Tõenäosusjaotuste lähendamine hulkadega

Meelis Käärik
Tartu Ülikool

Tõenäosusjaotuste lähendamine hulkadega on oluline ülesanne nii klassikalises statistikas ja tõenäosusteoorias kui mitmetes praktilistes valdkondades. Teatavasti on ka sellised tuntud karakteristikud nagu juhusliku suuruse keskväärtus ja mediaan tema parimad ühepunktilised lähendid vastavalt ruutkauguse või absoluutse kauguse mõttes. Käesolevas töös vaadeldakse selle klassikalise ülesande kaugeleulatuvat üldistust, seda nii lähendhulkade valiku, kaofunktsiooni kuju kui ka põhiruumi enda osas.

Võimalike lähendhulkade valik on väga lai. Tuntumad näited on k-punktilised hulgad, sirged, tasandid, aga ka erinevad kõverad ja pinnad. Kõige sügavamad tulemused on kirjanduses seni saadud juhul, kui klass $\mathcal{A}$ koosneb k-punktilistest hulkadest. Selliste lähendhulkadega on tegemist näiteks kvantimise (analoogsignaali muutmine digitaalseks) korral. Suhteliselt hästi on lahendatud optimaalsete k-hulkade olemasolu, nende koondumise jt. küsimused (vt. [12], [3], [8]). Lähendhulkade koondumise probleem tekib siis, kui lähendhulgad on optimaalsed mõõtude jada $\{P_n\}$ suhtes, kus P_n koondub (nõrgalt) mõõduks P. Kirjeldatud situatsioon on praktikas väga levinud: nimelt on statistikas enamasti vaja hinnata üldkogumi kriteeriume, aga võimalik on kasutada ainult valimist saadud informatsiooni. On ilmne, et valimi karakteristikute koondumine üldkogumile vastavaks karakteristikuks on igati soovitav omadus, mille tõestamine üldjuhul aga pole sugugi triviaalne ülesanne. Seejuures on teada fakt, et valimi tekitatud empiiriline jaotus koondub nõrgalt üldkogumile vastavaks jaotuseks tõenäosusega 1. Eelpool mainitud töödes ongi vaatluse all just empiirilised mõõdud P_n ning väitekirja üks ülesanne oligi käsitleda üldisemat juhtu, mis haaraks ka näiteks ergoodiliste protsesside poolt indutseeritud mõõte.

Seoses arvutustehnika võimsuse pideva kasvuga on järjest enam hakatud uurima ka keerulisemaid mudeleid kui k punktiga lähen-

Eesti Matemaatika Selts Aastaraamat 2006

damine. Jaotuste ringjoontega lähendamise probleem on kerkinud mitmetes eri valdkondades. Huvitavaks näiteks on siin ühe Antiik-Kreeka staadioni rekonstrueerimine osaliselt säilinud stardiraja järgi ([16]). Lähendamine ringjoonte ja ellipsitega on oluline teema füüsikas osakeste kiirendamisel magnetväljas, astronoomias ([11]), lennukitööstuses, metroloogias, helilainete uurimisel ([17]) ja mujal.

Matemaatiliselt saame uuritava ülesande kirja panna järgmiselt. Olgu antud juhuslik element X jaotusega P separaablil meetrilisel ruumil (S, d) ja olgu $\mathcal{A}$ ruumi S teatud alamhulkade hulk. Lähendhulka $A \in \mathcal{A}$ nimetame optimaalseks (jaotuse P mõttes), kui ta minimiseerib järgmise kaofunktsiooni:

$$W(A, P) = E\varphi(d(X, A)),$$

kus $d(x, A)$ on kaugus punkti x ja lähendhulga A vahel ja hälbefunktsioon φ annab lähendhulgale tema kauguse põhjal "hinde".

Töös püstitatud põhiküsimused olid:

1) millal optimaalne lähendhulk üldse leidub?
2) kas kaofunktsioonide infiimumväärtuste ("parimate hinnete") jada koondub?
3) kas optimaalsete lähendhulkade jada $\{A_n\}$ koondub?

Paneme tähele, et "parimate hinnete" käitumise uurimine on sageli olulisemgi kui optimaalsete lähendite endi uurimine. Nimelt, tihti on meil lihtsalt vaja jaotuse P_n põhjal leida jaotuse P jaoks "piisavalt hea hindega" lähendit. Sellise ülesande püstituse korral ei olegi niivõrd oluline see, kas P_n-optimaalsed lähendid koonduvad, tähtis on, et just "hinnete" jada koonduks (vt. näiteks [18]).

Toodud probleeme on üsna laialdaselt uuritud ka varem ([12], [1], [3], [4], [13], [14], [2], [8], [9]). Töö eesmärk oli üldistada seniseid tulemusi antud valdkonnas kahes põhisuunas:

- valides võimalikult laia lähendhulkade klassi $\mathcal{A}$;
- tehes võimalikult vähe kitsendusi mõõtudele $\{P_n\}$ ja P, et kaasata ka praktikas olulisi mitte-empiirilisi mõõte (kaasatud on näiteks ka ergoodiliste protsesside poolt genereeritud mõõtude jadad).

