Երևանի պետական համալսարան

Վիճակագրական հեռավորություններ

Հեղինակ՝ Վանուհի Բաղդասարյան

2025

Contents

Ներածություն **4**

1 Ստատիստիկական հեռավորությունների սահմանումներ և հատկություններ **5**

1.1 Կուլբակ–Լեյբլերի դիվերգենցիա 5

1.1.1 Սահմանում 5

1.1.2 Հատկություններ 5

1.2 Ընդհանուր վարիացիոն հեռավորություն 7

1.2.1 Սահմանում 7

1.2.2 Հատկություններ 7

1.3 Հելլինգերի հեռավորություն 7

1.3.1 Սահմանում 8

1.3.2 Հատկություններ 8

1.4 Յենսեն–Շենոնի դիվերգենցիա 8

1.4.1 Սահմանում 8

1.4.2 Հատկություն 9

1.5 Վասերստեյնի հեռավորություն 9

1.5.1 Սահմանում 9

2 Ստատիստիկական հեռավորությունների կիրառությունները **10**

2.1 Կուլբակ–Լեյբլերի դիվերգենցիա 10

2.2 Յենսեն-Շենոնի դիվերգենցիա 11

2.3 Վասերստեյն հեռավորություն 11

2.4 Ընդհանուր վարիացիոն հեռավորություն 12

2.5 Հելլինգերի հեռավորություն 12

3 Համեմատություն հեռավորությունների միջև և հիմնական անհավասարություններ **13**

3.1 Պինսկերի անհավասարությունը 13

3.2 Ընդհանուր վարիացիոն և Հելլինգերի հեռավորությունների կապը 13

4 Թվային փորձեր և արդյունքներ **15**

Եզրակացություն **20**

Գրականության ցանկ **21**

Ներածություն

Ստատիստիկայում և տվյալների գիտությունում կարևոր խնդիր է հավանականությ
բաշխումների միջև տարբերության քանակական չափումը:

Վիճակագրական հեռավորությունները և դիվերգենցիաները հանդիսանում
են գործիքներ, որոնք հնարավորություն են տալիս քանակապես գնահատել
բաշխումների միջև տարբերությունները՝ տարբեր զգայունությամբ և
տեսական հատկություններով: Տարբեր հեռավորություններ շեշտադրում
են բաշխումների տարբեր հատկանիշներ, ինչի պատճառով դրանց ընտրությունը
կախված է կոնկրետ խնդրի բնույթից:

Այս աշխատանքում ուսումնասիրվում են հետևյալ վիճակագրական
հեռավորությունները՝ Կուլբակ–Լեյբլերի և Յենսեն–Շենոնի դիվերգենցիաները,
ինչպես նաև Ընդհանուր վարիացիոն, Հելլինգերի և Վասերստեյնի հեռավորությունն
Ներկայացվում են դրանց սահմանումները, հիմնական հատկությունները,
փոխհարաբերությունները և կիրառությունները:

Բացի տեսական ուսումնասիրությունից, իրականացվում է նաև թվային
փորձ, որի ընթացքում տվյալների հիման վրա կառուցված էմպիրիկ
բաշխումը համեմատվում է նույն տվյալներից գնահատված տեսական
բաշխման հետ՝ տարբեր հեռավորությունների միջոցով: Քննարկվում
է նաև նմուշի (sample-ի) չափի ազդեցությունը ստացված չափումների
վրա:

Chapter 1

Ստատիստիկական հեռավորությունների սահմանումներ և հատկություններ

1.1 Կուլբակ-Լեյբլերի դիվերգենցիա

(Kullback–Leibler divergence, also called relative entropy)

1.1.1 Սահմանում

P և Q-ն Ω հավանականային տարածության վրա սահմանված հավանականային բաշխումներ են:

Կուլբակ–Լեյբլերի դիվերգենցիան P և Q բաշխումների միջև դիսկրետ դեպքում սահմանվում է հետևյալ կերպ՝

