ΑΝΑΛΥΤΙΚΟ ΥΠΟΜΝΗΜΑ:
ΕΠΙΣΤΗΜΟΝΙΚΗ ΚΑΙ ΕΡΕΥΝΗΤΙΚΗ ΔΡΑΣΤΗΡΙΟΤΗΤΑ

Φουσκάκης Δημήτρης

(τελευταία ενημέρωση 12 Απριλίου 2022)

Δημοσιεύσεις με Κρίση σε Περιοδικά

Δ1. **Fouskakis, D.** and Draper, D. (1999). Tabu Search - Book review, *Journal of the Royal Statistical Society Series D*, **48,** 616-619.

Πρόκειται για μια κριτική του βιβλίου "Tabu Search" του 1997, των F. Glover και M. Laguna (*Kluwer Academic Publishers*). Παρουσιάζονται τα θετικά και αρνητικά σημεία του βιβλίου αυτού, του μοναδικού έως τώρα βιβλίου για τη Μέθοδο Απαγορευμένης Αναζήτησης (*Tabu Search*), μία μέθοδος η οποία δεν είναι ευρέως γνωστή στο στατιστικό κοινό.

Δ2. Draper, D. and **Fouskakis, D.** (2000). A case study of stochastic optimization in health policy: problem formulation and preliminary results. *Journal of Global Optimization*, **18,** 399-416.

Χρησιμοποιούμε την Μπεϋζιανή Θεωρία Αποφάσεων για να λύσουμε ένα πρόβλημα επιλογής μεταβλητών, το οποίο προκύπτει στην προσπάθεια μας να μετρήσουμε έμμεσα την ποιότητα της νοσοκομειακής περίθαλψης, συγκρίνοντας το πραγματικό με το αναμενόμενο ποσοστό θνησιμότητας ασθενών εντός των πρώτων 30 ημερών νοσηλείας τους, λαμβάνοντας υπ' όψιν την βαρύτητα της ασθένειας τους κατά την εισαγωγή τους στο νοσοκομείο. Για την επιλογή των κατάλληλων μεταβλητών, η μέθοδος μας λαμβάνει υπ' όψιν πέραν από την ακρίβεια της πρόβλεψης του ποσοστού θνησιμότητας και το κόστος συλλογής των δεδομένων. Για την εύρεση του κατάλληλου υποδείγματος, μεγιστοποιείται η αναμενόμενη συνάρτηση χρησιμότητας, χρησιμοποιώντας *Monte Carlo* και *Cross Validation* τεχνικές. Επιπλέον, εξερευνούμε τον γεωμετρικό χώρο όλων των δυνατών λύσεων, και εν συνεχεία συγκρίνουμε διάφορους αλγόριθμους στοχαστικής βελτιστοποίησης, όπως τη ***Μέθοδο Προσομοιούμενης Ανόπτησης*** (*Simulated Annealing (SA)*), το ***Γενετικό Αλγόριθμος*** (*Genetic Algorithm (GA)*), τη ***Μέθοδος Απαγορευμένης Αναζήτησης*** (*Tabu Search (TS)*), τον ***Αλγόριθμος Αποδοχής Κατωφλιού*** (*Threshold Acceptance (TA)*) και τον ***Ακατάστατο Αλγόριθμο Προσομοιούμενης Ανόπτησης*** (*Messy Simulated Annealing (MSA)*). Συζητείται επίσης ο ρόλος του *N*, του αριθμού επανάληψεων της *Cross Validation* τεχνικής, στο πρόβλημα μεγιστοποίησης. Τα πρώτα αποτελέσματα υποδεικνύουν ότι ο *TS* υπερτερεί των *TA* και *SA*, ενώ ο *MSA* και ο *GA* έχουν την χειρότερη επίδοση.

Δ3. Gunnell, D., Harrison, G., Rasmussen, F., **Fouskakis, D.** and Tynelius, P. (2002). Associations between pre-morbid intellectual performance, early-life exposures and early onset schizophrenia: a cohort. *British Journal of Psychiatry*, **181**, 298-305.

Διάφορες έρευνες έχουν δείξει την ύπαρξη σημαντικής συσχέτιση μεταξύ της χαμηλής διανοητικής απόδοσης και της ανάπτυξης σχιζοφρένειας. Σε μια μελέτη κοορτής, δείγμα 197.613 Σουηδών ανδρών παρακολουθήθηκε από την ηλικία 18 ετών και για τα επόμενα 8

έτη. Συλλεχθήκαν δεδομένα που αφορούν την γέννησης τους, την εκπαίδευση των γονέων τους, τη διανοητική τους απόδοση κατά την διάρκεια της στρατιωτικής τους θητείας καθώς και τον αριθμό των εισαγωγών τους σε νοσοκομεία λόγω ψυχιατρικών διαταραχών.

Από τους 109.643 άνδρες με πλήρη δεδομένα, οι 60 (0.05%) ανέπτυξαν σχιζοφρένεια και οι 92 (0.08%) μη-συναισθηματική, μη-σχιζοφρενή ψύχωση. Ο μέγιστος κίνδυνος για σχιζοφρένειας παρατηρήθηκε στους απογόνους μορφωμένων γονέων που οι ίδιοι είχαν χαμηλό διανοητικό δείκτη.

Τα αποτελέσματα συμφωνούν με αυτά προηγούμενων ερευνών και δείχνουν ότι η χαμηλή διανοητική απόδοση συνδέεται με τις πρόωρες ψυχωτικές αναταραχές.

Δ4. **Fouskakis, D.** and Draper, D. (2001). Stochastic optimization: a review. *International Statistical Review*, **70**, 315-349.

Στην εργασία αυτή, αναλύουμε λεπτομερώς τρεις μεθόδους στοχαστικής βελτιστοποίησης, τη ***Μέθοδο Προσομοιούμενης Ανόπτησης*** (*Simulated Annealing (SA)*), το ***Γενετικό Αλγόριθμος*** (*Genetic Algorithm (GA)*), τη ***Μέθοδος Απαγορευμένης Αναζήτησης*** (*Tabu Search (TS)*). Σε κάθε περίπτωση, παρουσιάζεται ο ακριβής αλγόριθμος, προτερήματα και μειονεκτήματα, καθώς επίσης και προτεινόμενες από την βιβλιογραφία αρχικές τιμές. Για την υλοποίηση των εν λόγω αλγορίθμων χρησιμοποιούμε ένα πρόβλημα επιλογής μεταβλητών σε ένα μοντέλο λογιστικής παλινδρόμησης, το οποίο προκύπτει από την προσπάθεια μας να μετρήσουμε έμμεσα την ποιότητα της νοσοκομειακής περίθαλψης, συγκρίνοντας το πραγματικό με το αναμενόμενο ποσοστό θνησιμότητας ασθενών, λαμβάνοντας υπόψιν μας την βαρύτητα της ασθένειας τους κατά την εισαγωγή τους στο νοσοκομείο.

Δ5. Harrison, G., **Fouskakis, D.,** Rasmussen, F., Tynelius, P. and Gunnell, D. (2003). Association between psychotic disorder and urban place of birth is not mediated by obstetric complications or childhood socio-economic position: a cohort study. *Psychological Medicine*, **33**, 1-9.

Αν και ο αστικός τόπος γέννησης αποτελεί, βάσει δημοσιεύσεων, παράγοντα κινδύνου για τη σχιζοφρένεια, το μέγεθος του κινδύνου είναι πιθανό να ποικίλει στους διαφορετικούς πληθυσμούς. Σε αυτήν την εργασία έχουμε δείγμα 659.310 ανδρών και γυναικών από την Σουηδία, και εξετάζουμε την συσχέτιση του τόπου γεννήσεως με την ανάπτυξη μελλοντικών κρουσμάτων σχιζοφρένειας, αφού λάβουμε υπ' όψιν διάφορες μαιευτικές επιπλοκές, την εμβρυϊκή διατροφή και το επίπεδο εκπαίδευσης των γονέων. Ο μέγιστος κίνδυνος μη-συναισθηματικής ψύχωσης παρατηρήθηκε στους ανθρώπους με αστικό τόπο γέννησης.

Δ6. Gunnell, D., Rasmussen, F., **Fouskakis, D.,** Tynelius, P. and Harrison, G. (2003). Patterns of fetal and childhood growth and the development of psychosis in young males: a cohort study. *American Journal of Epidemiology*, **158**, 291-300.

Η σημασία της συσχέτισης των διαφόρων παραγόντων στην προ-γενετική και παιδική περίοδο με τον κίνδυνο ανάπτυξης σχιζοφρένειας είναι ασαφής. Έρευνες προτείνουν ότι οι μαιευτικές επιπλοκές, η περιοχή και η εποχή της γέννησης μπορούν να είναι σημαντικοί παράγοντες κινδύνου.

Τα δεδομένα μας αποτελούνται από 330.000 Σουηδούς άνδρες ηλικίας 17-25. Από τους 247.814 άνδρες με πλήρη στοιχεία, 204 ανέπτυξαν τις μη-συναισθηματικές ψυχώσεις (80

περιπτώσεις σχιζοφρένιας και 124 μη-συναισθηματικής, μη-σχιζοφρενούς ψύχωσης). Παρατηρήθηκε μια μη γραμμική σχέση μεταξύ του βάρους κατά την γέννηση και της ανάπτυξης σχιζοφρένειας. Τα νεογνά με πολύ χαμηλό (< 2.5kg) ή με πολύ υψηλό (> 4.0kg) βάρος είχαν τον μεγαλύτερο κίνδυνο. Επίσης νεογνά γεννημένα τους θερινούς μήνες είχαν σχεδόν τις μισές πιθανότητες ανάπτυξης σχιζοφρένιας έναντι εκείνων που γεννήθηκαν το φθινόπωρο ή το χειμώνα.

Δ7. **Fouskakis, D.,** Gunnell, D., Rasmussen, F., Tynelius, P., Sipos, A. and Harrison, G. (2004). Is the season of birth association with psychosis due to seasonal variations in foetal growth or other related exposures? A cohort study. *Acta Psychiatrica Scandinavica*, **109**, 1-5.

Πολλές έρευνες έχουν δείξει συσχέτιση μεταξύ εποχής γέννησης και ανάπτυξης ψυχώσεων. Σε μια μελέτη κοορτής, 695.310 ανδρών και γυναικών από την Σουηδία γεννημένων το 1973 - 1980, δεν εντοπίσθηκε συσχέτιση μεταξύ εποχής γεννήσεως και κρουσμάτων ψυχώσεων.

Δ8. Korkolis, D., Tsoli, E., **Fouskakis, D.**, Yiotis, J., Koullias, G.J., Giannopoulos, D., Papalambros, E., Patsounis, E., Asimacopoulos, P. and Gorgoulis, V.G. (2004). Tumor Histology and Stage but not P53, Her2-neu or Cathepsin-D Expression are Independent Prognostic Factors in Breast Cancer Patients. *Anticancer Research*, **24**, 2061-2068.

Έχουμε πραγματοποιήσει μια στατιστική μελέτη της προγνωστικής αξίας των συνήθων παραγόντων και των διερευνητικών παραγόντων *p53*, *Her2-neu* και *Cathepsin-D* στους ασθενείς με καρκίνο στο στήθος. Η ανάλυσή μας επεκτάθηκε για να καθορίσει τις συσχετίσεις των *p53* και *Her2-neu* με τον κίνδυνο θανάτου και την υποτροπή στους ασθενείς με και χωρίς μεταστάσεις λεμφαδένων. Προέκυψε ότι η επιβίωση σχετίζεται με τον ιστολογικό τύπο και τη σταδιοποίηση του καρκίνου.

Δ9. Naska, A., **Fouskakis, D.**, Oikonomou, E., Almeida, M.D.V., Berg, M.A., Gedrich, K., Moreiras, O., Nelson, M., Trygg, K., Turrini, A., Remaut, A.M., Volatier, J.L., Trichopoulou, A. and DAFNE participants (2006). Dietary patterns and their socio-demographic determinants in ten European countries. Data from the DAFNE databank. *European Journal of Clinical Nutrition*, **60**, 181-190.

