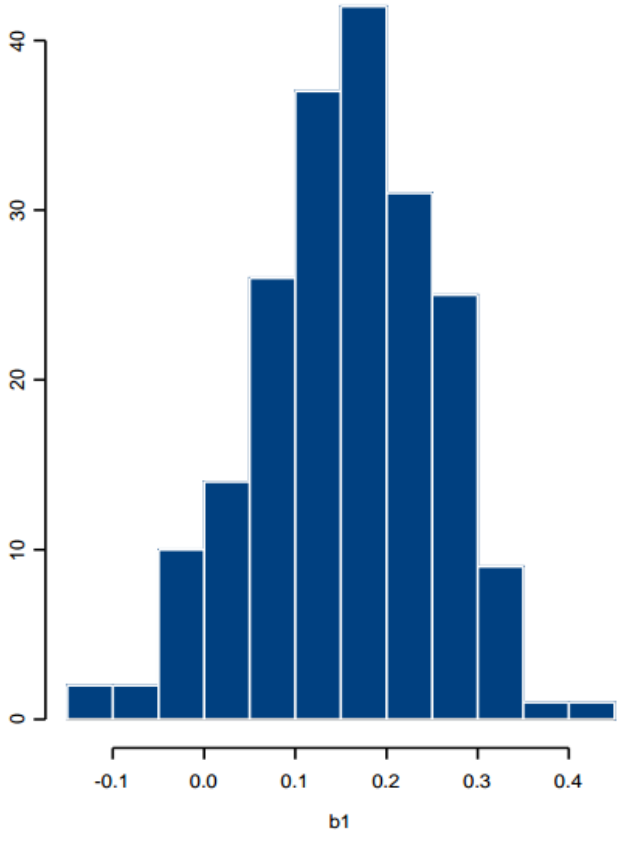


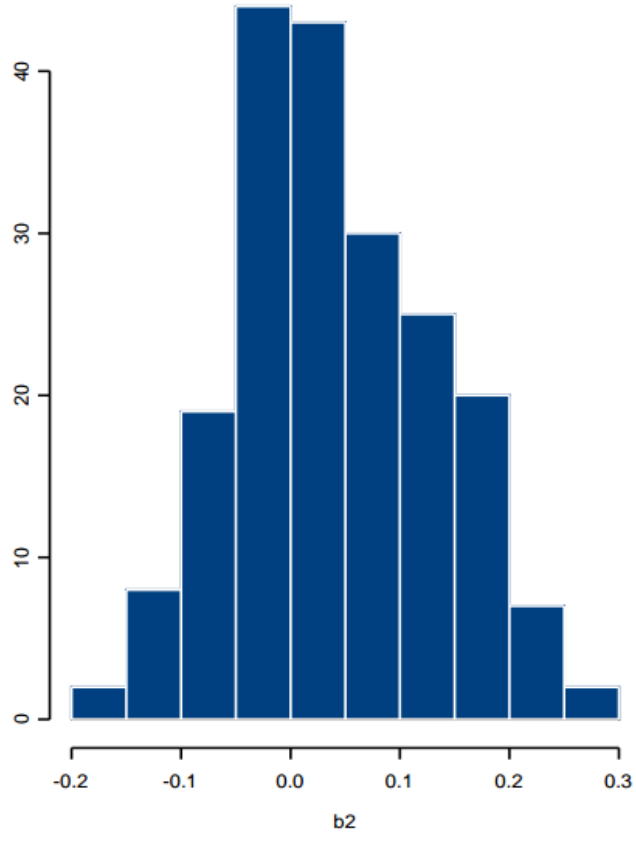
Παράδειγμα: Ετήσια μέγιστη ροή - Αποτελέσματα

θ	μέσος	τυπ. σφάλμα	95% δ.ε.	τιμή από το δείγμα
$\hat{\beta}_1$	0.155	0.0964	(-0.032, 0.333)	0.112
$\hat{\beta}_1$	0.042	0.0923	(-0.118, 0.226)	0.063
$\hat{\sigma}^2$	1.669	0.415	(1.087, 2.533)	1.579

Παράδειγμα: Ετήσια μέγιστη ροή (συνέχεια)

