

Esildis ja põhjendus akadeemikukandidaadi Jüri Lemberi esitamiseks matemaatika alal

Teen ettepaneku Tartu Ülikooli Senatile esitada Eesti Teaduste Akadeemia akadeemikukandidaadiks matemaatika alal tõenäosusteooria professor Jüri Lemberi.

Jüri Lember (JL) on Eesti juhtiv tõenäosusteoreetik ja (teoreetiline) statistik ning rahvusvaheliselt tunnustatud teadlane. Ta on Tartu Ülikooli (TÜ) stohhastika töögrupi liider; JL juhtimisel on sellest grupist saanud maailma juhtiv keskus mitmekordsete Markovi mudelite uurimisel. JL on andnud suure panuse tõenäosusteooria õpetamisesse ja arengusse Eestis, ta on toonud Tartusse uusi uurimissuundi, töötanud välja uusi ja kaasaegseid kursusi ning loonud palju eestikeelset õppematerjali. Tema teaduslik haare on lai ning ta on teinud edukalt koostööd mitmete juhtivate matemaatikutega.

JL on publitseerinud üle 50 artikli, neist 36 ETISE kategoorias 1.1 (6 artiklit kategoorias 3.1, 6 artiklit kategoorias 1.2), tema h -indeks on Google Scholaris ja RG järgi 16.

Kandidaadi andmed

Jüri Lember

Sündinud 01.06.1968

https://www.etis.ee/CV/Jüri_Lember/est/

Eluloolised andmed

JL on sündinud 1. juunil 1968. aastal Tallinnas, lõpetas Tallinna 17. keskkooli (1986), TÜ rakendusmatemaatika eriala (1994) ning kaitses doktorikraadi TÛs (1999). Doktorantuuri ajal ja peale seda (1994 - 2000) töötas JL TÜ Matemaatilise statistika instituudis teadurina; 2001-2003 järel doktorantuur rahvusvahelises uurimisinstituudis Eurandom (Eindhoveni tehnikaülikool, Holland). Peale järel doktorantuuri naases JL TÜ Matemaatilise statistika instituuti, kus töötas algul lektori ja hiljem vanemteadurina.

Alates 01.09.2019 on JL TÜ tõenäosusteooria professor.

Külalisõppejõu ja -teadlasena on JL töötanud veel Bielefeldi ülikoolis Saksamaal, Aalto ülikoolis Soomes, Telecom Sud-Pariisi ülikoolis Prantsusmaal.

Teadustöö põhisuunad

JL teadustöö valdkond on stohhastika (tõenäosusteooria ja matemaatiline statistika) ning uuritavad teadusteemad ulatuvad teoreetilisest tõenäosusteooriast matemaatilise statistika ja andmeanalüüsini. Selline lai haare on muljetavaldav, enamasti kipuvad matemaatikud kapselduma ühte kitsasse valdkonda. Alljärgnevas ülevaade mõnest olulisemast uurimissuunast.

M-hinnangute mõjus. Mõjus (consistency) on matemaatilise statistika keskne mõiste, mis (lihtsustatult) tähendab seda, et valimi mahu kasvades läheneb andmete põhjal saadud mingi näitaja hinnang üldkogumi vastavale näitajale. See on ülioluline omadus, mis muuhulgas on ka kõiksugu keskmiste arvutamise teoreetiline põhjendus. M-hinnangud on teatavat tüüpi hinnangute klass, enamik praktikas kasutatavaid hinnanguid on M-hinnangud (näiteks suurima tõepära hinnangud, vähima ruutkeskmise hinnangud, k-keskmised). Seega (M-)hinnangute mõjususe tõestus on asümptootilise statistika oluline haru ning enamasti põhinevad tõestused kompaktsusel. Oma doktoritöös ning sellega seotud artiklites uuris JL võimalusi mõjususe tõestamiseks üldistes meetrilistes ruumides ilma kompaktsuseta. Kuigi peamine uurimisobjekt (uuritav hinnang) neis töödes olid k-keskmised - tavalise (aritmeetilise) keskmise k-punktilised üldistused, rakenduvad neis artiklites välja töötatud meetodid ka muudele M-hinnangutele. Sellealaste tööde oluline tulemus on side asümptootilise statistika ja pealtnäha kaugelseisva lähendamisteooria vahel. Päril viimastel aastatel on JL selle temaatika juurde tagasi pöördunud - koos kolleegidega TÜ-st välja töötatud uus meetod normaaljaotuste segu parameetrite hindamiseks. Normaaljaotuste segu on väga levinud mudel mille parameetrite suurima tõepära hinnang klassikalisel kujul puudub. Pakutav uus hinnang asendab seda, kuid selle mõjususe tõestamine on ebastandardne ja nõuab suuresti uut lähenemist.