Σκοπός της συγκεκριμένης εργασίας είναι να περιγράψει διαιτητικά πρότυπα δέκα Ευρωπαϊκών χωρών, λαμβάνοντας υπόψη κοινωνικό-οικονομικούς παράγοντες και χρησιμοποιώντας τα δεδομένα από την Πανευρωπαϊκή έρευνα *DAFNE*. Χρησιμοποιώντας πολυμεταβλητές στατιστικές τεχνικές, καταλήξαμε στο συμπέρασμα ότι οι διατροφικές διαφορές βόρειο - Ευρωπαίων και νότιο – Ευρωπαίων έχουν μειωθεί.

Δ10. Kokolakis, G., Nanopoulos, Ph. and **Fouskakis, D.** (2006). Bregman Divergences in the (m × k) – Partitioning Problem. *Computational Statistics and Data Analysis*, **51**, 668-678.

Η επιλογή ενός κατάλληλου μέτρου για να εκφράσει τον βαθμό της πληροφορίας που χάνεται κατά την μικρο-ομαδοποίηση (*micro-aggregation*) στατιστικών δεδομένων έχει βαρύνουσα σημασία. Αντί των συνήθων, και μάλλον περιορισμένων, "μέτρων απώλειας πληροφορίας" που βασίζονται στην Ευκλείδεια Απόσταση (όπως το κριτήριο $L = SSW/SST$), είτε στην *Kullback-Liebler Divergence* $J = \int \ln\{P/Q\}dP$) και σε παραλλαγές αυτών, στην

εργασία αυτή κάνουμε εφαρμογή των "*Bregman Divergences*". Πρόκειται για μια γενική κλάση πληροφοριακών μέτρων "απόκλισης τυχαίων μεταβλητών και κατανομών", η οποία παράγεται από το σύνολο σχεδόν όλων των κυρτών συναρτήσεων, εμπεριέχει τα περισσότερα από τα γνωστά πληροφοριακά μέτρα, και έχει την Πυθαγόρεια ιδιότητα. Επίσης, όπως πολύ πρόσφατα αποδείχθηκε, είναι η γενικότερη κλάση πληροφοριακών μέτρων απόκλισης τυχαίων μεταβλητών και κατανομών τα οποία "ταυτίζουν" την βέλτιστη λύση με την δεσμευμένη μέση τιμή. Στην εργασία αυτή αποδεικνύεται ένα αρκετά ενδιαφέρον και γενικό αποτέλεσμα: Στο πρόβλημα της διαμέρισης ενός συνόλου στατιστικών δεδομένων σε ομάδες με δεδομένο μέγεθος, απ' όπου και η ονομασία *(m × k) – Partitioning Problem*, η βέλτιστη διαμέριση διαμορφώνεται: (α) από κέντρα των ομάδων και την επιλεγείσα κυρτή συνάρτηση, μέσω της οποίας κατασκευάστηκε η εφαρμοζόμενη *Bregman Divergence*, και (β) από το μέτρο κατανομής πιθανότητας των σημείων. Έτσι, η βέλτιστη ($m \times k$) – διαμέριση δεν είναι κατ'ανάγκη κυρτή, εκτός εάν πρόκειται για ομοιόμορφα κατανεμημένα μέτρα πιθανότητας. Επιπλέον παρουσιάζεται ένα κριτήριο το οποίο αποδεικνύεται ότι αποτελεί αναγκαία συνθήκη για το "βέλτιστο" της λύσης και με βάση αυτό επιτυγχάνουμε βελτίωση προγενέστερης μεθοδολογίας προσδιορισμού της βέλτιστης διαμέρισης.

Δ11. Kokolakis, G. and **Fouskakis, D.** (2008). On the Discrepancy Measures for the Optimal Equal Probability Partitioning in Bayesian Multivariate Micro-aggregation. *Journal of Classification*, **25**, 209-224.

Στην εργασία αυτή μελετάμε το πρόβλημα της Προστασίας των Προσωπικών Στατιστικών Δεδομένων (*SDC: Statistical Disclosure Control*) με Μπεϋζιανή προσέγγιση και τη μέθοδο της μικρο-ομαδοποίησης (*micro-aggregation*). Το κριτήριο είναι η ελαχιστοποίηση πληροφοριακών μέτρων απόκλισης ύστερων κατανομών. Με την ελαχιστοποίηση των μέτρων αυτών επιτυγχάνεται η διατήρηση του στατιστικού πληροφοριακού περιεχομένου των δεδομένων με παράλληλη προστασία των προσωπικών δεδομένων. Τα πληροφοριακά μέτρα που εφαρμόζονται εδώ είναι: η *Kullback-Liebler Divergence*, η *Symmetric Kullback-Liebler Divergence*, η *Hellinger's Distance*, η *Bhattacharyya Distance* και η *Chernoff's Distance*. Παρότι, όπως είναι γνωστό, τα παραπάνω πληροφοριακά μέτρα απόκλισης δεν οδηγούν κατ' ανάγκη σε ταυτόσημα στατιστικά συμπεράσματα, εν τούτοις στην εργασία αυτή διαπιστώνεται ότι εφαρμοζόμενα στις ύστερες κατανομές των παραμέτρων $\boldsymbol{\mu}$ και $\boldsymbol{\Sigma}$ των πολυμεταβλητών Κανονικών κατανομών, πριν και μετά την μικρο-ομαδοποίηση, μας οδηγούν στον ίδιο χαρακτηρισμό της βέλτιστης διαμέρισης. Η εν λόγω προσέγγιση μας επιτρέπει να δώσουμε "ρεαλιστική" απάντηση στο γνωστό ως *k-means partitioning problem*, του οποίου η υπολογιστική πολυπλοκότητα είναι *NP-hard*, κατασκευάζοντας προσεγγιστικά βέλτιστες διαμερίσεις στατιστικών δεδομένων σε χρόνο πολυωνιμικό.

Δ12. **Fouskakis, D.** and Draper, D. (2008). Comparing stochastic optimization methods for variable selection in binary outcome prediction with application to health policy. *Journal of the American Statistical Association*, **103**, 1367-1381.

Στο πεδίο της αξιολόγησης της ποιότητας των υπηρεσιών Υγείας, η σοβαρότητα της κατάστασης των ασθενών κατά την εισαγωγή τους σε μια Νοσοκομειακή μονάδα εκτιμάται κυρίως με χρήση λογιστικής παλινδρόμησης με μεταβλητή απόκρισης τη θνησιμότητα τους μετά από 30 ημέρες νοσηλείας και επεξηγηματικές μεταβλητές ένα σχετικά μεγάλο αριθμό δεικτών νοσηρότητας (περίπου της τάξης των 100). Χρησιμοποιώντας συνηθισμένες μεθόδους κλιμακωτών διαδικασιών επιλογής μεταβλητών καταλήγουμε σε ένα βέλτιστο

υποσύνολο δεικτών (συνήθως της τάξης των 10-20) που τελικά χρησιμοποιούνται για την εκτίμηση της αναμενόμενης πιθανότητας επιβίωσης του κάθε ασθενή.

Με την παραπάνω μέθοδο δεν λαμβάνεται όμως υπόψη το κόστος συλλογής κάθε δείκτη, το οποίο είναι ιδιαίτερα σημαντικό στην σύγκριση και αξιολόγηση μεγάλου πλήθους νοσοκομειακών μονάδων. Η εισαγωγή του κόστους στο πρόβλημα επιλογής μεταβλητών δημιουργεί ένα σύνθετο πρόβλημα βελτιστοποίησης.

Το παραπάνω πρόβλημα μπορεί να διατυπωθεί στα πλαίσια της Μπεϋζιανής Θεωρίας Βέλτιστων Αποφάσεων. Για την επίλυση του πολύπλοκου προβλήματος βελτιστοποίησης χρησιμοποιούμε στοχαστικούς αλγόριθμους βελτιστοποίησης, όπως η ***Μέθοδος Προσομοιούμενης Ανόπτησης*** (*Simulated Annealing (SA)*), ο ***Γενετικός Αλγόριθμος*** (*Genetic Algorithm (GA)*) και η ***Μέθοδος Απαγορευμένης Αναζήτησης*** (*Tabu Search (TS)*).

Διαπιστώνεται ότι: α) η καλύτερη εκδοχή του *GA* υπερτερεί των καλύτερων εκδοχών του *TS*, με την διαφορά να μεγαλώνει καθώς αυξάνεται ο αριθμός των υπό μελέτη μεταβλητών *p*, β) ο *GA* και ο *TS* είναι σαφώς ανώτεροι του *SA* για αυτό το πρόβλημα για όλες τις τιμές του αριθμού *p* που μελετήθηκαν, και γ) τα 'κατάλληλα' υποσύνολα των δεικτών που συμβιβάζουν τις έννοιες κόστος και ακρίβεια πρόβλεψης, δημιουργούν μεγάλη μείωση του κόστους στα προγράμματα ποιοτικού έλεγχου.

Η εργασία αυτή α) δημιουργεί νέες ιδέες στο κομμάτι της επιλογής μεταβλητών στα γενικευμένα γραμμικά μοντέλα, β) συνθέτει νέες αντιλήψεις για τα πλεονεκτήματα και τα μειονεκτήματα των συναγωνιζόμενων μεθόδων μεγιστοποίησης και γ) παράγει αποτελέσματα που μπορούν να χρησιμοποιηθούν άμεσα στον τομέα της πολιτικής της υγείας.

Δ13. **Fouskakis, D.**, Ntzoufras, I. and Draper, D. (2009). Bayesian variable selection using a cost-adjusted BIC, with application to cost-effective measurement of quality of health care. *Annals of Applied Statistics*, **3**, 663-690.

Στην παρούσα ερευνητική εργασία, συνεχίζουμε τη μελέτη του προβλήματος της εργασίας Δ12, χρησιμοποιώντας αυτή την φορά τον εκ-των-υστέρων λόγο σχετικών πιθανοτήτων των υποδειγμάτων για την αξιολόγηση των υπό εξέταση μεταβλητών και μοντέλων. Προτείνουμε εκ-των-προτέρων κατανομές και πιθανότητες που λαμβάνουν υπόψη τους το κόστος της κάθε επεξηγηματικής μεταβλητής και καταλήγουν σε εκ-των-υστέρων πιθανότητες υποδειγμάτων που αντιστοιχούν σε μια βασισμένη στο κόστος γενίκευση του Κριτηρίου Πληροφορίας κατά *Bayes* (*BIC*). Χρησιμοποιούμε τον αλγόριθμο προσομοίωσης Αντιστρέψιμου Άλματος *Monte Carlo* με χρήση Μαρκοβιανών αλυσίδων (*Reversible Jump MCMC*) για τη διερεύνηση του χώρου των υποδειγμάτων και συγκρίνουμε τα αποτελέσματά μας με αυτά που προκύπτουν από δυο παραλλαγές του αλγορίθμου προσομοίωσης Σύνθεσης Υποδειγμάτων *Monte Carlo* με χρήση Μαρκοβιανών αλυσίδων (*MCMC Model Composition, MC*^{3}). Αρχικά μειώνουμε τη διάσταση του χώρου των υπό μελέτη υποδειγμάτων αφαιρώντας μεταβλητές με χαμηλές περιθώριες εκ-των-υστέρων πιθανότητες και εν συνεχεία εκτιμούμε τις εκ-των υστέρων πιθανότητες των υποδειγμάτων στο χώρο με τη μικρότερη διάσταση.

Η προτεινόμενη μεθοδολογία έχει ως αποτέλεσμα την επιλογή υποδειγμάτων στα οποία παρατηρείται αξιοσημείωτη μείωση στο κόστος και στη διάσταση και μόνο μικρή μείωση στην προβλεπτική τους ικανότητα, σε σύγκριση με τα μοντέλα που προκύπτουν από την ανάλυση που δε λαμβάνει υπόψη της το κόστος. Τα αποτελέσματα μας μπορούν εύκολα να εφαρμοστούν και σε προβλήματα πέραν της αξιολόγησης νοσοκομειακών μονάδων, όπως για παράδειγμα, η εξέταση του ποσοστού διαρροής στην εκπαίδευση, της μελέτης του δείκτη διατήρησης πελατών (*retention rate*) μιας εταιρείας και στην κατασκευή δεικτών (βασισμένων στο κόστος) της πιστοληπτικής ικανότητας των πελατών μιας τράπεζας.