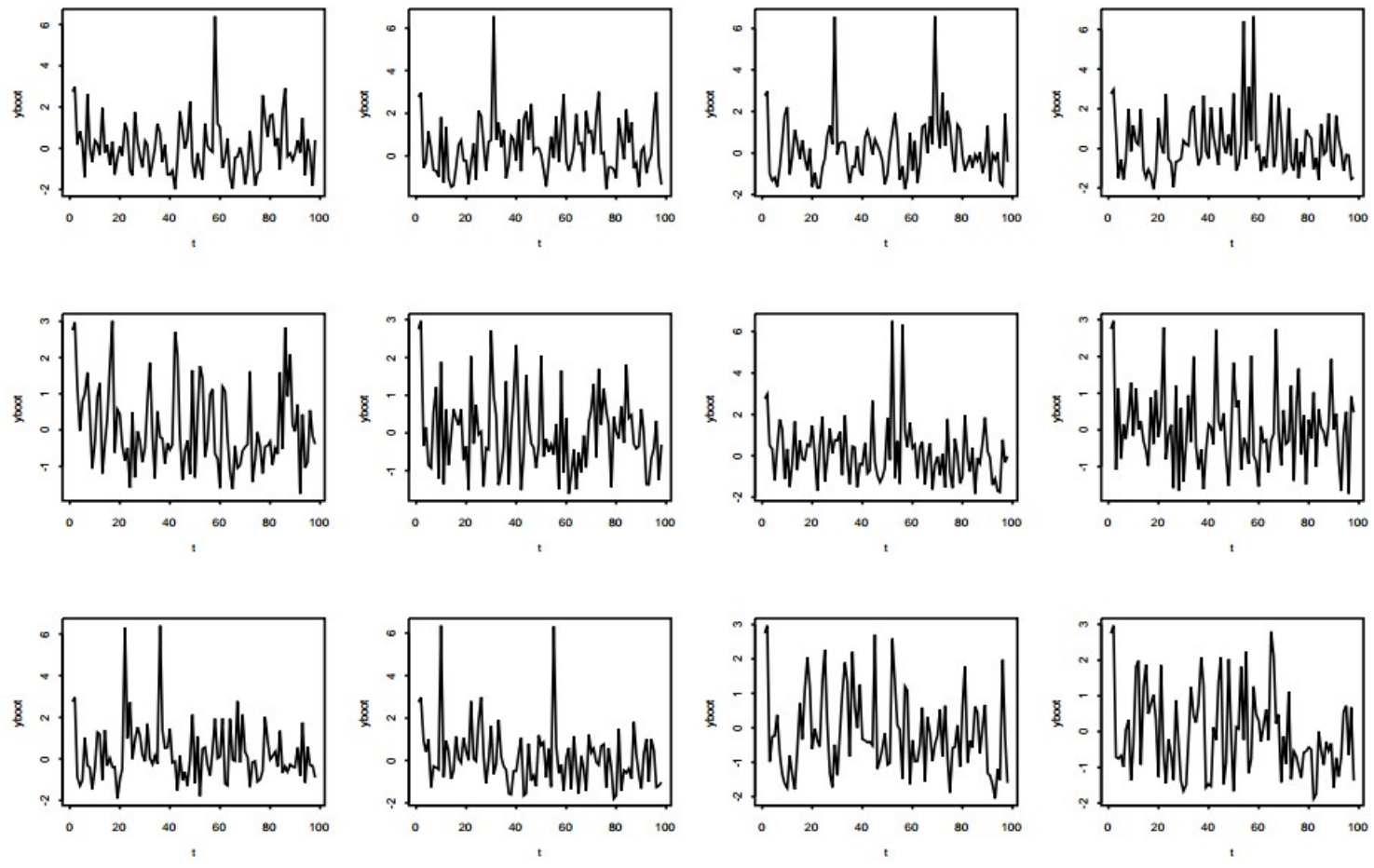
Ιστόγραμμα των *Bootstrap* τιμών για τις παραμέτρους:



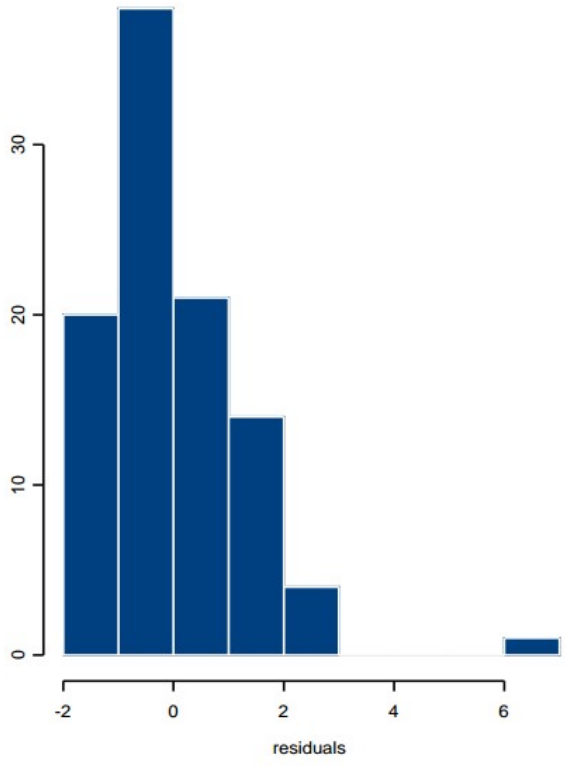


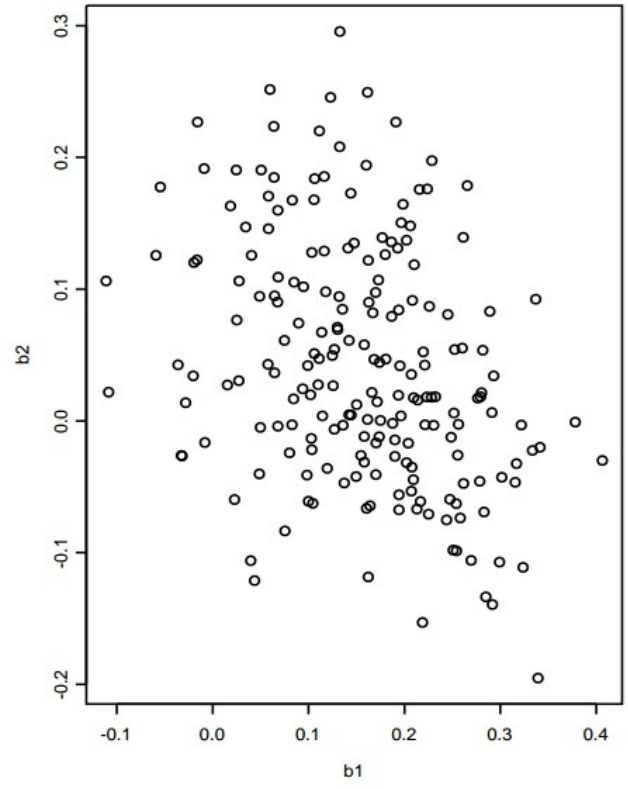
Παράδειγμα: Ετήσια μέγιστη ροή (συνέχεια)

Μερικές *Bootstrap* χρονολογικές σειρές που κατασκευάσαμε με τον αλγόριθμο:



Παράδειγμα: Ετήσια μέγιστη ροή - Κατάλοιπα





Bootstrap για εξαρτημένα δεδομένα

Η *Bootstrap* βασίζεται στη δημιουργία ανεξάρτητων δειγμάτων από την $\hat{F}_n$. Για εξαρτημένα δεδομένα η τυπική *Bootstrap* δεν μπορεί να εφαρμοστεί. Η ιδέα είναι ότι χρειάζεται να μιμηθούμε την εξάρτηση των δεδομένων. Αυτό μπορεί να γίνει με 2 τρόπους:

- *Moving Block Bootstrap*: Για παράδειγμα θεωρείστε τα δεδομένα: $(x_1, x_2, \ldots, x_{12})$. Κατασκευάζουμε 4 *block* από 3 παρατηρήσεις το κάθε ένα, συγκεκριμένα $y_1 = (x_1, x_2, x_3)$, $y_2 = (x_4, x_5, x_6)$, $y_3 = (x_7, x_8, x_9)$, και $y_4 = (x_{10}, x_{11}, x_{12})$. Τότε κάνουμε δειγματοληψία από τα y_i αντί για τις αρχικές παρατηρήσεις. Κρατάμε κάποιο κομμάτι της εξάρτησης αλλά το χάνουμε όταν ενώνουμε τα *block*. Κατα κάποιο τρόπο προσθέτουμε λευκό θόρυβο στα δεδομένα. Παρατηρείστε ότι η εξάρτηση για *lag* ≥ 3 εξαφανίζεται.
- *Overlapping blocks*: Κατασκευάζουμε τα *blocks* με επικάλυψη, π.χ. ορίζουμε $y_1 = (x_1, x_2, x_3)$, $y_2 = (x_2, x_3, x_4), \ldots, y_{11} = (x_{11}, x_{12}, x_1), y_{12} = (x_{12}, x_1, x_2)$. Αυτό προσθέτει λιγότερο λευκό θόρυβι αλλά ακόμα χάνουμε πληροφορία αναφορικά με την εξάρτηση.