Juhusliku koodi taastamine. Kujutlegem kahepoolset binaarset (nullide ja ühtede)jada - koodi (scenery) - mida mööda juhuslikult eksleb koodilugeja. Juhuslik ekslemine tähendab, et koodilugeja (piltlikult) viskab münte otsustamiseks kas ta teeb sammu paremale või vasemale. Peale sammu tegemist registreerib koodilugeja selle numbri (0 või 1), millele ta sattus. Nii eksleb koodilugeja lõpmatult ning kui vasakul ja paremale astumise tõenäosus on võrdne, siis loeb ta iga koodinumbrit lõpmata palju kordi. Probleem on taastada algne kood (teatud täpsuseni) teadmata koodilugeja eksleja trajektoori. Et asi oleks põnevam, oletagem et ka loetav kood on juhuslik, saadud näiteks mündivisete tulemusena. Probleemi peeti lahendamatuks kuni H. Matzinger oma doktoritöös (1990.-ndatel) näitas, et põhimõtteliselt on koodi restaureerimine võimalik. See läbimurre avas diskreetse tõenäosusteooria uue peatüki. JL liitus Matzingeri grupiga järel doktorantuuris, koos on avaldatud mitu artiklit. Kuigi lihtsalt sõnastatav, nõuab selle lahendamine tihti uut lähenemisviisi ja tehnikat, mida saab kasutada ka muudes matemaatika valdkondades. Näiteks JL ja Matzingeri Fourier teisendustel põhinev perioodilise koodi taastamine pakub huvi algebraistidele.

Mitteparameetiline Bayesi statistika. Arvutusvõimsuste kasvades muutuvad Bayesi meetodid statistikas iga aastaga järjest populaarsemaks. Mitteparameetrilises analüüsis konstrueeritakse eelmöödud lõpmatummõõtmelistel ruumidel ning eesmärk on need möödud konstrueerida nii, et järelmöödu (posterior) koondumine tegelikult jaotuseks oleks

võimalikult kiire See uurimissund tugineb suuresti A. van der Vaardi ja S. Ghosali läbimurdeartiklil (2000), milles nad esitasid üldise raamistiku järeldamõõdu koondumiseks. Seega oli nullindate alguseks küll selge kuidas antud mudeli korral konstrueerida parima võimaliku koondumiskiirusega eelmõõt, kuid ei teatud, kuidas kombineerida erinevate koondumiskiirustega mudeleid ehk kuidas (ja kas on üldse võimalik) leida mudelite-vahelised nn adapteerivad eelmõõdud. Suuresti järel doktorantuuris tehtud uurimistöö tulemusena esitas JL teadaolevalt esimesed adapteerivad eelmõõdud (splainidel ja Hölderite ruumidel). Vastavad artiklid (kaasautorid A. van der Vaart ja S. Ghosal) ilmusid aastaid hiljem ja on siiani JLi enim tsiteeritud uurimused. Aastal 2017 ilmus A. van der Vaarti ja S. Ghosali fundamentaalne koguteos "Fundamentals of Nonparametric Bayesian Inference", mis postuleeris uurimisvaldkonna *status quo*. Adapteerivatele eelmõõdudele ja muuhulgas ka JL vastavasisulistele tulemustele on pühendatud üks selle raamatu peatükk.

Juhuslike jadade võrdlus. Jadade sarnasuse mõõtmine on kesksel kohal paljude teadusharudes nagu näiteks molekulaarbioloogia (DNA-järjestused, valgud) või lingvistika (tekstid). Enamasti mõõdetakse jadade sarnasust läbi nn sarnasusmõõdiku, vast enamlevinum mõõdik on pikima ühisjada (*longest common subsequence*) pikkus - mida pikem ühisjada, seda sarnasemad jadad. Samas on ka teineteisega mitteseotud jadalatel teatav sarnasus ning võtmeküsimus on sisulise sarnasuse eristamine juhuslikust kokkulangemisest. See aga eeldab, et jadad on modelleeritud juhuslikena ning appi tuleb tõenäosusteooria. Sel teemal JL publitseerinud üle kümne artikli, peamine kaasautor H. Matzinger.