Δ14. Kokolakis, G. and **Fouskakis, D.** (2009). Importance Partitioning in Micro-Aggregation. *Computational Statistics and Data Analysis*, **53**, 2439-2445.

Στην εργασία αυτή αναπτύσσεται μια νέα τεχνική βέλτιστης μικρο-ομαδοποίησης (*optimal micro-aggregation*) στατιστικών δεδομένων. Το κριτήριο βελτιστοποίησης που τίθεται εδώ είναι αυτό της ελαχιστοποίησης της Απώλειας Πληροφορίας $L = SSW/SST$, όπου $SST = SSB + SSW$, δηλαδή αυτό της μεγιστοποίησης του "Μεταξύ Ομάδων Αθροίσματος Τετραγώνων" (*SSB: Between Group Sum of Squares*), ή ισοδύναμα αυτό της ελαχιστοποίησης του "Εντός Ομάδων Αθροίσματος Τετραγώνων" (*SSW: Within Group Sum of Squares*). Η ως άνω ισοδυναμία μας επιτρέπει να αναπτύξουμε μια πρωτότυπη τεχνική δύο βημάτων κατά την οποία, στο πρώτο βήμα μεγιστοποιείται το *SSB* και στο δεύτερο ελαχιστοποιείται το *SSW*. Αποδεικνύεται ότι το μεν κριτήριο μεγιστοποίησης του *SSB* μας οδηγεί κάθε φορά στην πλέον "ακραία" ομάδα με βάση το "μέτρο σημαντικότητας" (*importance measure*) της ομάδας αυτής στην διαμόρφωση του *SST*. Το κριτήριο ελαχιστοποίησης του *SSW* μας οδηγεί κάθε φορά στην πλέον "συμπαγή" ομάδα με βάση το "μέτρο ομοιότητας" (*similarity measure*) των στοιχείων αυτής. Η εναλλαγή των ως άνω δύο κριτηρίων σε διαδοχικά βήματα μας επιτρέπει να κατασκευάσουμε εξαιρετικά "συμπαγείς" ομάδες, δηλαδή διαμερίσεις με ελάχιστη Απώλεια Πληροφορίας. Διαπιστώνεται ότι η κατασκευή αυτή επιτυγχάνεται σε πολυωνυμικό χρόνο παρότι πρόκειται για υπολογιστικό πρόβλημα εξαιρετικά υψηλής πολυπλοκότητας (*NP-hard*).

Δ15. **Fouskakis, D.**, Ntzoufras, I. and Draper, D. (2009). Population-based reversible-jump Markov chain Monte Carlo for Bayesian variable selection and evaluation under cost limit restrictions. *Journal of the Royal Statistical Society, Series C: Applied Statistics*, **58**, 663-690.

Η ποιότητα των υπηρεσιών υγείας αποτελεί σημαντική περιοχή έρευνας και ανάπτυξης για πολλές σύγχρονες κοινωνίες. Ένας έμμεσος τρόπος για την αξιολόγηση νοσοκομείων είναι η σύγκριση των παρατηρούμενων συντελεστών θνησιμότητας με τους αντίστοιχους αναμενόμενους συντελεστές τους για έναν αριθμό νοσοκομείων, με δεδομένη τη νόσο κατά την εισαγωγή των ασθενών. Η νοσηρότητα των ασθενών κατά την εισαγωγή τους υπολογίζεται τυπικά με την χρήση της λογιστικής παλινδρόμησης με μεταβλητή απόκρισης τη θνησιμότητα του ασθενούς, μέσα σε ένα διάστημα 30 ημερών από την ημερομηνία εισαγωγής. Για να επιτευχθεί η κατασκευή ενός αποτελεσματικού μοντέλου μπορούν να χρησιμοποιηθούν οι κλασικές μέθοδοι επιλογής μεταβλητών έτσι ώστε να βρεθεί το 'βέλτιστο' υποσύνολο 10–20 δεικτών ή μεταβλητών που θα χρησιμοποιηθούν ως επεξηγηματικές μεταβλητές.

Όταν ο στόχος είναι η δημιουργία μίας κλίμακας ασθένειας, η οποία μελλοντικά θα μπορεί να χρησιμοποιηθεί για τη μέτρηση της ποιότητας των παροχών σε ένα νέο σύνολο ασθενών, οι παραδοσιακές μέθοδοι επιλογής μεταβλητών μέσω της συνάρτησης ωφέλειας (ή χρησιμότητας) επιλέγουν ένα υποσύνολο μεταβλητών το οποίο δεν είναι βέλτιστο εφόσον δεν λαμβάνονται υπ' όψη τα διαφορετικά κόστη συλλογής πληροφοριών για κάθε διαθέσιμη επεξηγηματική μεταβλητή.

Σε αυτή την ερευνητική εργασία, προτείνεται ένας αλγόριθμος που μπορεί να εφαρμοστεί υπό την ύπαρξη ενός περιορισμού κόστους. Εξετάζεται η πρακτική σημαντικότητα των μεταβλητών υπό τον περιορισμό ενός ολικού κόστους. Η διερεύνηση γίνεται μεταξύ των μοντέλων με κόστη που δεν ξεπερνούν τα όρια ενός προϋπολογισμού. Η εφαρμογή των παραδοσιακών αλγορίθμων διερεύνησης υποδειγμάτων συνήθως αποτυχαίνουν στις περιπτώσεις όπου το καλύτερο μοντέλο βρίσκεται εκτός των ορίων του

κόστους που έχει τεθεί ή αν υπάρχουν προβλήματα πολυσυγγραμικότητας μεταξύ επεξηγηματικών μεταβλητών με υψηλό κόστος συλλογής. Ο λόγος αποτυχίας των μεθόδων αυτών είναι η ύπαρξη πολλών τοπικών μεγίστων στο χώρο των εκ–των–υστέρων πιθανοτήτων των μοντέλων σε περιοχές που δεν επικοινωνούν άμεσα κατά τη διάρκεια της διερεύνησης του χώρου των μοντέλων από τον αλγόριθμο λόγω των περιορισμών κόστους που έχουν τεθεί. Για το λόγο αυτό προτείνεται μια βελτιωμένη έκδοση του αλγορίθμου αναστρέψιμου άλματος, βασισμένη στους αλγορίθμους πολλαπλών παράλληλων πληθυσμών – αλυσίδων. Ο προτεινόμενος αλγόριθμος εκτιμάει αποτελεσματικά τις εκ–των–υστέρων πιθανότητες των μοντέλων και κινείται επιτυχημένα μεταξύ περιοχών που πριν δεν επικοινωνούσαν λόγω των περιορισμών κόστους.

Δ16. Daskalakis, G., Simou, M., Zacharakis, D., Detorakis, S., Akrivos, N., Papantoniou, N., **Fouskakis D.** and Antsaklis, A. (2011). Impact of placenta previa on obstetric outcome. *International Journal Gynaecology and Obstetrics*, **114**, 238-241.

Σκοπός της παρούσας εργασίας είναι να καταγραφεί η μητρική και νεογνική έκβαση για τους διαφορετικούς τύπους προδρομικού πλακούντα (ΠΠ), χρησιμοποιώντας μία αναδρομική μελέτη 132 μονήρων κυήσεων με προδρομικό πλακούντα. Η επίπτωση του προδρομικού πλακούντα για το χρονικό διάστημα που μελετήθηκε ήταν 0.96% (132 στις 13705 γεννήσεις). Από τις γυναίκες με προδρομικό πλακούντα το 51.5% είχε επιπωματικό ΠΠ, το 20.5% είχε επιχείλιο ΠΠ, το 5.3% είχε παραχείλιο ΠΠ και το 22.7% είχε πλακούντα χαμηλής πρόσφυσης. Οι περισσότερες γυναίκες (93.9%) γέννησαν με καισαρική τομή. Συνολικά στο 19.7% των περιπτώσεων ακολούθησε μαιευτική ολική υστερεκτομή, ενώ από αυτές τις γυναίκες το 92.3% είχε επιπωματικό ΠΠ. Γυναίκες με 2 ή περισσότερες καισαρικές τομές στο ιστορικό τους παρουσίασαν αυξημένο κίνδυνο για μαιευτική υστερεκτομή (*p-τιμή < 0.01*). Η ηλικία κύησης κατά τη γέννηση αποτέλεσε επίσης σημαντικό παράγοντα του *Apgar Score* στο 5^0 λεπτό. Τέλος γυναίκες με επιχείλιο προδρομικό πλακούντα γέννησαν νεογνά με μικρότερα *Apgar Scores* συγκριτικά με περιπτώσεις επιπωματικού ΠΠ (*p-τιμή = 0.02*).

Με βάση την εν λόγω στατιστική ανάλυση καταλήγουμε στο συμπέρασμα ότι το ιστορικό πολλαπλών προηγηθείσων καισαρικών τομών αυξάνει τον κίνδυνο για μαιευτική ολική υστερεκτομή. Ο τύπος του προδρομικού πλακούντα δεν σχετίζεται με διαφορές στο νεογνικό και μητρικό περιγεννητικό αποτέλεσμα, εκτός από το ότι, όπως φάνηκε στην εν λόγω μελέτη, νεογνά στην ομάδα του επιχείλιου πλακούντα είχαν μικρότερα *Apgar Scores* σε σχέση με τα νεογνά στην ομάδα του επιπωματικού προδρομικού πλακούντα.

Δ17. **Fouskakis, D.** (2012). Bayesian variable selection in generalized linear models using a combination of stochastic optimization methods. *European Journal of Operational Research*, **220**, 414-422.

Στην παρούσα εργασία η χρήση στοχαστικών αλγόριθμων βελτιστοποίησης για την αποτελεσματική διερεύνηση του χώρου των υποδειγμάτων προτείνεται, στο πρόβλημα επιλογής μεταβλητών κατά *Bayes* στα γενικευμένα γραμμικά μοντέλα. Συνδυάζοντας πτυχές τριών ευρέως γνωστών στοχαστικών αλγόριθμων βελτιστοποίησης (***Μέθοδος Προσομοιούμενης Ανόπτησης*** "Simulated Annealing (SA)", ***Γενετικός Αλγόριθμος*** "Genetic Algorithm (GA)" και ***Μέθοδος Απαγορευμένης Αναζήτησης*** "Tabu Search (TS)") καταλήγουμε σε έναν αποτελεσματικό αλγόριθμο διερεύνησης του χώρου των υποδειγμάτων. Χρησιμοποιώντας κατάλληλες εκ-των-προτέρων κατανομές, οι εκ-των-υστέρων πιθανότητες των υπό μελέτη υποδειγμάτων χρησιμοποιούνται για την αξιολόγηση - κατάταξή τους, ενώ σε περιπτώσεις που οι εκ-των-υστέρων πιθανότητες των υποδειγμάτων

δεν μπορούν να υπολογιστούν σε κλειστή μορφή εκτιμούνται με την βοήθεια της προσέγγισης κατά Laplace. Ο προτεινόμενος αλγόριθμος υλοποιείται σε παραδείγματα κανονικών γραμμικών και λογιστικών παλινδρομικών μοντέλων, με προσομοιωμένα και πραγματικά δεδομένα. Με αρκετά χαμηλό υπολογιστικό κόστος, ο προτεινόμενος αλγόριθμος υπερτερεί σαφώς γνωστών αλγόριθμων *MCMC*, όπως ο αλγορίθμους προσομοίωσης Σύνθεσης Υποδειγμάτων Monte Carlo με χρήση Μαρκοβιανών αλυσίδων (*MCMC Model Composition, MC*^{3}), ενώ επιπλέον υπερτερεί και των απλών εκδοχών των *SA*, *GA* και *TS*.

Δ18. Spanos, A., Theoharis, G., Karageorgopoulos, D.E., Peppas, G., **Fouskakis, D.** and Falagas, M.E. (2012). Surveillance of community outbreaks of respiratory tract infections based on house-call visits in the metropolitan area of Athens, Greece. *PLos ONE*, **7**, e40310.