Bootstrap για εξαρτημένα δεδομένα (2)

- Παραμετρική *Bootstrap*: Εφαρμόζουμε ένα παραμετρικό μοντέλο έτσι ώστε να αιχμαλωτίσουμε την δομή της εξάρτησης. Για παράδειγμα, για γραμμικές σχέσεις, με βάση τη θεωρία, μπορούμε να βρούμε ένα κατάλληλο μοντέλο *AR* το οποίο προσεγγίζει αρκετά ικανοποιητικά την δομή εξάρτησης της σειράς. Ή εναλλακτικά μπορούμε να συνδιάσουμε παραμετρικές ιδέες με *block bootstrapping*, με το να προσαρμόσουμε ένα μοντέλο και να χρησιμοποιήσουμε *block bootstrap* στα κατάλοιπα αυτού του μοντέλου.
- Υπάρχουν άλλες πολύ ποιο πολύπλοκες μέθοδοι κατάλληλες για συγκεκριμένους τύπους δεδομένων (π.χ. χωρική εξάρτηση κλπ).

Περισσότερες εφαρμογές

Η *bootstrap* βρίσκει εφαρμογές και σε πολλές άλλες περιπτώσεις τις οποίες δεν θα αναλύσουμε. Ενδεικτικά αναφέρουμε:

- Λογοκριμένα δεδομένα
- Προβλήματα με *Missing Data*
- Πεπερασμένοι Πληθυσμοί
- κλπ

Ωστόσο πριν κάποιος εφαρμόσει *bootstrap* κάποιος πρέπει να είναι σίγουρος ότι τα *bootstrap* δείγματα αποτελούν καλή εκτίμηση της άγνωστης πυκνότητας του πληθυσμού. Διαφορετικά η *bootstrap* δεν δουλεύει!

Bag of Little Bootstraps

- Μία νέα διαδικασία η οποία συνδιάζει τη *bootstrap* και το *subsampling*, και παίρνει το καλύτερο και από τις 2 μεθόδους.
- Δουλεύει με μικρά υποσύνολα των δεδομένων.
- Αλλά δεν απαιτεί αναλυτική αλλαγή κλίμακας.
- Δουλεύει καλά στην πράξη!

