Ε Θ Ν Ι Κ Ο Μ Ε Τ Σ Ο Β Ι Ο Π Ο Λ Υ Τ Ε Χ Ν Ε Ι Ο
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ & ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
Τομέας Μαθηματικών
Πολυτεχνειούπολη – Ζωγράφου ΑΘΗΝΑ - 157 80
ΤΗΛ. : 772 1774
FAX : 772 1775

(α) ΔΠΜΣ ΣΤΙΣ ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ
(β) ΔΠΜΣ ΣΤΗΝ ΜΑΘΗΜΑΤΙΚΗ ΠΡΟΤΥΠΟΠΟΙΗΣΗ ΣΕ ΣΥΓΧΡΟΝΕΣ ΤΕΧΝΟΛΟΓΙΕΣ ΚΑΙ ΣΤΗΝ ΟΙΚΟΝΟΜΙΑ
(γ) ΔΠΜΣ ΣΤΗΝ ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ
(δ) 8ο ΕΞΑΜΗΝΟ ΤΟΥ ΠΡΟΠΤΥΧΙΑΚΟΥ ΠΡΟΓΡΑΜΜΑΤΟΣ ΣΠΟΥΔΩΝ ΤΗΣ ΣΕΜΦΕ

ΤΙΤΛΟΣ ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΜΑΘΗΜΑΤΟΣ: *ΥΠΟΛΟΓΙΣΤΙΚΗ ΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΣΤΟΧΑΣΤΙΚΗ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ*
ΤΙΤΛΟΣ ΠΡΟΠΤΥΧΙΑΚΟΥ ΜΑΘΗΜΑΤΟΣ: *ΥΠΟΛΟΓΙΣΤΙΚΕΣ ΜΕΘΟΔΟΙ ΣΤΗ ΣΤΑΤΙΣΤΙΚΗ δ*
ΔΙΔΑΣΚΟΝΤΕΣ: *ΔΗΜΗΤΡΗΣ ΦΟΥΣΚΑΚΗΣ (τηλ: 210 7721702 – email: fouskakis@math.ntua.gr) & ΜΙΧΑΛΗΣ ΛΟΥΛΑΚΗΣ (τηλ: 210 7721689 – email: loulakis@math.ntua.gr)*

ΕΡΓΑΣΙΑ

1. (α) Γράψτε μια συνάρτηση στην R που εκτιμά με την απλή μέθοδο *Monte Carlo* τον όγκο της μοναδιαίας σφαίρας σε d διαστάσεις $B_d = \{(x_1,\ldots,x_d): x_1^2+\cdots+x_d^2<1\}$ και το % σφάλμα της εκτίμησης αυτής σε σχέση με την ακριβή τιμή V_d του όγκου της σφαίρας. Η συνάρτηση σας θα πρέπει να επιλέγει τυχαία και ομοιόμορφα $N = 10^5$ σημεία στον κύβο $[-1, 1]^d$ και να ελέγχει πόσα από αυτά πέφτουν μέσα στη σφαίρα B_d. Θα πρέπει επίσης να ξεκινά από $d = 2$ και να ανεβαίνει κατά μία διάσταση έως ότου κανένα από τα N σημεία στον κύβο να μην πέσει μέσα στη σφαίρα. Για την ακριβή τιμή V_d, του όγκου της σφαίρας, μπορείτε να χρησιμοποιήσετε ότι $V_1 = 2, V_2 = \pi$ και $V_{d+2} = \frac{2\pi}{d+2}V_d$, για $d = 1, 2, \ldots$

(β) Γράψτε μια συνάρτηση στην R που εκτιμά τον όγκο V_d με τη βοήθεια του αλγορίθμου *Metropolis-Hastings*. Η μαρκοβιανή αλυσίδα που θα θεωρήσετε να έχει χώρο καταστάσεων τον κύλινδρο $C_d = B_{d-1} \times [-1, 1]$ και αναλλοίωτη κατανομή την ομοιόμορφη στον C_d. Μπορείτε να πάρετε ως προτεινόμενη κατανομή κάθε βήματος εκείνη που

- από τις d -1 πρώτες συντεταγμένες επιλέγει τυχαία μία και την μεταβάλλει κατά μια ομοιόμορφη μεταβλητή στο (-ε, ε)

και

- την τελευταία συντεταγμένη την επιλέγει ομοιόμορφα στο (-1, 1).

Αν στο εργοδικό θεώρημα θεωρήσετε τη δείκτρια συνάρτηση της σφαίρας B_d, θα πάρετε μια εκτίμηση για το λόγο $\frac{V_d}{2 \times V_{d-1}}$. Ξεκινώντας από d = 3, μπορείτε έτσι να εκτιμήσετε επαγωγικά τον όγκο V_d για d = 3, 4,… μέχρι τη διάσταση d_{max} για την οποία το σφάλμα της εκτίμησης ξεπερνά για πρώτη φορά το 10%.