Kuigi sarnasusmõõdikud on lihtsalt deneeritavad, puuduvad neil enamasti analüütilised kujud ning see teeb nende teoreetilise uurimise erakordselt raskes. Näiteks on suhteliselt lihtne näidata, et kahe sõltumatu sama pikkusega i.i.d. jada sarnasusmõõdiku (näiteks pikima ühisjada pikkuse) ning jada pikkuse suhe läheneb konstandile (nn Chvatal-Sankovi konstant). Konstandi olemasolu on lihtne näidata, kuid selle täpne väärtus on teadmata, kuigi seda on intensiivselt püütud leida juba üle 50 aasta.

Suur osa JL sellealastest uurimistöödest on pühendatud sarnasusmõõdiku dispersiooni DL asümptootikale: tõestada, et leiduvad konstandid $0 < c < C$ nii, et iga n korral $c < DL < C$. Väljakutse on just alumise tõkke c olemasolu tõestus, selle näitamiseks töötasid JL ja H. Matzinger välja täiesti uude lähenemise.

Lisaks dispersiooni asümptootikale on JL uurinud ka muude momentide asümptootikat, keskmise (suhtelise) skoori koondumiskiirust ning suboptimaalseid joondusi. Selle valdkonna oluline tulemus on alternatiivse nn ekstremaalsetel joondustel põhineva sarnasusmõõdiku väljatöötamine ja selle omaduste uurimine. Ekstremaalsed joondused on suuresti läbi uurimata kui suure potentsiaaliga töövahend.

Mitmekordsed Markovi mudelid. (Mitmekordsed) Markovi mudelid on viimasel kahel aastakümnel olnud JL ehk domineerivaim uurimisteema. Ka see suund sai alguse järel doktorantuuris, ajendatuna tollaegsetest varjatud Markovi ahelatel põhinevatel kõnetuvastussüsteemidest. Nende treenimine põhines nn Viterbi joondusel, kuid selle joonduse ja seeläbi ka treeningsüsteemi omadusi polnud selleks ajaks keegi sisuliselt uurinud. See motiveeris JLi konstrueerima nn Viterbi protsessi, mis omakorda võimaldas tõestada

Viterbi joonduse ning seeläbi ka treeningsüsteemi teatavad (asümptootilised) omadused. Neid omadusi arvesse võttes töötasid JL ja A. Koloydenko välja nn parandatud Viterbi treeningu varjatud Markovi ahela (*hidden Markov models*, HMM) parameetrite hindamiseks.

Viterbi joondus on saanud oma nime kuulsalt Viterbi algoritmilt, mis on populaarseim töövahend segmenteerimiseks varjatud ahela realisatsiooni prognoosimiseks andmete põhjal. Viterbi joondus pole muud kui suurima (tingliku) tõenäosusega realisatsioon ning selle joonduse kasutamine segmenteerimises tähendab kõikvõimalikest stsenaariumitest kõige tõenäolisema välja valimine. Intuiitselt arusaadav kuid tihti valedele järeldustele viiv tee. Nii näiteks pole see parim joondus prognoosivigade (keskmise) arvu mõttes; selles mõttes on parim üks hoopis teine joondus PMAP joondus mis aga võib olla väga väikese tõenäosusega. Praktikas pakub tihti huvi prognoos, millel oleksid mõlema head omadused: suur tõenäosus ja väike vigade arv. Leidmaks sellist prognoosi töötas JL välja tehiseõppel (*statistical learning*) põhineva üldise segmenteerimisteooria ja pakkus välja nn hübriidjoondused. Hübriidjoondused kombineerivad Viterbi ja PMAP joondusehäid omadusi ning neid on võimalik lihtsalt leida Viterbi algoritmiga sarnase algoritmiga. Viimastel aastatel on hübriidjoondused leidnud tee rakendusteni. Hiljuti kasutasid Taani teadlased hübriidjoondusi DNA arhailiste segmentide identifitseerimiseks.