Στην παρούσα εργασία, στα πλαίσια της επιδημιολογικής επιτήρησης γρίπης και παρεμφερών νοσημάτων, εφαρμόζονται στατιστικές μέθοδοι γραμμικής παλινδρόμησης και συσσωρευτικών αθροισμάτων οι οποίες χρησιμοποιούν μικρό αριθμό πρόσφατων ιστορικών δεδομένων για την αναγνώριση εξάρσεων λοιμώξεων του άνω και κάτω αναπνευστικού συστήματος. Τα δεδομένα για την εν λόγω μελέτη αφορούν στο ποσοστό των διαγνώσεων για λοιμώξεις ανωτέρου ή κατωτέρου αναπνευστικού ως προς το συνολικό αριθμό των διαγνώσεων για κατ'οίκον ιατρικές επισκέψεις που πραγματοποιήθηκαν από το ιδιωτικό δίκτυο των SOS Ιατρών στην ευρύτερη περιοχή της Αθήνας μεταξύ 1 Ιανουαρίου 2000 και 12 Οκτωβρίου 2008. Τα ευρήματα της ανάλυσης υποδεικνύουν ότι η παρούσα μεθοδολογία μπορεί να χρησιμοποιηθεί αξιόπιστα για το σκοπό της έγκαιρης αναγνώρισης εξάρσεων λοιμώξεων του αναπνευστικού. Η συγκεκριμένη μεθοδολογία μπορεί να αποδειχθεί ιδιαίτερα χρήσιμη σε περιπτώσεις που δεν είναι διαθέσιμη κάποια μεγάλη ή αξιόπιστη βάση δεδομένων για την εφαρμογή των συνηθισμένων στατιστικών μεθόδων.

Δ19. **Fouskakis, D.** and Ntzoufras, I. (2013). Computation for intrinsic variable selection in normal regression models via expected-posterior prior. *Statistics and Computing*, **23**, 491-499.

Στην παρούσα εργασία έμφαση δίνεται στο πρόβλημα επιλογής μεταβλητών κατά *Bayes* στα γενικά γραμμικά μοντέλα παλινδρόμησης, χρησιμοποιώντας μεταγενέστερες αναμενόμενες εκ των προτέρων (*expected-posterior prior*) κατανομές. Παρουσιάζεται ένας αποτελεσματικός αλγόριθμος *MCMC* για προσομοίωση από την εκ των υστέρων κατανομή, καθώς επίσης και διάφορες *Monte Carlo* εκτιμήτριες των περιθώριων πιθανοφανειών και των εκ των υστέρων πιθανοτήτων των υποδειγμάτων. Επιπλέον, για μεγάλους χώρους υποδειγμάτων, παρουσιάζεται ένας αλγόριθμος αναζήτησης υποδειγμάτων βασιζόμενος στους αλγορίθμους προσομοίωσης Σύνθεσης Υποδειγμάτων Monte Carlo με χρήση Μαρκοβιανών αλυσίδων (*MCMC Model Composition, MC*^{3}). Η προτεινόμενη μεθοδολογία υλοποιείται σε δύο παραδείγματα με πραγματικά δεδομένα, τα οποία έχουν χρησιμοποιηθεί κατά το παρελθόν στη σχετική βιβλιογραφία της αντικειμενικής επιλογής μεταβλητών. Και στα δύο παραδείγματα, η αβεβαιότητα που απορρέει από διαφορετικά διδακτικά (*training*) δείγματα λαμβάνεται υπόψη.

Δ20. **Fouskakis, D.**, Ntzoufras, I. and Draper, D. (2015). Power-Expected-Posterior Priors for Variable Selection in Gaussian Linear Models. *Bayesian Analysis,* **10**, 75-107.

Στο πλαίσιο των μεταγενέστερων αναμενόμενων εκ των προτέρων (*expected-posterior prior - EPP*) κατανομών, συνδυάζουμε στοιχεία των δυναμικών και των μοναδιαίας

πληροφορίας πρότερων κατανομών έτσι ώστε ταυτόχρονα (α) να δημιουργήσουμε μία ελάχιστα πληροφοριακή πρότερη κατανομή και (β) να μειώσουμε την επίδραση των διδακτικών δειγμάτων. Η προκύπτουσα δυναμικά-μεταγενέστερη-αναμενόμενη εκ των προτέρων κατανομή *(power-expected-posterior (PEP) prior)* είναι μη ευαίσθητη στο μέγεθος n* των διδακτικών δειγμάτων, λόγω του τρόπου κατασκευής της (πρότερη μοναδιαίας πληροφορίας). Οπότε θέτοντας το n* ίσο με το μέγεθος (n) του πλήρους αρχικού μας δείγματός απαλλασσόμαστε από τη διαδικασία επιλογής διδακτικού δείγματος. Η εν λόγω τεχνική προωθεί την σταθερότητα του προκύπτοντας παράγοντα Bayes, αφαιρεί την αυθαιρεσία που προκύπτει από την επιλογή ενός διδακτικού δείγματος και αυξάνει σημαντικά την υπολογιστική ταχύτητα, επιτρέποντας πολλά περισσότερα μοντέλα να συγκριθούνε. Στην παρούσα εργασία, έμφαση δίνεται στο κανονικό γραμμικό μοντέλο παλινδρόμησης και αναπτύσσουμε την μέθοδο μας κάτω από δύο διαφορετικές αρχικές πρότερες κατανομές: την πρότερη του Jeffreys, με βάση την οποία δημιουργείται η J-PEP ύστερη κατανομή, και η πρότερη κατανομή του Zellner (*Zellner's g-prior*), με βάση την οποία δημιουργείται η Z-PEP ύστερη κατανομή. Η πρώτη επιλογή αρχικής πρότερης είναι η συνηθέστερη επιλογή στην βιβλιογραφία της αντικειμενικής επιλογής μεταβλητών που σχετίζεται με την εν λόγω εργασία, ενώ η δεύτερη απλοποιεί και επιτυγχάνει τους υπολογισμούς λόγω της συζυγούς δομής της (επιπλέον η J-PEP ύστερη είναι ειδική περίπτωση της Z-PEP ύστερης). Αποδεικνύουμε ότι, κάτω από την πρώτη επιλογή αρχικής πρότερης, ο παραγόμενος παράγοντας Bayes ασυμπτωτικά έχει την ίδια συμπεριφορά με το παράγοντα Bayes που προκύπτει από το Schwartz's BIC κριτήριο, και αυτό εξασφαλίζει την συνέπεια της μεθοδολογίας μας. Συγκρίνουμε την αποδοτικότητα της μεθόδου μας με αυτή μεθόδων που χρησιμοποιούνται συνήθως στην περιοχή της αντικειμενικής επιλογής μεταβλητών, σε δύο παραδείγματα με προσομοιωμένα και πραγματικά δεδομένα, με την αποδοτικότητα. Η PEP πρότερη κατανομή, λόγω της δομής της (πρότερη κατανομή μοναδιαίας πληροφορίας) οδηγεί σε μια διαδικασία επιλογής μεταβλητών (1) η οποία είναι συστηματικά πιο φειδωλή, σε σχέση με την διαδικασία που προκύπτει έπειτα από χρήση της EPP με διδακτικά δείγματα ελάχιστου μεγέθους, χωρίς να θυσιάζει μέρος της απόδοσής της για την επίτευξη της εν λόγω φειδωλότητας, (2) η οποία είναι ανθεκτική στο μέγεθος του διδακτικού δείγματος, και συνεπώς απολαμβάνει τα πλεονεκτήματα που περιγράφηκαν παραπάνω για την αποφυγή αυθαίρετης δημιουργίας διδακτικών δειγμάτων και (3) η οποία προσδιορίζει το μοντέλο με την μέγιστη ύστερη πιθανότητα και με καλή προβλεπτική απόδοση. Επιπλέον η PEP πρότερη κατανομή είναι διάχυτη (μη πληροφοριακή) ακόμα και αν όταν το μέγεθος δείγματος n δεν είναι πολύ μεγαλύτερο του αριθμού επεξηγηματικών μεταβλητών p, περίπτωση κατά την οποία η EPP είναι πολύ πιο πληροφοριακή από ότι έχει σχεδιαστεί.

Δ21. **Fouskakis, D.,** Petrakos, G. and Vavouras, I. (2016). A Bayesian Hierarchical Model for Comparative Evaluation of Teaching Quality Indicators in Higher Education. *Journal of Applied Statistics*, **43**, 195-211.

Σκοπός του άρθρου είναι η αντιμετώπιση του προβλήματος ποσοτικοποίησης των προτιμήσεων των φοιτητών σχετικά με την ποιότητα της εκπαίδευσης και των ακαδημαϊκών μαθημάτων, κάτω από ορισμένους περιορισμούς που υπεισέρχονται ώστε να κατασκευαστεί ένα μοντέλο για τον προσδιορισμό, την αξιολόγηση και την παρακολούθηση των βασικών συνιστωσών της συνολικής ακαδημαϊκής ποιότητας. Αφού εξετάσουμε τα πλεονεκτήματα και τους περιορισμούς της ανάλυσης σύζευξης και του μοντέλου παλινδρόμησης τυχαίων συντελεστών τα οποία έχουν χρησιμοποιηθεί σε παρόμοια προβλήματα στο παρελθόν, προτείνουμε ένα Μπεϋζιανό μοντέλο Βήτα παλινδρόμησης με μια Dirichlet πρότερη κατανομή για τους συντελεστές του μοντέλου. Η εν λόγω προσέγγιση όχι μόνο επιτρέπει την

ενσωμάτωση πληροφοριακής εκ των προτέρων κατανομής (όταν αυτή είναι διαθέσιμη), αλλά επιπλέον παρέχει φιλικά προς το χρήστη αποτελέσματα και άμεση ερμηνεία των πιθανοτήτων όλων των εμπλεκόμενων ποσοτήτων. Ακόμη, αποτελεί έναν φυσικό τρόπο για να εφαρμόσει κανείς τους συνήθεις περιορισμούς για τα βάρη/συντελεστές του μοντέλου. Το συγκεκριμένο μοντέλο εφαρμόστηκε σε δεδομένα που συλλέχθηκαν το 2009 και το 2013 από προπτυχιακούς φοιτητές του Παντείου Πανεπιστημίου Αθηνών και, εκτός από την κατασκευή ενός εργαλείου αξιολόγησης και παρακολούθησης της ποιότητας διδασκαλίας, έδωσε υλικό για μια προκαταρκτική συζήτηση πάνω στη συσχέτιση των διαφορετικών προτιμήσεων των φοιτητών μεταξύ δύο χρονικών περιόδων με την τρέχουσα Ελληνική χρηματοοικονομική κρίση.

Δ22. Charitidou, E., **Fouskakis, D.** and Ntzoufras, I. (2015). Bayesian Transformation Selection: Moving Towards a Transformed Gaussian Universe. *Canadian Journal of Statistics*, forthcoming, arXiv: 1312.3482.

Το πρόβλημα της επιλογής μετασχηματισμού εξετάζεται πλήρως από την Μπεϋζιανή σκοπιά. Οι κάτωθι οικογένειες μετασχηματισμού λαμβάνονται υπόψη με σκοπό να προσεγγίσουμε την κανονικότητα όσον αφορά στην κατανομή ενός συνόλου δεδομένων: Box-Cox, Modulus, Yeo & Johnson και Dual. Αλγόριθμοι MCMC κατασκευάστηκαν ώστε να προσομοιώσουμε δείγμα από την εκ των υστέρων κατανομή της παραμέτρου μετασχηματισμού λ_T συνδεόμενη με την εκάστοτε οικογένεια μετασχηματισμού T. Διερευνούμε διαφορετικές προσεγγίσεις με στόχο την κατασκευή συμβατών πρότερων κατανομών για την παράμετρο λ_T μεταξύ των οικογενειών, κάνοντας χρήση μιας δυναμικής πρότερης κατανομής μοναδιαίας πληροφορίας αλλά και μιας κανονικής εκ των προτέρων κατανομής με μοναδιαία βαρύτητα. Η επιλογή και ανάδειξη της βέλτιστης οικογένειας βασίζεται στον υπολογισμό των εκ των υστέρων πιθανοτήτων των οικογενειών. Χρησιμοποιώντας προσομοιωμένα δεδομένα, αναδεικνύεται η αποτελεσματικότητα της μεθοδολογίας που περιγράφεται στην παρούσα εργασία. Παρά το γεγονός ότι δεν αναδείχθηκε μία οικογένεια η οποία να προσφέρει καθολική λύση, σύνολα δεδομένων με συγκεκριμένα χαρακτηριστικά φαίνεται να ωφελούνται από συγκεκριμένες οικογένειες. Για παράδειγμα, ασύμμετρες κατανομές σχετίζονται με την οικογένεια Box-Cox ενώ κατανομές με παχιές ουρές ωφελούνται περισσότερο από το μετασχηματισμό Modulus.