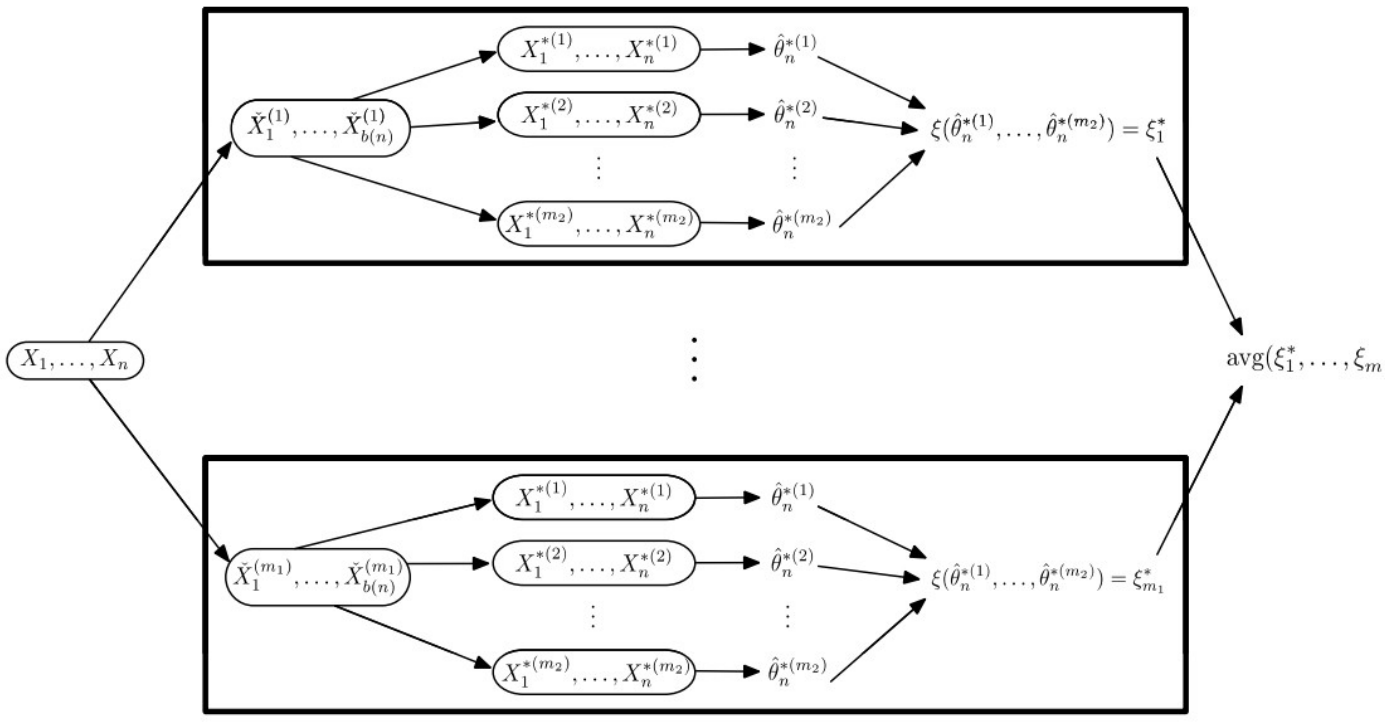
Bag of Little Bootstraps: η ιδέα

- πάρε ένα υποσύνολο των δεδομένων μεγέθους $b < n$
- *Bootstrap* n φορές με επανάθεση από αυτό το υποσύνολο.
- Επανέλαβε τη διαδικασία με διαφορετικά υποσύνολα.

Bag of Little Bootstraps: η ιδέα

- Μπορούμε (και οφείλουμε!) να κάνουμε δειγματοληψία με επανάθεση n φορές, και όχι b φορές.
- Άν επαναλάβουμε αυτή τη διαδικασία για ένα δεδομένο υποσύνολο μπορούμε να κατασκευάσουμε *Bootstraps* δ.ε. στη σωστή κλίμακα—χωρίς να είναι απαραίτητη αλλαγή κλίμακας!
- Αυτό το κάνουμε (με παράλληλο *processing*) για πολλαπλά υποσύνολα και συνδιάζουμε τα αποτελέσματα (π.χ. με το μέσο).

BLB


$X_1, \ldots, X_n$
$\check{X}_1^{(1)}, \ldots, \check{X}_{b(n)}^{(1)}$
$X_1^{*(1)}, \ldots, X_n^{*(1)}$
$X_1^{*(2)}, \ldots, X_n^{*(2)}$
$X_1^{*(m_2)}, \ldots, X_n^{*(m_2)}$
$\hat{\theta}_n^{*(1)}$
$\hat{\theta}_n^{*(2)}$
$\hat{\theta}_n^{*(m_2)}$
$\xi(\hat{\theta}_n^{*(1)}, \ldots, \hat{\theta}_n^{*(m_2)}) = \xi_1^*$
$\check{X}_1^{(m_1)}, \ldots, \check{X}_{b(n)}^{(m_1)}$
$X_1^{*(1)}, \ldots, X_n^{*(1)}$
$X_1^{*(2)}, \ldots, X_n^{*(2)}$
$X_1^{*(m_2)}, \ldots, X_n^{*(m_2)}$
$\hat{\theta}_n^{*(1)}$
$\hat{\theta}_n^{*(2)}$
$\hat{\theta}_n^{*(m_2)}$
$\xi(\hat{\theta}_n^{*(1)}, \ldots, \hat{\theta}_n^{*(m_2)}) = \xi_{m_1}^*$
$\text{avg}(\xi_1^*, \ldots, \xi_m$

Ιδιότητες

Η *BLB* όπως και η *bootstrap*, κάτω από τις ίδιες συνθήκες οι οποίες ίσχυαν και πριν στην *bootstrap* μοιράζονται τις ιδιότητες

- συνεπείς ασυμπτωτικά
- σωστές σε μεγάλη τάξη
- Γρήγορη σύγκλιση

Άρα στην πράξη δεν χάνουμε κάτι αλλά έχουμε ένα διπλό εργαλείο.