HMM on lihtne ja väga populaarne kahedimensionaalne (vaatlused ja varjatud tunnus) juhuslik protsess, mille oluline tunnus on Markovi omadus. Just see omadus on kõikide ülalkirjeldatud algoritmide alus. Selgub aga, et (kahedimensionaalne) Markovi omadus on palju laiemal mudelite klassil kui HMM. Selliseid nn kahekordsed Markovi mudelid on modelleerimisel palju paindlikumad kui HMM, kuid HMMide töövahendid (algoritmid) on ikka rakendatavad. Kahekordsed (üldisemalt mitmekordsed) Markovi mudelid on viimastel aastatel JL peamised uurimisobjektid; tema juhitud töögrupp Tartus ja W. Pieczynski juhitud töögrupp Pariisis on ilmselt maailma juhtivad keskused selliste mudelite uurimisel. Teoreetilise uurimistöö üks eesmärk on ülalkirjeldatud ja seni vaid HMMide korral tõestatud tulemuste (nt Viterbi protsessi konstrueerimine, asümptootilised riskid, eksponentsiaalne unustamine jne) üldistamine. JL peamine kaasautor sellel teel on J. Sova. Et kahekordsed Markovi mudelid on üllatavalt lai klass juhuslikke protsesse, pakuvad nad huvi ka eelpoolmainitud juhuslike jadade võrdlemisel. Enamasti modelleeritakse võrreldavaid jadu sõltumatuna, kuid kahekordsed Markovi mudelid võimaldavad modelleerida sõltuvust nii jadade kui ka ühe jada elementide vahel. JL (koos kaasautoritega) oli esimene, kes võttis selliseid mudelid kasutusele jadade võrdluses.

Evolutsioonimudelid. Need mudelid modelleerivad populatsiooni geneetilise struktuuri muutumist ajas. Koos C. Watkinsiga töötas JL välja päris uut tüüpi evolutsioonimodeli, mis üldistab erinevaid Moran-tüüpi mudeleid. Selle mudeli abil tõestas JL nn faasinihke - väikesed muutused parameetrites põhjustab suuri erinevusi (piir)populatsioonis. Koos Itaalia tõenäosusteoreetikute F. Zucca ja D. Bertacchiga üldistas JL mudelit (ja tõestas faasinihke) juhule, kus võimalikke genotüüpe on mitteloenduv hulk. See on matemaatiliselt keerukam kuid bioloogiliselt märksa realistlikum mudel. Faasinihe on nii matemaatiliselt ning ka sisuliselt huvipakkuv fenomen vajab edasist uurimist.

Teaduskorralduslik ja ühiskondlik tegevus

Alates aastast 2004 (peale järeldoktorantuuri) on JL olnud järgmiste grandiprojektide juht: ETF 5694 (2004-2007), ETF 7553 (2008-2011), ETF 9288 (2012-2016), SF0180015s12 (2012-2014, projekt hõlmas kogu tollase Matemaatilise statistika instituudi teadustööd), IUT34-5 (2015-2020, institutsionaalne uurimistoetus, hõlmas kogu Matemaatika ja statistika instituudi stohhastika-alast uurimistööd). Lisaks sellele on JL olnud põhitäitja muudes (nii Eesti kui ka rahvusvahelistes) grandiprojektides, näiteks PRG865 (2020-24).

JL on osalenud mitmete rahvusvaheliste konverentside ja teadusürituste organiseerimisel. Neist olulisim on vast Tartus toimunud rahvusvaheline statistikakonverents *Nordstat 2018*, kus JL oli programmikomitee juht; konverentsidel Nordstat 2010, 2023 organiseeris JL sektsiooni ja järgmisel, Nordstat 2026 konverentsil on JL üks kutsutud (*keynote*) ettekandjatest. Eelmise aasta sügisel toimus Tartus esmakordselt Markovi mudelitele pühendatud rahvusvaheline töötuba (*Tartu Workshop on Markov Modelling*), JL oli selle peamine korraldaja. Samuti lööb JL kaasa iga aasta maikuus toimuva Soome tõenäosusteooria suvekooli organiseerimises, 2023. aastal oli ta üks kutsutud lektoritest (minikursus Markovi mudelistest).

Aastast 2023 on JL maineka teadusajakirja "*Scandinavian Journal of Statistics*" toimetuskolleegiumi liige. Ajakiri pühendab eelpoolmainitud Markovi mudelite töötoale spetsiaalse sektsiooni ja JL on selle peatoimetaja.

Aastast 2021 on JL Eesti matemaatika seltsi president. Ligi 370 liikmega on matemaatika selts Eesti erialaseltsidest üks suurimaid, TAgA assotsieerunud seltsidest on suurema liikmeskonnaga vaid loodusuurijate selts ning emakeele selts. Seltsi presidentuur hõlmab suhtlemist ajakirjandusega, matemaatikõpetajatega, kooliõpilastega (eriti olümpiaadidel ja muudel matemaatikavõistlustel osalejatega), ürituste organiseerimist ja muud seesugust.