Δ23. **Fouskakis, D.** and Ntzoufras, I. (2016). Limiting behavior of the Jeffreys power-expected-posterior Bayes factor in Gaussian linear models. *Brazilian Journal of Probability and Statistics,* **30**, 299-320.

Οι μεταγενέστερες αναμενόμενες εκ των προτέρων (*expected-posterior prior - EPP*) κατανομές έχει αποδειχθεί ότι είναι ιδιαίτερα χρήσιμες για ελέγχους υποθέσεων που αφορούν τους συντελεστές κανονικών γραμμικών μοντέλων. Ένα από τα προτερήματα των EPP είναι ότι ακαταλληλότητα αρχικών πρότερων κατανομών δεν δημιουργεί απροσδιοριστία. Ωστόσο βασίζονται σε ένα ή και περισσότερα διδακτικά δείγματα, τα οποία μπορεί να επηρεάσουν την προκύπτουσα ύστερη κατανομή. Οι δυναμικά-μεταγενέστερες-αναμενόμενες εκ των προτέρων κατανομές *(power-expected-posterior (PEP) prior)* είναι ελάχιστα πληροφοριακές πρότερες οι οποίες ελαττώνουν την επίδραση των διδακτικών δειγμάτων στις EPP, συνδυάζοντας πτυχές από τις δυναμικές και τις μοναδιαίας πληροφορίας πρότερες κατανομές. Στην παρούσα εργασία αποδεικνύουμε σε ένα κανονικό γραμμικό μοντέλο, κάτω από ασθενείς συνθήκες που αφορούν τον πίνακα σχεδιασμού, την συνέπεια του παράγοντα Bayes όταν χρησιμοποιούμε τις PEP με την πρότερη κατανομή του Jeffreys ως αρχική εκ των προτέρων κατανομή.

Δ24. **Fouskakis, D.** and Ntzoufras, I. (2016). Power-conditional-expected priors. Using g-priors with random imaginary data for variable selection. *Journal of Computational and Graphical Statistics,* **25**, 647-664.

Η πρότερη κατανομή του Zellner (*Zellner's g-prior*) καθώς και οι πρόσφατες ιεραρχικές της επεκτάσεις (*hyper-g prior)* είναι συνηθισμένες επιλογές πρότερων κατανομών σε Μπεϋζιανά προβλήματα επιλογής μεταβλητών. Οι εν λόγω πρότερες κατανομές μπορούν να εκφραστούν ως δυναμικές πρότερες κατανομές με προκαθορισμένη συλλογή φανταστικών δεδομένων. Στην παρούσα εργασία, χρησιμοποιούμε ιδέες από τις δυναμικές-μεταγενέστερες-αναμενόμενες εκ των προτέρων κατανομές *(power-expected-posterior (PEP) priors)* ώστε να εισάγουμε στην πρότερη κατανομή του Zellner ένα επιπλέον ιεραρχικό επίπεδο στο οποίο η αβεβαιότητα που απορρέει από τα φανταστικά δεδομένα εξηγείται. Σε ένα κανονικό γραμμικό μοντέλο παλινδρόμησης, η προκύπτουσα *δυναμική-υπό συνθήκη-μεταγενέστερη-αναμενόμενη εκ-των-προτέρων κατανομή (power-conditional-expected-posterior (PCEP) prior)* είναι μία συζυγής *normal-inverse gamma* πρότερη κατανομή, η οποία δημιουργεί μία συνεπή διαδικασία επιλογής μεταβλητών και δίνει μεγαλύτερο βάρος σε πιο φειδωλά μοντέλα από ότι η *g-prior* και η *hyper-g prior* σε δείγματα μικρού μεγέθους. Η προτεινόμενη μεθοδολογία υλοποιείται σε δύο παραδείγματα με προσομοιωμένα και πραγματικά δεδομένα και συγκριτικά αποτελέσματα με την *g-prior* και την *hyper-g prior* παρουσιάζονται.

Δ25. **Fouskakis, D.** and Ntzoufras I. (2017). Information consistency of the Jeffreys power-expected-posterior prior in Gaussian linear models. *Metron,* **75***,* 371-380.

Οι δυναμικά-μεταγενέστερες-αναμενόμενες εκ των προτέρων κατανομές *(power-expected-posterior (PEP) priors)* έχουν προσφάτως αναπτυχθεί ως γενικεύσεις των μεταγενέστερων αναμενόμενων εκ των προτέρων (*expected-posterior prior - EPP*) κατανομών για το πρόβλημα επιλογής επεξηγηματικών μεταβλητών σε κανονικά μοντέλα γραμμικής παλινδρόμησης. Είναι ελάχιστα πληροφοριακές πρότερες κατανομές οι οποίες μειώνουν την επίδραση των διδακτικών δεδομένων συνδυάζοντας στοιχεία των δυναμικών και των μοναδιαίας πληροφορίας πρότερων κατανομών. Στην παρούσα εργασία αποδεικνύουμε την συνέπεια πληροφορίας της PEP μεθόδου, όταν χρησιμοποιούμε την Jeffreys ως αρχική εκ των προτέρων κατανομή σε ένα κανονικό γραμμικό μοντέλο.

Δ26. **Fouskakis, D.**, Ntzoufras I. and Perrakis K. (2018). Power-expected-posterior priors in generalized linear models. *Bayesian Analysis,* **13**, 721-748.

Η δυναμικά-μεταγενέστερη-αναμενόμενη εκ των προτέρων κατανομή *(power-expected-posterior (PEP) prior)* που αναπτύχθηκε για το πρόβλημα επιλογής επεξηγηματικών μεταβλητών σε κανονικά μοντέλα γραμμικής παλινδρόμησης αποτελεί μία αντικειμενική, αυτόματη, συνεπή και φειδωλή μέθοδο επιλογής μοντέλων. Συγχρόνως επιλύει προβλήματα θεωρητικά αλλά και υπολογιστικής φύσης, χρησιμοποιώντας διδακτικά δεδομένα. Συγκεκριμένα με τη χρήση της παρούσας μεθόδου, (α) απαλλασσόμαστε από τη διαδικασία επιλογής του διδακτικού δείγματος καθώς επίσης και τον υπολογισμό μέσον όρων ως προς όλα τα πιθανά διδακτικά δείγματα, (β) ελαττώνεται η επίδραση των διδακτικών δειγμάτων στην τελική ύστερη κατανομή. Λόγω των παραπάνω, για μεγάλου μεγέθους δείγματα, μπορούμε να χρησιμοποιήσουμε, όταν χρειαστεί, ασυμπτωτικές προσεγγίσεις οι οποίες σε περίπλοκα μοντέλα δύναται να μειώσουν δραστικά το υπολογιστικό φορτίο. Στην παρούσα εργασία γενικεύουμε τη χρήση

των δυναμικά-μεταγενέστερων-αναμενόμενων εκ των προτέρων κατανομών σε γενικευμένα γραμμικά μοντέλα, παρουσιάζοντας δύο νέους ορισμούς της PEP πρότερης κατανομής, τους οποίους μπορούμε να εφαρμόσουμε σε οποιαδήποτε μοντέλα. Ιεραρχικές επεκτάσεις της παραμέτρου δύναμης, η οποία ρυθμίζει της επίδραση των διδακτικών δεδομένων, επίσης μελετώνται. Οι τελικές PEP πρότερες κατανομές τότε μπορούν προσεγγιστικά να γράφουν ως μία διπλή μίξη g-prior κατανομών. Εμπειρικά αποτελέσματα υποδεικνύουν ότι οι προτεινόμενες μέθοδοί είναι αποτελεσματικές όταν ο σκοπός είναι η φειδωλή συμπερασματολογία.

Δ27. Charitidou, E., **Fouskakis, D.**, and Ntzoufras, I. (2018). Objective Bayesian transformation and Variable Selection using Default Bayes Factors. *Statistics and Computing*, **28***,* 579-594.

Στο παρόν άρθρο, το πρόβλημα της ταυτόχρονης επιλογής μετασχηματισμού και επεξηγηματικών μεταβλητών αντιμετωπίζεται διεξοδικά μέσω αντικειμενικών Μπεϋζιανών προσεγγίσεων χρησιμοποιώντας εναλλακτικές μορφές του παράγοντα Μπέυζ. Τέσσερις μονοπαραμετρικές οικογένειες μετασχηματισμών (Box-Cox, Modulus, Yeo-Johnson και Dual), συνοδευόμενες από τον δείκτη T, αξιολογούνται και συγκρίνονται. Η εκμαίευση υποκειμενικής πρότερης πληροφορίας για την παράμετρο μετασχηματισμού λ_T, για κάθε οικογένεια T, είναι μια μη τετριμμένη διαδικασία. Επιπρόσθετα, δεν μπορεί να εύκολα υπάρξει πρότερη πληροφορία για την παράμετρο λ_T και συνεπώς μια αντικειμενική μέθοδος είναι αναγκαία. Ο ενδογενής παράγοντας Μπέυζ (IBF) και ο κλασματικός παράγοντας Μπέυζ (FBF) επιτρέπουν την ενσωμάτωση μη γνήσιων πρότερων κατανομών για την παράμετρο λ_T. Η συμπεριφορά κάθε προσέγγισης μελετάται χρησιμοποιώντας ένα παράδειγμα αναφοράς με προσομοιωμένα δεδομένα καθώς και δύο παραδείγματα με δεδομένα πραγματικών προβλημάτων.

Δ28. Consonni, G., **Fouskakis, D.**, Liseo, B. and Ntzoufras, I. (2018). Prior Distributions for Objective Bayesian Analysis. *Bayesian Analysis,* **13**, 627-679.

Στην παρούσα εργασία κάνουμε μία ανασκόπηση των πρότερων κατανομών της αντικειμενικής Μπεϋζιανής συμπερασματολογίας, εστιαζόμενοι κυρίως σε δημοσιευμένο ερευνητικό έργο το οποίο δεν έχει επαρκώς καλυφθεί σε προγενέστερες ανασκοπικές δημοσιεύσεις. Διαφοροποιούμε τις πρότερες κατανομές ανάλογα αν χρησιμοποιούνται για εκτίμηση (ή πρόβλεψη) ή για επιλογή μοντέλων και εστιάζουμε κυρίως στην δεύτερη περίπτωση. Εξετάζουμε διάφορα θεμελιώδη ζητήματα, παρουσιάζουμε πρόσφατες συνεισφορές σε διακριτό παραμετρικό χώρο καθώς και σε μοντέλα μεγάλων διαστάσεων, απαραίτητα κριτήρια κατασκευής πρότερων για επιλογή μοντέλων, πρότερες για επιλογή επεξηγηματικών μεταβλητών σε προβλήματα παλινδρόμησης και πρότερες για τον χώρο των υπό σύγκριση μοντέλων. Το άρθρο στο τέλος καταλήγει με μια σύντομη συνοπτική συζήτηση για σημεία αξίας περαιτέρω έρευνας.

Δ29. **Fouskakis, D.** (2019). Priors via Imaginary Training Samples of Sufficient Statistics for Objective Bayesian Hypothesis Testing. *Metron*, **77**, 179-199.