Πειραματιστείτε με την τιμή του ε και τα βήματα για το *burn-in* ώστε με την *MCMC* εκτιμήτρια που χρησιμοποιεί N βήματα της αλυσίδας να φτάσετε σε όσο το δυνατόν μεγαλύτερη τιμή για το d_{max}. Στην αναφορά σας να δώσετε τις τιμές που χρησιμοποιήσατε για τις παραπάνω παραμέτρους, τις εκτιμήσεις σας για τον όγκο της σφαίρας d διαστάσεων και το % σφάλμα της εκτίμησης για d = 2,3,…,d_{max}.

2. Στη βιβλιοθήκη `MASS` της R θα βρείτε τα δεδομένα `mcycle`. Αρχικά κατεβάστε και φορτώστε τη βιβλιοθήκη και εν συνεχεία με χρήση της εντολής `data(mcycle)` φορτώστε τα δεδομένα. Τα δεδομένα έχουν πληροφορία για n = 133 προσομοιώσεις ατυχήματος μοτοσυκλέτας, που χρησιμοποιήθηκαν για τη δοκιμή κρανών. Σκοπός σας είναι να χρησιμοποιήσετε ένα μοντέλο παλινδρόμησης με μεταβλητή απόκρισης την `mcycle$accel` (επιτάχυνση κεφαλής σε g) και επεξηγηματική μεταβλητή την `mcycle$times` (χρόνος σε χιλιοστά του δευτερολέπτου από τη στιγμή της κρούσης).

(α) Αρχικά, έστω ότι θέλετε να εκτιμήσετε την σ.π.π. f(x) από όπου προέρχονται οι τιμές `mcycle$times`. Έστω ότι θέλετε να χρησιμοποιήσετε Κανονικό πυρήνα. Βρείτε το βέλτιστο πλάτος h μεγιστοποιώντας την *cross-validated* πιθανοφάνεια, με χρήση δικού σας κώδικα στην R. Προβείτε σε ένα διάγραμμα της εκτιμώμενης f(x) για το h που βρήκατε, χρησιμοποιώντας την έτοιμη συνάρτηση `density` με Κανονικό πυρήνα και σχολιάστε το αποτέλεσμα που πήρατε.

(β) Εν συνεχεία θέλετε να προσαρμόστε το μοντέλο μη παραμετρικής παλινδρόμησης (*Nadaraya-Watson*) με μεταβλητή απόκρισης την `mcycle$accel` και επεξηγηματική μεταβλητή την `mcycle$times`. Για την εύρεση του h χρησιμοποιήστε *leave-one-out CV* με χρήση δικού σας κώδικα στην R. Για εύρος τιμών h από το 1.01 μέχρι και το 10 (με βήμα 0.01) υπολογίστε το CV-MSE του *Nadaraya-Watson* εκτιμητή με Κανονικό πυρήνα, για κάθε h, και διαλέξτε το h εκείνο που το ελαχιστοποιεί. Παρουσιάστε το διάγραμμα διασποράς και συγχρόνως την προσαρμοσμένη καμπύλη (για το h που βρήκατε) και σχολιάστε.

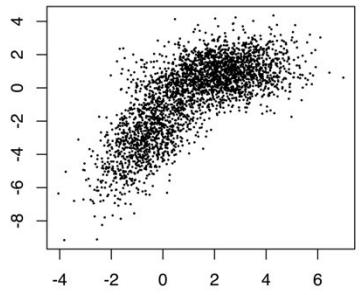
3. Στο αρχείο

http://www.math.ntua.gr/~fouskakis/Computational_Stats/Assign/datafile.csv

βρίσκονται οι συντεταγμένες 3000 σημείων, τα οποία απεικονίζονται στο παρακάτω Σχήμα. Χρησιμοποιώντας την εντολή

```
data<-unname(as.matrix(read.csv2("datafile.csv")))
```

διαβάστε τα δεδομένα στην R.



(α) Να προσαρμόσετε στα δεδομένα του αρχείου ένα μοντέλο κανονικής κατανομής σε δύο διαστάσεις $N(\mathbf{m},\Sigma)$, βρίσκοντας αναλυτικά τις εκτιμήτριες μέγιστης πιθανοφάνειας των άγνωστων παραμέτρων **m** και Σ.

(β) Με τη βοήθεια της R, να προσομοιώσετε 3000 ανεξάρτητες παρατηρήσεις από την κατανομή που εκτιμήσατε στο ερώτημα (α) και να συγκρίνετε το διάγραμμά τους με το δοσμένο Σχήμα.