Eriala populariseerimiseks on JL on kirjutanud artikleid ja pidanud ettekandeid laiemale publikule (teoreetilise bioloogia suvekool, Eesti matemaatika päevad, ajakiri UT, raadiosaade "Kukkuv õun").

Pedagoogiline tegevus

JL juhendamisel on kaitstud üks doktoritöö (J. Sova), 17 magistritööd, kaks diplomitööd (Bielefeldi ülikool, võrdne magistrikraadiga) ning paarkümmend bakalaureusetööd.

TÜs on JL regulaarselt õpetanud aastast 2003, peamiselt tõenäosusteooria põhikursust. Aastal 2006 hakkas JL esimesena Tartus süstemaatiliselt õpetama tehisõppe/masinõppe (*machine/statistical learning*) kursust ja umbes samal ajal lõi ta uue informatsiooniteooria kursuse, ka seda ainet polnud Tartus enne õpetatud. Aastatel 2012 ja 2013 luges ta neid kursusi (tehiseõpe ja informatsiooniteooria) Aalto ülikoolis (Helsingi). JL on õpetanud ka Bielefeldi ülikoolis (mööduteooria). Paar aastat tagasi lõi JL uue ingliskeelse kursuse Bayesi statistikast. JLi kursusi iseloomustavad põhjalikud loengukonspektid, näiteks tema tõenäosusteooria kursuse konspekt (ca 250 lk) on kindlasti mahukaim ning põhjalikum eestikeelne tõenäosusteooria õppematerjal.

Preemiad ja tunnustused

1994 A. Humala noore teadlase preemia.

JL on kaks korda kandideerinud EV riiklikule teaduspreemiale (2012 ja 2022).

10 olulisimat teaduspublikatsiooni

1. Jüri Lember. "On minimizing sequences for k-centres." *Journal of Approximation Theory* 120.1 (2003): 20-35. [https://doi.org/10.1016/S0021-9045\(02\)00010-2](https://doi.org/10.1016/S0021-9045(02)00010-2)
2. Heinrich Matzinger, Jüri Lember. "Reconstruction of periodic sceneries seen along a random walk." *Stochastic Processes and their Applications* 116.11 (2006): 1584-1599. <https://doi.org/10.1016/j.spa.2006.03.006>
3. Subhashis Ghosal, Jüri Lember, and Aad van der Vaart. "Nonparametric Bayesian model selection and averaging." *Electronic Journal of Statistics* 2 (2008): 63-89. DOI: 10.1214/07-EJS090
4. Jüri Lember, Alexey Koloydenko. "The adjusted Viterbi training for hidden Markov models." *Bernoulli* 14.1 (2008): 180-206. DOI: 10.3150/07-BEJ1055.
5. Jüri Lember, Heinrich Matzinger. "Standard Deviation of the Longest Common Subsequence." *The Annals of Probability* (2009): 1192-1235. DOI: 10.1214/08-AOP436
6. Jüri Lember, Alexey Koloydenko. "A constructive proof of the existence of Viterbi processes." *IEEE Transactions on Information Theory* 56.4 (2010): 2017-2033. <https://doi.org/10.1109/TIT.2010.2040897>
7. Jüri Lember, Alexey Koloydenko. "Bridging Viterbi and posterior decoding: a generalized risk approach to hidden path inference based on hidden Markov models." *The Journal of Machine Learning Research* 15.1 (2014): 1-58. <https://jmlr.org/papers/volume15/lemb14a/lemb14a.pdf>
8. Jüri Lember, Heinrich Matzinger, Anna Vollmer. "Optimal alignments of longest common subsequences and their path properties." *Bernoulli* 20.3 (2014): 1292-1343. DOI: 10.3150/13-BEJ522
9. Jüri Lember, Joonas Sova. "Existence of infinite Viterbi path for pairwise Markov models." *Stochastic Processes and their Applications* 130.3 (2020): 1388-1425. <https://doi.org/10.1016/j.spa.2019.05.004>.
10. Bertacchi, Daniela, Jüri Lember, Fabio Zucca. "An evolution model with uncountably many alleles." *ALEA* 20 (2023): 1041-1075. DOI: 10.30757/ALEA.v20-38

(allkirjastatud digitaalselt)

Meelis Käärik

Tõenäosusteooria kaasprofessor

Matemaatika ja statistika instituudi juhataja