Στην παρούσα εργασία εξετάζουμε εκ νέου την χρήση των μεταγενέστερων αναμενόμενων εκ των προτέρων (*expected-posterior prior - EPP*) κατανομών στην περιοχή της "αντικειμενικής" Μπεϋζιανής σύγκρισης μοντέλων και ελέγχου υποθέσεων. Επιπλέον χρησιμοποιούμε την προσφάτως δημιουργηθείσα μέθοδο των δυναμικά-μεταγενέστερων-αναμενόμενων εκ των προτέρων κατανομών *(power-expected-posterior (PEP) prior)*, κατά

την οποία συνδυάζονται στοιχεία των δυναμικών και των μοναδιαίας πληροφορίας πρότερων κατανομών έτσι να δημιουργηθεί μία ελάχιστα πληροφοριακή πρότερη κατανομή που είναι ανθεκτική στο μέγεθος των διδακτικών δειγμάτων. Διερευνούμε την χρήση επαρκών στατιστικών ώστε να ορίσουμε εκ νέου τις EPP και PEP prior. Με τον τρόπο αυτόν μπορούμε να μειώσουμε την διάσταση του προβλήματος, οπότε και την υπολογιστική πολυπλοκότητα, μιας και σε μη επιλύσιμα προβλήματα προσομοιώνουμε δείγματα επαρκών στατιστικών χρησιμοποιώντας την πραγματική δειγματοληπτική τους κατανομή αντί να γεννάμε πλήρη σετς διδακτικών δειγμάτων. Παρουσιάζουμε συγκρίσεις μεταξύ διαφορετικών προσεγγίσεων και ολοκληρώνουμε με μία συζήτηση σχετικά με την ερμηνεία και τα πλεονεκτήματα της χρήσης επαρκών στατιστικών, στον σχηματισμό της πρότερης κατανομής, βασιζομένων σε διδακτικά δείγματα. Η εν λόγω ερευνητική εργασία επιπλέον συνεισφέρει στη θεωρητική κατανόηση των μεθοδολογιών της "αντικειμενικής" Μπεϋζιανής σύγκρισης μοντέλων, οι οποίες τα τελευταία χρόνια γίνονται όλο και πιο πρακτικές.

Δ30. **Fouskakis, D.**, Ntzoufras I. and Perrakis K. (2019). Variations of power-expected-posterior priors in normal regressions models. *Computational Statistics and Data Analysis*, **143**, 1-26.

Η δυναμικά-μεταγενέστερη-αναμενόμενη εκ των προτέρων κατανομή *(power-expected-posterior (PEP) prior)* είναι μία ``αντικειμενική" πρότερη κατανομή στα κανονικά γραμμικά μοντέλα που οδηγεί σε μία συνεπή διαδικασία επιλογής μοντέλων δίνοντας μεγαλύτερο βάρος σε πιο φειδωλά μοντέλα. Προσφάτως, δύο νέες μορφές της PEP prior έχουν προταθεί, οι οποίες γενικεύουν την δυνατότητα εφαρμογής της εν λόγω πρότερης κατανομής σε μοντέλα ευρύτερων κλάσεων. Στην παρούσα εργασία εξετάζουμε τις ιδιότητες των δύο νέων αυτών παραλλαγών, στα κανονικά γραμμικά μοντέλα, δίνοντας μεγαλύτερη έμφαση στις πρότερες διασπορές και στην συνέπεια των διαδικασιών επιλογής μοντέλου. Τα αποτελέσματα δείχνουν ότι και οι δύο αυτές προτεινόμενες μορφές της PEP prior έχουν μεγαλύτερες διασπορές από την πρότερη κατανομή μοναδιαίας πληροφορίας του Zellner και καταλήγουν σε συνεπείς διαδικασίες μιας και η ασυμπτωτική συμπεριφορά των αντιστοίχων περιθώριων πιθανοφανειών ταιριάζει με αυτή της διαδικασίας BIC.

Δ31. **Fouskakis, D.**, Innocent, J.K. and Pericchi, L. (2020). Bayes factors consistency for nested linear models based on the Jeffreys power-expected-posterior prior with increasing dimensions. *Statistical Theory and Related Fields,* **4**, 162-171.

Στην παρούσα εργασία εξετάζουμε, σε ένα κανονικό γραμμικό μοντέλο, την συνέπεια του παράγοντα Bayes όταν χρησιμοποιούμε τις δυναμικά-μεταγενέστερες-αναμενόμενες εκ των προτέρων κατανομές *(power-expected-posterior (PEP) prior)*, για διάφορες τιμές της δυναμικής παραμέτρου και όταν η διάσταση του πλήρες μοντέλου, με όλες τις επεξηγηματικές μεταβλητές παραμένει σταθερή ή είναι τάξης O(n), όπου n δηλώνει το μέγεθος του δείγματος.

Δ32. Petrakis, N., Peluso, S., **Fouskakis, D.** and Consonni, G. (2020). Objective Methods for Graphical Structural Learning. *Statistica Neerlandica,* **74**, 420-438.

Στην εργασία αυτή εφαρμόζουμε δύο τεκμηριωμένες αντικειμενικές Μπεϋζιανές μεθοδολογίες στο πρόβλημα επιλογής γραφικών μοντέλων. Οι μέθοδοι αυτές βασίζονται στις μεταγενέστερες-αναμενόμενες και στις δυναμικά-μεταγενέστερες-αναμενόμενες εκ

των προτέρων κατανομές, οι οποίες δεν χρησιμοποιούν δύο φορές τα δεδομένα, όπως συμβαίνει με άλλες υπάρχουσες μεθόδους. Ξεκινώντας από μία ακατάλληλη μη-πληροφοριακή πρότερη κατανομή, εκτιμούμε παράγοντες Μπέϋες και λόγους εκ των υστέρων πιθανοτήτων. Σε αρκετά σενάρια προσομοιώσεων, χρησιμοποιώντας διαφορετικούς συνδυασμούς μεγέθους δείγματος και κόμβων, παρατηρούμε ότι οι προτεινόμενες μεθοδολογίες παρουσιάζουν όμοιες ή και καλύτερες επιδόσεις συγκριτικά με ήδη υπάρχουσες μεθόδους. Κλείνοντας, παρουσιάζουμε μία εφαρμογή σε πραγματικά δεδομένα πρωτεϊνών, όπου τα προβλεπόμενα αποτελέσματα συμβαδίζουν με την ήδη υπάρχουσα βιβλιογραφία.

Δ33. **Fouskakis, D.** and Ntzoufras, I. (2020). Bayesian Model Averaging using Power-Expected-Posterior Priors. *Econometrics.* **8**, 17.

Στην παρούσα εργασία εστιάζουμε στην Μπεϋζιανή Στάθμιση Μοντέλων (Bayesian Model Averaging – BMA) χρησιμοποιώντας την δυναμικά-μεταγενέστερη-αναμενόμενη εκ των προτέρων κατανομή *(power-expected-posterior (PEP) prior),* στην περιοχή της αντικειμενικής Μπεϋζιανής μεθοδολογίας, στα πολλαπλά γραμμικά μοντέλα. Δίνεται ο BMA σημειακός εκτιμητής της προβλεπόμενης τιμής και παρουσιάζονται υπολογιστικές λύσεις καθώς και στρατηγικές αξιολόγησης της προβλεπτικής ακρίβειας. Συγκρίνουμε την αποτελεσματικότητα της μεθόδου μας με αυτήν άλλων παρεμφερών μεθόδων σε προσομοιωμένα δεδομένα καθώς και σε πραγματικά δεδομένα από τον οικονομικό κλάδο.

Δ34. **Fouskakis, D.**, Petrakos, G. and Rotous, I. (2020). A Bayesian Longitudinal Model for Quantifying Student's Preferences Regarding Teaching Quality Indicators. *Metron,* **78**, 255-270.

Σκοπός της εργασίας είναι να εκτιμήσει την εκ των υστέρων μέση τιμή και να αναλύσει την εκ των υστέρων μεταβλητότητα των προτιμήσεων των σπουδαστών σχετικά με την ποιότητα των ακαδημαϊκών μαθημάτων σε έναν χρονικό ορίζοντα δέκα ετών. Τα αποτελέσματα βασίζονται σε μια διαχρονική δειγματοληπτική έρευνα, στην οποία το δείγμα απαρτιζόταν από φοιτητές Ελληνικού Πανεπιστημίου κατά την διάρκεια της οικονομικής κρίσης της Ελλάδας από το 2009 μέχρι το 2018. Για την ανάλυση των δεδομένων, προσαρμόστηκε ένα Μπεϋζιανό ιεραρχικό μοντέλο Βήτα παλινδρόμησης με Dirichlet πρότερη κατανομή για τους συντελεστές του μοντέλου, οι οποίοι αντιστοιχούν σε είκοσι ποιοτικούς δείκτες. Με το εν λόγω μοντέλο μπορούμε να υλοποιήσουμε τους συνήθης περιορισμούς, οι συντελεστές του μοντέλου ερμηνεύονται ως βάρη και ως εκ τούτου μετρούν την σχετική προτίμηση που οι φοιτητές δίνουν στα είκοσι διαφορετικά χαρακτηριστικά. Εκτιμώντας εκ των υστέρων μέσους και άλλα μέτρα θέσης και μεταβλητότητας, για όλες τις χρονικές περιόδους, στα βάρη του μοντέλου, ερχόμαστε σε συμπεράσματα σχετικά με τις αλλαγές και τα διαφορετικά πρότυπα που ακολουθούν οι αντιλήψεις των φοιτητών σχετικά με την ποιότητα των ακαδημαϊκών μαθημάτων στην διάρκεια των δέκα ετών.

Δ35. Dedos, S. G. and **Fouskakis, D.** (2021). Dataset and Validation of the Approaches to Study Skills Inventory for Students. *Scientific Data,* **8**, 158.

Μέσω της εκτενούς έρευνας σχετικά με τους τρόπους με τους οποίους οι φοιτητές/τριες προσεγγίζουν τη μελέτη και τη μάθηση, έχουν προκύψει πολλά εργαλεία μέτρησης τέτοιων δεξιοτήτων και προσεγγίσεων. Ένα από αυτά ονομάζεται Προσεγγίσεις

και Κατάλογος Μαθησιακών Δεξιοτήτων για Φοιτητές (Approaches to Study Skills Inventory for Students (ASSIST)). Το εργαλείο αυτό διακρίνει τους φοιτητές/τριες ως έχοντες ενδελεχή, στρατηγική και/ή επιφανειακή προσέγγιση ως προς τη μάθηση και μελέτη τους. Στο άρθρο αυτό παρουσιάζουμε τα αποτελέσματα μια 8ετούς έρευνας για την πιστοποίηση του παραπάνω εργαλείου σε ένα δείγμα 1181 φοιτητών/τριών σε ένα τμήμα ενός ελληνικού πανεπιστημίου. Η ανάλυση αξιοπιστίας των αποτελεσμάτων έδειξε ότι η ανωτέρω κατηγοριοποίηση ισχύει και για το μεγάλο δείγμα φοιτητών/τριων της έρευνάς μας.

Δ36. **Fouskakis, D.** and Ntzoufras, I. (2020). Power-Expected-Posterior Priors as Mixtures of g-Priors. *Bayesian Analysis (accepted).*

Μια από τις κύριες μεθόδους που χρησιμοποιούνται για την κατασκευή πρότερων κατανομών υπό την αντικειμενική Μπεϋζιανή μεθοδολογία είναι αυτή που χρησιμοποιεί τυχαία φανταστικά δεδομένα. Υπό την εν λόγω σκοπιά, η μεταγενέστερη-αναμενόμενη εκ των προτέρων κατανομή *(expected-posterior prior (EEP))* προσφέρει πολλά πλεονεκτήματα, μεταξύ των οποίων έχει μια ωραία και απλή ερμηνεία και παρέχει έναν αποτελεσματικό τρόπο καθορισμού συμβατών πρότερων μεταξύ των υπό σύγκριση μοντέλων. Σε αυτή την εργασία μελετάμε την δυναμικά-μεταγενέστερη-αναμενόμενη εκ των προτέρων κατανομή *(power-expected-posterior (PEP) prior),* η οποία αποτελεί γενίκευση της EPP, στην αντικειμενική Μπεϋζιανή επιλογή κανονικά γραμμικών μοντέλων. Αποδεικνύουμε ότι μπορεί να αναπαρασταθεί ως μια μίξη g-πρότερων κατανομών, όπως και ένα ευρύ φάσμα άλλων πρότερων κατανομών σε κανονικά γραμμικά μοντέλα. Έτσι οι ύστερες κατανομές και οι παράγοντες Μπέϋς προκύπτουν σε κλειστή μορφή διατηρώντας συνεπώς της υπολογιστική ευχρηστία τους. Με την βοήθεια αυτού του αποτελέσματος αποδεικνύουμε ότι τα κριτήρια για αντικειμενική σύγκριση Μπεϋζιανών μοντέλων ισχύουν κάτω από την PEP. Συγκρίσεις με άλλες μίξεις g-πρότερων κατανομών γίνονται και τα αποτελέσματα παρουσιάζονται σε προσομοιωμένα και πραγματικά δεδομένα.