Töö tulemused võib jagada kahele põhilise üldistuse tasemele. Esmalt on uuritud jaotuste lähendamist suvalise lähendhulkade klassi ja separaabli meetrilise ruumi S korral. Sellistel üldistel eeldustel õnnestus tõestada kaofunktsioonide optimaalsete väärtuste ("parimate hinnete") koondumine – üks püstitatud põhieesmärke. Toodud teoreem (koos mitmete kaasnevate tulemustega) üldistab paljusid varasemaid samalaadseid tulemusi (vt. [12], [1], [14], [18]).

Edasi on lähemalt vaadeldud kahte võrdlemisi laia lähendhulkade klassi: teatud tüüpi tõkestatud hulgad ja parameetrilised hulgad. Mõlema klassi korral said positiivse vastuse kaks ülejäänud põhiküsimust: tõestatud on optimaalsete lähendhulkade leidumine ja koondumine lõplikumõõtmelises normeeritud ruumis.

Saadud tulemused on edasiarenduseks artikli (Käärik, 2000) tulemustele, kus vaadeldi jaotuste lähendamist sfääridega, samuti on nad üldisema iseloomuga võrreldes teiste varasemate töödega selles vallas (vrdl. [12], [1], [2], [15]).

Väitekirjaga seotud teemadel on tehtud ettekanded Vilniuse Ülikooli matemaatika-informaatikateaduskonna tõenäosusteooria seminaris (detsembris 2004) ja Leedu Teaduste Akadeemia Matemaatika ja Informaatika Instituudi tõenäosusteooria seminaris (detsembris 2004), samuti konverentsil *9th International Vilnius Conference on Probability Theory and Mathematical Statistics* (juunis 2006). Lisaks väitekirjale on sel teemal avaldatud ka kolm artiklit: [5], [6], [7].

Kirjandus

1. E.F. Abaya, G.L. Wise, *Convergence of vector quantizers with applications to optimal quantization*, SIAM J. Appl. Math. **44** (1984), 183–189.

2. J. Averous, M. Meste, *Median balls: an extension of the interquantile intervals to multivariate distributions*, J. Multivariate Anal. **63** (1997), 222–241.

3. J.A. Cuesta, C. Matrán, *The strong law of large numbers for k-means and best possible nets of Banach valued random*

variables, Probab. Theory Related Fields, **78** (1988), 523–534.

4. J.A Cuesta, C. Matrán, *Uniform consistency of r-means*, Statist. Probab. Lett. **6** (1989), 65–71.

5. M. Käärik, *Approximation of distributions by sphere*, Multivariate Statistics. New Trends in Probability and Statistics, Vol. 5, VSP/TEV, Vilnius-Utrecht-Tokyo 2000, 61–66.

6. M. Käärik, K. Pärna, *Approximation of distributions by parametric sets*, Acta Appl. Math. **78** (2003), 175–183.

7. M. Käärik, K. Pärna, *Fitting parametric sets to probability distributions*, Acta Comment. Univ. Tartu. Math. **8** (2004), 101–112.

8. J. Lember, *Consistency of empirical k-centres*, Doctoral Dissertation. Tartu Ülikooli Kirjastus, Tartu 1999.

9. J. Lember, *Consistency of k-centres via metric projection*, Limit Theorems in Probability and Statistics II, Budapest 2002, 335–350.

10. J. Lember, K. Pärna, *Strong consistency of k-centres in reflexive spaces*, Probability Theory and Mathematical Statistics, VSP/TEV, Vilnius-Utrecht-Tokyo 1999, 441–452

11. Y. Nievergelt, *A tutorial history of least squares with applications to astronomy and geodesy*, J. Comput. Appl. Math., **121** (2000), 37–72.

12. D. Pollard, *Strong consistency of k-means clustering*, Ann. Statist. **9** (1981), 135–140.

13. K. Pärna, *Strong consistency of k-means clustering criterion in separable metric spaces*, Tartu Riikl. Ülik. Toimetised **733** (1986), 86–96.

14. K. Pärna, *On the stability of k-means clustering criterion in separable metric spaces*, Tartu Riikl. Ülik. Toimetised **798** (1988), 19–36.

15. K. Pärna, J. Lember, A. Viiart, *Approximating of distributions by sets*, Classification in the Information Age, Springer-Verlag 1999, Heidelberg 215–224.

16. C. Rorres, D.G. Romano, *Finding the center of a circular starting line in an ancient Greek stadium*, SIAM Review **39**, (1997), 745–754

17. H. Späth, *Least-squares fitting by circles*, Computing **57** (1996), 179–185.

18. V. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.

