Εργασία Απαραμετρικής Στατιστικής (Μεταπτυχιακό)

Εργασία: Έστω $X_1, X_2, \ldots, X_n$ ένα τυχαίο δείγμα από κάποια απόλυτα συνεχή κατανομή. Ο στόχος είναι να εκτιμήσουμε την άγνωστη συνάρτηση πυκνότητας πιθανότητας f που βρίσκεται πίσω από την παραγωγή των δεδομένων. Έστω

$$\begin{aligned} f_1(x) &= 2\,(1 - 2\,|x - 0.5|\,)^+\,, \\ f_2(x) &= 2\cos^2(2\pi x)\,\mathbf{1}_{[0,1]}(x). \end{aligned}$$

Το πρόβλημα εκτίμησης της f θα πρέπει να εξεταστεί ξεχωριστά για κάθε μία από τις παραπάνω 2 περιπτώσεις, δηλαδή όταν $f = f_i$ για κάθε $i = 1, 2$. Παρακάτω αναφέρονται όλα τα ερωτήματα που πρέπει να απαντηθούν ξεχωριστά για κάθε μία από τις παραπάνω πυκνότητες. Ο κώδικας στην R πρέπει να συνοδεύεται από επαρκή σχόλια.

Ερωτήματα:

(1) Να γίνει ένα γράφημα της f.

(2) Να προταθεί ένας τρόπος προσομοίωσης από την f και να παραχθούν 100 δεδομένα από τη συγκεκριμένη κατανομή.

(3) Να γίνει ένας έλεγχος Kolmogorov-Smirnov ότι τα δεδομένα που έχετε παράγει προέρχονται πράγματι από την κατανομή με συνάρτηση πυκνότητας f. Να αξιολογηθεί η $p-value$ του ελέγχου.

(4) Να γίνει εκτίμηση πυκνότητας για $n = 20$ (χρησιμοποιώντας τα 20 πρώτα δεδομένα που έχετε παράγει) και $n = 100$ με τη μέθοδο του ιστογράμματος (ισομήκη διαστήματα) και με τη μέθοδο του πυρήνα, επιλέγοντας ως πυρήνα τον Epanechnikov και άλλον ένα πυρήνα δική σας επιλογής. Οι εκτιμήσεις πυκνότητας και στις 3 περιπτώσεις πρέπει να γίνουν με τη μέθοδο του cross-validation, είτε με τη βοήθεια έτοιμου πακέτου, είτε με δικό σας κώδικα.

(5) Στο παραπάνω ερώτημα υπάρχει μία βέλτιστη επιλογή της παραμέτρου ομαλοποίησης h^* σε κάθε τύπο εκτίμησης (3 περιπτώσεις). Επιλέξτε 2 ακόμα τιμές του h, ας πούμε h^+ και h^-, που να επιφέρουν αισθητό oversmoothing και undersmoothing αντίστοιχα. Συγκρίνετε τις τιμές των cross-validation σκορ, $\widehat{J}^*$, $\widehat{J}^+$ και $\widehat{J}^-$ αντίστοιχα.

(6) Η εκτιμήτρια $\widehat{J}^*$, είναι μία εκτιμήτρια του ελάχιστου ρίσκου. Να γίνει εφαρμογή της μεθόδου bootstrap για να εκτιμηθεί η μεροληψία της και το τυπικό σφάλμα αυτής της εκτιμήτριας για $n = 20$ και $n = 100$.

(7) Σε αυτό το ερώτημα να χρησιμοποιηθεί ως εκτιμήτρια πυκνότητας μόνο αυτή που δίνεται από τη μέθοδο με πυρήνες και συγκεκριμένα με πυρήνα Epanechnikov. Να υπολογιστεί η συνάρτηση κινδύνου

$$R(h) = \mathbf{E}(J(h)) = \mathbf{E}\left(\int \hat{f}^2_{n,h}(x)\,dx\right) - 2\mathbf{E}\left(\int \hat{f}_{n,h}(x) f(x)\,dx\right),$$

όταν $f = f_1$.

(8) Να βρεθεί το h^* που ελαχιστοποιεί την $R(h)$ με τη βοήθεια αριθμητικής βελτιστοποίησης. Να γίνει σύγκριση με το h^*_{cv} (ερώτημα 5) και το ασυμπτωτικά βέλτιστο h^*_∞.