(γ) Θέλουμε τώρα να προσαρμόσουμε στα δεδομένα του αρχείου ένα μοντέλο μείξης κανονικών κατανομών

$$f_X(x) = p f_1(x) + (1-p) f_2(x),$$

όπου f_1, f_2 είναι οι συναρτήσεις πυκνότητας πιθανότητας των κανονικών κατανομών δύο διαστάσεων $N(\mathbf{m_1},\Sigma_1)$ και $N(\mathbf{m_2},\Sigma_2)$ και $p \in (0,1)$. Πειστείτε αρχικά ότι δεν μπορείτε να υπολογίσετε αναλυτικά τις εκτιμήτριες μέγιστης πιθανοφάνειας των άγνωστων παραμέτρων p, $\mathbf{m_1}$, Σ_1, $\mathbf{m_2}$, και Σ_2 από τις δοσμένες παρατηρήσεις. Στη συνέχεια, θεωρώντας την λανθάνουσα μεταβλητή $Z \sim$ Bernoulli(p) που κωδικοποιεί την κατανομή της μείξης από την οποία προέρχεται κάθε παρατήρηση, υλοποιήστε στην R τον αλγόριθμο *Expectation-Maximization* για να εκτιμήσετε τις άγνωστες παραμέτρους, αρχικοποιώντας τις όπως νομίζετε, και τερματίζοντας τον αλγόριθμο όταν το άθροισμα της απόλυτης μεταβολής όλων των παραμέτρων κατά την εκτέλεση ενός βήματος είναι μικρότερο από 10^{-6}. Να δείξετε σε έναν πίνακα τις τρέχουσες τιμές των άγνωστων παραμέτρων σε κάθε βήμα του αλγορίθμου μέχρι τον τερματισμό.

Υπόδειξη: Στο βήμα *Maximation* ίσως βρείτε χρήσιμες τις αναλυτικές εκφράσεις που υπολογίσατε στο ερώτημα (α).

(δ) Με τη βοήθεια της R, να προσομοιώσετε 3000 ανεξάρτητες παρατηρήσεις από το μοντέλο μείξης που εκτιμήσατε στο ερώτημα (γ) και να συγκρίνετε το διάγραμμά τους με εκείνο που κατασκευάσατε στο ερώτημα (β) και με το δοσμένο Σχήμα.

4. Στην βιβλιοθήκη `faraway` της R θα βρείτε τα δεδομένα `fat`. Αρχικά κατεβάστε και φορτώστε τη βιβλιοθήκη και εν συνεχεία με χρήση της εντολής `data(fat)` φορτώστε τα δεδομένα. Τα δεδομένα έχουν πληροφορία για n = 252 άνδρες, σχετικά με ένα ποσοτικό χαρακτηριστικό `fat$brozek`, που μετρά το ποσοστό σωματικού λίπους χρησιμοποιώντας την εξίσωση *Brozek*, και p = 17 επεξηγηματικές μεταβλητές (για περισσότερες πληροφορίες επισκεφτείτε τον σύνδεσμο https://rdrr.io/cran/faraway/man/fat.html). Αρχικά τυποποιήστε τις τιμές των επεξηγηματικών μεταβλητών, ώστε να έχουν άθροισμα 0 και άθροισμα τετραγώνων ίσο με 1 (προσοχή αυτό δεν σημαίνει ότι η τυπική απόκλιση είναι 1). Μην χρησιμοποιήσετε τις επεξηγηματικές μεταβλητές `siri`, `density` και `free` σε κανένα από τα παρακάτω ερωτήματα.

(α) Τυχαία χρησιμοποιήστε το 4/5 (περίπου) των παρατηρήσεων σας ως *training data* και τις υπόλοιπες ως *test data*.

(β) Εφαρμόστε τη μεθοδολογία *Lasso* με τη βοήθεια της βιβλιοθήκης `glmnet` της R στα *training data* και σχολιάστε τα αποτελέσματα.

(γ) Χρησιμοποιώντας *cross-validation* επιλέξτε την παράμετρο ποινής λ, με χρήση της έτοιμης συνάρτησης `cv.glmnet`, που ελαχιστοποιεί το CV-MSE. Χρησιμοποιώντας την εν λόγω τιμή καταλήξτε σε ένα μοντέλο, χωρίς κάποιες από τις επεξηγηματικές μεταβλητές, το οποίο καλέστε M1. Επαναλάβετε την ίδια διαδικασία επιλέγοντας ως λ την τιμή που ελαχιστοποιεί το CV-MSE με σφάλμα εντός μιας τυπικής απόκλισης από την ελάχιστη τιμή. Καλέστε το μοντέλο αυτό M2. Συγκρίνετε το MSE των μοντέλων M1 και M2 στα *test data* και επιλέξτε το καλύτερο. Καλέστε αυτό το μοντέλο M_Lasso.