Δ37. Tzoumerkas, G., Fouskakis, D. and Ntzoufras, I. (2022). A Comparison of Power-Expected-Posterior Priors in Shrinkage Regression. *Journal of Statistical Theory and Practice*, **16**, 61.

Η δομή των Δυναμικά-Μεταγενέστερων-Αναμενόμενων εκ των προτέρων κατανομών (Power-Expected- Posterior (PEP) priors) μας προσφέρει μια κατάλληλη και αντικειμενική μεθοδολογία, όσον αφορά το πρόβλημα της μπεϋζιανής επιλογής μεταβλητών, σε μοντέλα γραμμικής παλινδρόμησης. Η PEP πρότερη διαθέτει όλα τα πλεονεκτήματα των μεταγενέστερων αναμενόμενων εκ των προτέρων κατανομών (EPP). Επιπροσθέτως, αποφεύγει την αυθαίρετη επιλογή διδακτικού (φανταστικού) δείγματος, ενώ συγχρόνως μετριάζει την επίδρασή του στην τελική πρότερη. Υπό τη μεθοδολογία των PEP πρότερων μια αρχική (συνήθως προκαθορισμένη) πρότερη ανανεώνεται χρησιμοποιώντας φανταστικά δεδομένα. Η εργασία επικεντρώνει σε κανονικά μοντέλα γραμμικής παλινδρόμησης, όπου το πλήθος των παρατηρήσεων n είναι μικρότερο από το πλήθος των επεξηγηματικών μεταβλητών p. Παρουσιάζουμε τη μεθοδολογία των PEP πρότερων χρησιμοποιώντας διαφορετικές αρχικές πρότερες συρρίκνωσης. Εν συνεχεία, παρουσιάζουμε την υπολογιστική μέθοδο και πραγματοποιούμε συγκρίσεις με τη χρήση προσομοιωμένων και πραγματικών δεδομένων.

Δημοσιεύσεις Υποβληθείσες για Κρίση σε Περιοδικά

Υ1.

Δημοσιεύσεις σε Πρακτικά Συνεδρίων

Π1. **Fouskakis, D.** and Draper, D. (1998). Stochastic optimization methods for cost-effective quality assessment in health. COMPSTAT 1998, Proceedings in Computational Statistics, Short Communications and Posters. Harpenden: IACR-Rothamsted.

Στην παρούσα εργασία χρησιμοποιούμε Μπεϋζιανή θεωρία αποφάσεων για την επίλυση ενός μεγάλης κλίμακας προβλήματος βελτιστοποίησης το οποίο προκύπτει κατά την εκτίμηση της ποιότητας νοσοκομειακής περίθαλψης χρησιμοποιώντας δεδομένα από μια μεγάλη έρευνα της δεκαετίας του 80 στις ΗΠΑ. Χρησιμοποιούμε τη ***Μέθοδο Προσομοιούμενης Ανόπτησης*** (*Simulated Annealing (SA)*), το ***Γενετικό Αλγόριθμο*** (*Genetic Algorithm (GA)*), τη ***Μέθοδο Απαγορευμένης Αναζήτησης*** (*Tabu Search (TS)*) και τη ***Μέθοδο Προσομοιούμενης Μεταβολής Θερμοκρασίας*** (*Simulated tempering (ST)*) για την εύρεση του 'βέλτιστου' υποσυνόλου επεξηγηματικών μεταβλητών.

Π2. Charitidou, E., **Fouskakis, D.** and Ntzoufras, I. (2013). On Bayesian Transformation Selection: Problem Formulation and Preliminary Results. *Proceedings of the 26th Panhellenic Statistics Conference*, 253-260.

Στην παρούσα εργασία, εξετάζουμε το πρόβλημα της επιλογής οικογένειας μετασχηματισμού από την Μπεϋζιανή σκοπιά, αφού προηγουμένως αναφερθούμε στην υπάρχουσα σχετική βιβλιογραφία συμπεριλαμβάνοντας κλασικές και Μπεϋζιανές προσεγγίσεις. Ο βασικός στόχος είναι να προσεγγίσουμε την κανονικότητα της κατανομής ενός συνόλου παρατηρήσεων επιστρατεύοντας τις εξής οικογένειες μετασχηματισμών: *Box-Cox, Modulus, Yeo & Johnson και Dual.* Για την προσομοίωση δείγματος από την εκ των υστέρων κατανομή της παραμέτρου μετασχηματισμού λ_T, η οποία συνδέεται με κάθε οικογένεια μετασχηματισμού T, κατασκευάστηκαν αλγόριθμοι Markov Chain Monte Carlo. Παράλληλα, παρουσιάζουμε δύο προσεγγίσεις κατασκευής συμβατών πρότερων κατανομών (μεταξύ των διαφορετικών οικογενειών μετασχηματισμών) για την παράμετρο λ_T, κάνοντας χρήση της δυναμικής πρότερης κατανομής και μιας κανονικής εκ των προτέρων κατανομής. Η σύγκριση της επίδοσης των οικογενειών βασίζεται στις εκ των υστέρων πιθανότητες των οικογενειών των μετασχηματισμών. Χρησιμοποιώντας προσομοιωμένα δεδομένα, διαφόρων κατανομών, αναδεικνύεται η αποτελεσματικότητα της εν λόγω μεθοδολογίας.

Π3. Charitidou, E., **Fouskakis, D.** and Ntzoufras, I. (2014). On Bayesian transformation selection: Problem formulation and preliminary results. In Lanzarone, E. & Ieva, F. eds. The Contribution to Young Researchers in Bayesian Statistics. *Research from BAYSM 2013, Springer Proceedings in Mathematics and Statistics*, Springer-Verlang, Berlin, **63**, 11-14.

Το πρόβλημα της επιλογής οικογένειας μετασχηματισμού εξετάζεται πλήρως από Μπεϋζιανή σκοπιά. Οι κάτωθι οικογένειες μετασχηματισμού λαμβάνονται υπόψη με σκοπό να προσεγγίσουμε την κανονικότητα όσον αφορά στην κατανομή ενός συνόλου δεδομένων: *Box-Cox, Modulus, Yeo & Johnson και Dual.* Αλγόριθμοι Markov Chain Monte Carlo κατασκευάστηκαν ώστε να προσομοιώσουμε δείγμα από την εκ των υστέρων κατανομή της παραμέτρου μετασχηματισμού λ_T η οποία συνδέεται με κάθε οικογένεια μετασχηματισμού T. Διερευνήσαμε διαφορετικές προσεγγίσεις με στόχο την κατασκευή συμβατών πρότερων κατανομών για την παράμετρο λ_T μεταξύ των οικογενειών, κάνοντας χρήση της δυναμικής πρότερης κατανομής και εναλλακτικά μιας κανονικής εκ των προτέρων κατανομής. Η επιλογή και ανάδειξη της βέλτιστης οικογένειας βασίζεται στον υπολογισμό των εκ των υστέρων πιθανοτήτων των οικογενειών-μοντέλων. Χρησιμοποιώντας προσομοιωμένα δεδομένα, αναδεικνύεται η αποτελεσματικότητα της μεθοδολογίας που περιγράφεται στην παρούσα εργασία.

Π4. Perrakis, K., **Fouskakis, D.** and Ntzoufras, I. (2015). Bayesian variable selection for generalized linear models using the power-conditional-expected-posterior-prior. In Frühwirth-Schnatter, S., Bitto, A., Kastner, G. & Posekany, A. eds. Bayesian Statistics from Methods to Models and Applications, *Research from BAYSM 2014, Springer Proceedings in Mathematics and Statistics*, Springer-Verlang, Berlin, **126**, 59-73.

Η δυναμική-υπό συνθήκη-αναμενόμενη-μεταγενέστερη (power-conditional-expected-posterior – PCEP) πρότερη κατανομή, η οποία αναπτύχθηκε για επιλογή μεταβλητών σε κανονικά μοντέλα γραμμικής παλινδρόμησης, συνδυάζει στοιχεία των δυναμικών πρότερων κατανομών και των μεταγενέστερων-αναμενόμενων πρότερων κατανομών, που στηρίζονται σε ένα προκαθορισμένο δείγμα «φανταστικών» δεδομένων και οδηγεί σε μια συνεπή διαδικασία επιλογής μεταβλητών που δίνει περισσότερο βάρος σε φειδωλά μοντέλα. Στην παρούσα εργασία επεκτείνουμε την PCEP μεθοδολογία σε γενικευμένα γραμμικά μοντέλα (generalized linear models - GLMs). Δίνουμε τον ορισμό της PCEP πρότερης κατανομής στο πλαίσιο των GLMs, εξηγούμε την σχέση της με άλλες ευρέως διαδεδομένες εκ-των-προτέρων κατανομές και παρουσιάζουμε διάφορες εκφράσεις της PCEP ύστερης κατανομής οι οποίες μπορούν να χρησιμοποιηθούν για συμπερασματολογία σε συγκεκριμένα μοντέλα καθώς και για επιλογή μεταβλητών. Η προτεινόμενη μεθοδολογία υλοποιείται σε ένα παράδειγμα λογιστικής παλινδρόμησης με δίτιμα δεδομένα Bernoulli. Τα αποτελέσματα υποδηλώνουν ότι η διαδικασία επιλογής μεταβλητών μέσω της PCEP πρότερης κατανομής καταλήγει σε φειδωλά μοντέλα λογιστικής παλινδρόμησης, όπως συμβαίνει και στην περίπτωση των κανονικών γραμμικών μοντέλων. Παρούσες δυσκολίες και πιθανές λύσεις σχετικά με την γενίκευση της μεθοδολογίας στην ευρύτερη οικογένεια των GLMs συζητούνται.

Π5. Mpousiou, D., Lamprou, D., Toumpis, M., Katsaounou, T., **Fouskakis, D.**, Moscholaki, M., Karathanasi, A., Gratziou, C., Zervas, E. and Katsaounou, P. (2018). The effect of parental smoking and smoking inside the house in the adolescents attitude towards smoking. European Respiratory Journal 2018 52: Suppl. 62, PA4571.

Η οικογένεια αποτελεί το πρώτο κοινωνικό περιβάλλον μέσα στο οποίο αναπτύσσεται το παιδί και διαμορφώνει την κοινωνική συμπεριφορά και τις στάσεις του. Μεγάλος αριθμός στάσεων, πεποιθήσεων και συμπεριφορών που σχετίζονται με την υγεία στους εφήβους και τους ενηλίκους διαμορφώνονται από την παιδική ηλικία και ξεκινούν από την οικογένεια. Οι γονείς μπορούν να ενισχύσουν θετικά ή αρνητικά τις συμπεριφορές των παιδιών τους ως προς το κάπνισμα. Το γονικό κάπνισμα ασκεί άμεση επίδραση στην

καπνιστική συμπεριφορά των παιδιών. Εξετάζουμε την επίδραση του καπνίσματος των γονέων καθώς και την επίδραση της τήρησης του κανόνα περί απαγόρευσης του καπνίσματος στο σπίτι, στην πρόθεση και στην καπνιστική συμπεριφορά εφήβων μαθητών.

Π6. Tzoumerkas, G. and **Fouskakis, D.** (2021). Using the Power-Expected-Posterior Prior in Shrinkage Regression: A Simulation Study. *Proceedings of the 33rd Panhellenic Statistics Conference*, 345-355.