(δ) Σε όλα τα δεδομένα προσαρμόστε το πολλαπλό γραμμικό μοντέλο χρησιμοποιώντας μόνο τις επεξηγηματικές μεταβλητές από το μοντέλο M_Lasso. Καλέστε αυτό το μοντέλο M3.

(ε) Εξερευνώντας πλήρως τον χώρο όλων των πιθανών μοντέλων στο πρόβλημα επιλογής επεξηγηματικών μεταβλητών, με τη βοήθεια δικής σας συνάρτησης στην R, βρείτε το μοντέλο εκείνο που ελαχιστοποιεί την τιμή του κριτηρίου BIC (για τον υπολογισμό του BIC μπορείτε να χρησιμοποιήσετε έτοιμη συνάρτηση της R). Καλέστε το εν λόγω μοντέλο M4. Χρησιμοποιήστε όλα τα διαθέσιμα δεδομένα και όλες τις επεξηγηματικές μεταβλητές (εκτός των `siri`, `density` και `free`). Αν το μοντέλο M4 έχει τις ίδιες επεξηγηματικές μεταβλητές με αυτές του M3, διαλέξτε ως M4 αυτό με την δεύτερη μικρότερη τιμή για το BIC.

(στ) Χρησιμοποιώντας *5-fold cross-validation* και την (*within fold*) ARMSE (*Average Root Mean Square Error*) συνάρτηση εξετάστε ποιο από τα δύο μοντέλα, M3 και M4, έχει την καλύτερη προβλεπτική ικανότητα, με χρήση δικού σας κώδικα στην R.

(ζ) Θεωρήστε το τελικό μοντέλο που επιλέξατε και όλα τα διαθέσιμα δεδομένα. Χρησιμοποιώντας 1000 *Bootstrap* δείγματα από τα υπόλοιπα, δώστε εκτιμήτριες και τυπικά σφάλματα για τους συντελεστές του εν λόγω μοντέλου χρησιμοποιώντας έτοιμες συναρτήσεις στην R από την βιβλιοθήκη `bootstrap`. Προβείτε σε συγκρίσεις και σχολιασμό με τα αποτελέσματα που παίρνετε με χρήση της εντολής `lm` στην R.

Οδηγίες

- Η εργασία θα πρέπει **να υποβληθεί ηλεκτρονικά στη σελίδα του μαθήματος στο HELIOS**, **μέχρι την Παρασκευή 1 Ιουλίου 2022 στις 13:00μμ. Καμιά εργασία δεν θα γίνει δεκτή μετά την ώρα αυτή.**
- Η εργασία που θα παραδώσετε πρέπει να είναι **σε pdf μορφή αφού πρώτα τη γράψετε υποχρεωτικά σε Latex**. Ο κώδικας θα πρέπει **υποχρεωτικά να είναι σε R**.
- Η εργασία σας μπορεί να είναι είτε στα Ελληνικά είτε στα Αγγλικά.
- Παρακαλώ χρησιμοποιήστε τον **ακόλουθο τίτλο στο pdf αρχείο σας**: Surname-Name.pdf, όπου Surname είναι το επώνυμό σας (με λατινικούς χαρακτήρες) και Name το όνομα σας (με λατινικούς χαρακτήρες). Π.χ. αν

παρέδιδε ο Δημήτρης Φουσκάκης εργασία, θα την ονόμαζε ως εξής: Fouskakis-Dimitris.pdf.

- Παρακαλώ χρησιμοποιήστε **ένα εξώφυλλο στο pdf αρχείο σας**, στο οποίο να υπάρχει κατάλληλος τίτλος και να αναγράφεται **υποχρεωτικά το ονοματεπώνυμο σας, το πρόγραμμα (προπτυχιακό ή μεταπτυχιακό που παρακολουθείτε) καθώς και το email σας και ο αριθμός μητρώου σας.**
- Θα πρέπει να **αποστείλετε ένα μόνο αρχείο**. Η εργασία θα πρέπει να περιλαμβάνει τους κώδικες της R, όχι σε παράρτημα αλλά στην απάντηση του κάθε ερωτήματος, με πλήρη επεξήγηση, γραφήματα και πλήρη περιγραφή των αποτελεσμάτων.
- Θα δοθεί ιδιαίτερη σημασία στην παρουσίαση της εργασίας. Η εργασία πρέπει να είναι κατανοητή και να περιγράφει οτιδήποτε χρησιμοποιήσατε πειστικά για κάποιον που δεν γνωρίζει πάρα πολλά για το αντικείμενο.

Ευχόμαστε Επιτυχία