Η Δυναμικά-Μεταγενέστερη-Αναμενόμενη εκ των προτέρων κατανομή (Power-Expected- Posterior (PEP) prior) μας παρέχει μια κατάλληλη μεθοδολογία, στο πρόβλημα της μπεϋζιανής επιλογής μεταβλητών σε μοντέλα γραμμικής παλινδρόμησης, με κανονικά κατανεμημένα σφάλματα. Η μεθοδολογία των PEP εκ των πρότερων κατανομών χρησιμοποιεί διδακτικά δεδομένα για να ανανεώσει μια αρχικά επιλεγμένη πρότερη (baseline prior) κατανομή. Όταν το μέγεθος του δείγματος n είναι μικρότερο του πλήθους των επεξηγηματικών μεταβλητών p, η επιλογή της αρχικής πρότερης κατανομής είναι σημαντική. Οι πρότερες κατανομές συρρίκνωσης (shrinkage priors) έχουν αξιοσημείωτες θεωρητικές ιδιότητες και μπορούν να χρησιμοποιηθούν σε τέτοιες περιπτώσεις. Χρησιμοποιώντας πρότερες κατανομές συρρίκνωσης, ως αρχικές πρότερες κατανομές στην PEP μεθοδολογία, δημιουργείται μια νέα κλάση μπεϋζιανών αντικειμενικών πρότερων κατανομών (PEP-Shrinkage), κατάλληλων για προβλήματα παλινδρόμησης με $n < p$. Στην εργασία αυτή παρουσιάζουμε συνοπτικά τη μεθοδολογία των PEP-Shrinkage πρότερων κατανομών. Σε προσομοιωμένα δεδομένα συγκρίνουμε τα αποτελέσματα που παίρνουμε όταν εφαρμόζουμε την PEP-Shrinkage μεθοδολογία, με διαφορετικές πρότερες κατανομές συρρίκνωσης ως αρχικές πρότερες κατανομές. Επιπρόσθετα, γίνονται συγκρίσεις στα αποτελέσματα των PEP-Shrinkage εκ των πρότερων κατανομών με αυτά που λαμβάνουμε από τις πρότερες συρρίκνωσης χωρίς την χρήση της PEP μεθοδολογίας.

Π7. Tzoumerkas, G. and **Fouskakis, D.** (2022). Power-Expected-Posterior Methodology with Baseline Shrinkage Priors. In Raffaele Argiento, Federico Camerlenghi, Sally Paganin (eds.), Methodological and Computational Contributions on Bayesian Statistics, *Research from BAYSM 2021, Springer Proceedings in Mathematics and Statistics*, Springer-Verlang, Berlin, accepted.

Οι δυναμικά μεταγενέστερες αναμενόμενες εκ των προτέρων κατανομές (Power-Expected-Posterior (PEP) priors) προσφέρουν μια κατάλληλη και αντικειμενική μεθοδολογία όσον αφορά τα προβλήματα Μπεϋζιανής επιλογής μεταβλητών σε μοντέλα γραμμικής παλινδρόμησης. Οι PEP πρότερες διαθέτουν όλα τα πλεονεκτήματα των μεταγενέστερων αναμενόμενων εκ των προτέρων κατανομών (EPP) και επιπλέον αποφεύγουν την αυθαίρετη επιλογή διδακτικού (φανταστικού) δείγματος, ενώ συγχρόνως μειώνουν την επίδρασή του στην τελική πρότερη. Υπό τη μεθοδολογία των PEP πρότερων μια αρχική (συνήθως προκαθορισμένη) πρότερη ανανεώνεται χρησιμοποιώντας φανταστικά δεδομένα. Η εργασία επικεντρώνει σε κανονικά μοντέλα γραμμικής παλινδρόμησης, όπου το πλήθος των παρατηρήσεων n είναι μικρότερο από το πλήθος των επεξηγηματικών μεταβλητών p. Παρουσιάζουμε τη μεθοδολογία των PEP πρότερων χρησιμοποιώντας διαφορετικές αρχικές πρότερες συρρίκνωσης και πραγματοποιούμε συγκρίσεις με τη χρήση προσομοιωμένων δεδομένων.

Συγγραφικό Έργο

1. Γ. Κοκολάκης και **Δ. Φουσκάκης** (2005). *Σημειώσεις Στατιστικής* (σελ. 196).

   Αναπτύσσεται κατά τρόπο συστηματικό και με το απαιτούμενο βάθος και έκταση η βασική Εισαγωγική Στατιστική Θεωρία. Οι σημειώσεις αυτές απευθύνονται στους σπουδαστές του 4ου εξαμήνου της σχολής Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών του ΕΜΠ. Καλύπτουν με την απαιτούμενη σαφήνεια, πληρότητα και αυστηρότητα την Περιγραφική Στατιστική, την Εκτιμητική Στατιστική Θεωρία, Την Κατασκευή Διαστημάτων Εμπιστοσύνης και την Θεωρία Ελέγχων Στατιστικών Υποθέσεων, με χρήση πολλών αποσαφηνιστικών παραδειγμάτων και εφαρμογών.

2. Γ. Κοκολάκης και **Δ. Φουσκάκης** (2009). *Στατιστική Θεωρία & Εφαρμογές* (σελ. 370). Εκδόσεις Συμεών. Αθήνα.

   Στο εν λόγω σύγγραμμα αναπτύσσεται σταδιακά και συστηματικά η απαιτούμενη μαθηματική υποδομή με βάση την οποία παρουσιάζεται μια μαθηματικά θεμελιωμένη μεθοδολογία της Στατιστικής Συμπερασματολογίας. Ταυτόχρονα δίνεται έμφαση στις προϋποθέσεις που διασφαλίζουν την εγκυρότητα μιας στατιστικής τεχνικής. Με το τρόπο αυτό δίνεται η δυνατότητα στον αναγνώστη να γνωρίζει τις συνθήκες κάτω από τις οποίες η εφαρμοζόμενη μέθοδος μπορεί να μην είναι κατάλληλη και τα συμπεράσματα που προκύπτουν να μην είναι ασφαλή. Μεγάλος αριθμός παραδειγμάτων και εφαρμογών αποσαφηνίζουν τη θεωρία και παρέχουν στον σπουδαστή ένα αρκετά ευρύ φάσμα εμπειριών σχετικά με την ανάλυση στατιστικών δεδομένων. Καλύπτονται ενότητες όπως Δειγματοληψία, Περιγραφική Στατιστική, Εκτιμητική Στατιστική Θεωρία, Έλεγχοι Υποθέσεων, Ανάλυση Παλινδρόμησης και Διασποράς, Ανάλυση Κατηγορικών Δεδομένων και Απαραμετρική Στατιστική Μεθοδολογία, ενώ ως Παράρτημα υπάρχουν Στοιχεία της Θεωρίας Πιθανοτήτων.

3. **Δ. Φουσκάκης** (2013). *Ανάλυση Δεδομένων με Χρήση της R* (σελ. 504). Εκδόσεις Τσότρας. Αθήνα.

   Το εν λόγω βιβλίο, εκτός από έναν αρκετά λεπτομερή οδηγό χρήσης της στατιστικής γλώσσας προγραμματισμού R, χρησιμοποιεί επιπλέον τη συγκεκριμένη γλώσσα για την υλοποίηση των στατιστικών μεθόδων που αναπτύσσει, καλύπτοντας μεταξύ άλλων μεθόδους περιγραφικής στατιστικής, κατασκευής γραφημάτων, προσομοίωσης, ελέγχων υποθέσεων, ανάλυσης παλινδρόμησης και ανάλυσης διασποράς. Η παρουσίαση των μεθόδων αυτών γίνεται με απλό και κατανοητό τρόπο αποφεύγοντας τη λεπτομερή μαθηματική θεμελίωση τους και έμφαση δίνεται στο πότε μπορούν να εφαρμοστούν, ποιες είναι οι προϋποθέσεις τους και με ποιους τρόπους τις ελέγχουμε και πώς ερμηνεύουμε τα αποτελέσματα που λαμβάνουμε μετά την εφαρμογή τους. Στη συνέχεια παρουσιάζεται ο τρόπος υλοποίησης τους στην R, ενώ παράλληλα συμπεριλαμβάνεται ένας μεγάλος αριθμός παραδειγμάτων από διάφορα ερευνητικά πεδία (π.χ. Ιατρική, Κοινωνικές Επιστήμες, Διοίκηση Επιχειρήσεων, Μηχανική, Οικονομικές Επιστήμες κλπ) για καλύτερη κατανόηση. Έτσι ο αναγνώστης οδηγείται με συγκροτημένο και εύληπτο τρόπο στην εμπέδωση της θεωρίας και των πρακτικών που εφαρμόζονται σε βασικά προβλήματα ανάλυσης δεδομένων.

4. **Δ. Φουσκάκης** (2013). *Ανάλυση Δεδομένων με Χρήση της R, 2η Έκδοση* (σελ. 862). Εκδόσεις Τσότρας. Αθήνα.

   Η νέα έγχρωμη έκδοση του βιβλίου είναι ανανεωμένη, βελτιωμένη και επαυξημένη. Συγκεκριμένα, έχουν γίνει προσθήκες στο Κεφάλαιο 2 όπου πλέον γίνεται πλήρης αναφορά

στις συναρτήσεις apply() καθώς και σε χρήση ελληνικών χαρακτήρων στην R. Το Κεφάλαιο 2 ολοκληρώνεται με μια νέα παράγραφο στην οποία παρουσιάζεται το RStudio και το R Markdown. Στο Κεφάλαιο 7 έχει προστεθεί υλικό σχετικά με την επίδραση που έχουν στους συντελεστές του γραμμικού μοντέλου "συνήθεις" γραμμικοί μετασχηματισμοί (κεντράρισμα, τυποποίηση ή κανονικοποίηση) στις τιμές της μεταβλητής απόκρισης και/ή στις τιμές των ποσοτικών επεξηγηματικών μεταβλητών. Στο ίδιο κεφάλαιο αναπτύσσονται περαιτέρω τα πολλαπλασιαστικά μοντέλα, με ή χωρίς εικονικές μεταβλητές, ενώ επιπλέον έχει προστεθεί υλικό σχετικά με τον συντελεστή προσδιορισμού και τους κινδύνους που κρύβει η λανθασμένη χρήση του ως μέτρο καλής προσαρμογής. Το Κεφάλαιο 7 ολοκληρώνεται με μία νέα παράγραφο σχετικά με τα συνηθέστερα μέτρα σύγκρισης γραμμικών μοντέλων. Στην παρούσα έκδοση έχουν προστεθεί επίσης δύο νέα κεφάλαια. Στο Κεφάλαιο 9 εισάγονται, περιγράφονται και αναλύονται οι δυνατότητες των βιβλιοθηκών ggplot2 και data.table της R, καθώς και της βιβλιοθήκης shiny του RStudio, οι οποίες χρησιμοποιούνται πολύ συχνά για τη διαγραμματική απεικόνιση, το χειρισμό δεδομένων μεγάλης κλίμακας και τη δημιουργία διαδραστικών διαδικτυακών εφαρμογών από την R, αντίστοιχα. Παρουσιάζονται οι συνηθέστεροι μηχανισμοί διαγραμματικής αναπαράστασης δεδομένων προερχόμενων από ποσοτικές ή κατηγορικές μεταβλητές, σε μία ή και περισσότερες διαστάσεις, καθώς επίσης και τα συνηθέστερα διαγράμματα που δημιουργούμε για να ανακαλύψουμε το είδος της εξάρτησης δεδομένων ιδίου ή διαφορετικού είδους, με χρήση της βιβλιοθήκης ggplot2. Επιπλέον, παρουσιάζονται, μέσω της βιβλιοθήκης data_table, οι συνηθέστεροι μηχανισμοί χειρισμού δεδομένων, όπως συνάθροιση, ομαδοποίηση, ταξινόμηση, επιλογή, κ.λπ. Τέλος, γίνεται αναφορά στη βασική δομή συγγραφής shiny εφαρμογών και παρουσιάζονται διάφορα εργαλεία διαμόρφωσης τους, μέσω παραδειγμάτων. Στο Κεφάλαιο 10 παρουσιάζεται η μεθοδολογία των μοντέλων της δίτιμης λογιστικής παλινδρόμησης, οι τρόποι προσαρμογής αυτών με τη βοήθεια της R, ενώ έμφαση δίνεται στην ερμηνεία των συντελεστών καθώς και στον έλεγχο των προϋποθέσεων τους. Επιπλέον, παρουσιάζονται οι έλεγχοι καλής προσαρμογής, ενώ αναφορά γίνεται και στο πρόβλημα της ταξινόμησης, καθώς και της χρήσης μεθόδων διασταυρωμένης επικύρωσης.