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \Omega} P(x) \, \log\left(\frac{P(x)}{Q(x)}\right) = \mathbb{E}_P\left[\log\left(\frac{P(X)}{Q(X)}\right)\right].$$

Անընդհատ դեպքում՝

$$D_{\mathrm{KL}}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \, \log\left(\frac{p(x)}{q(x)}\right) \, dx.$$

որտեղ p և q-ն՝ P, Q բաշխումներին համապատասխան խտության ֆունկցիաներն են:

1.1.2 Հատկություններ

Ոչ բացասական է

$$D_{\mathrm{KL}}(P \| Q) \geq 0,$$

որը Գիբբսի անհավասարության արդյունք է, և

$$D_{\mathrm{KL}}(P\|Q) = 0 \quad \text{այն և միայն այն դեպքում, երբ } P = Q.$$

Ապացույց (դիսկրետ դեպք) Օգտվելով Գիբբսի անհավասարությունից՝

$$-\sum_{i=1}^{n} p_i \log p_i \;\le\; -\sum_{i=1}^{n} p_i \log q_i.$$

որտեղ $P = \{p_1, \ldots, p_n\}$ և $Q = \{q_1, \ldots, q_n\}$ դիսկրետ հավանականային բաշխումներ են:

Եթե տարբերությունը դիտարկենք, կստանանք Կուլբակ–Լեյբլերի դիվերգենցիան, այսինքն Գրիբբսի անհավասարությունը կարող ենք գրել հետևյալ կերպ՝

$$D_{\mathrm{KL}}(P \,\|\, Q) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \;\ge\; 0,$$

Եվ հավասարությունը տեղի ունի այն և միայն այն դեպքում, երբ

$$p_i = q_i, \qquad i = 1, \ldots, n.$$

Ասիմետրիկություն

Ընդհանուր դեպքում

$$D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P),$$

և հետևաբար KL դիվերգենցիան չի հանդիսանում մետրիկ: Այնուամենայնիվ,

$$D_{\mathrm{KL}}(P\|Q) = 0 \quad \text{այն և միայն այն դեպքում, երբ } P = Q.$$

Ադդիտիվություն

$$D_{\mathrm{KL}}(P\|Q) = D_{\mathrm{KL}}(P_1\|Q_1) + D_{\mathrm{KL}}(P_2\|Q_2).$$

որտեղ P_1 և P_2-ը անկախ բաշխումներ են և

$$P = P_1 \times P_2,$$

ինչպես նաև Q_1 և Q_2-ը անկախ բաշխումներ են և

$$Q = Q_1 \times Q_2.$$

Ուռուցիկություն

$D_{\mathrm{KL}}(P\|Q)$-ը ուռուցիկ է հավանականային բաշխումների զույգի նկատմամբ, մասնավորապես, եթե (P_1, Q_1) և (P_2, Q_2) հավանականային բաշխումների զույգեր են, ապա կամայական $\lambda \in [0,1]$ համար տեղի ունի հետևյալ անհավասարությունը՝

$$D_{\mathrm{KL}}\big(\lambda P_1 + (1-\lambda)P_2 \,\|\, \lambda Q_1 + (1-\lambda)Q_2\big) \le \lambda D_{\mathrm{KL}}(P_1\|Q_1) + (1-\lambda) D_{\mathrm{KL}}(P_2\|Q_2), \qquad 0 \le \lambda \le 1.$$

1.2 Ընդհանուր վարիացիոն հեռավորություն

(Total variation distance)

1.2.1 Սահմանում

Ω-ն հավանականային տարածություն է և A⊂Ω պատահույթ։

Ընդհանուր վարիացիոն հեռավորությունը μ և ν հավանականային բաշխումների միջև սահմանվում է հետևյալ կերպ՝

$$d_{\mathrm{TV}}(\mu,\nu) \;:=\; \sup_{A\subset\Omega} |\mu(A)-\nu(A)|.$$

Եթե Ω-ն հաշվելի տարածություն է, ապա ընդհանուր տատանումների հեռավորությունը կստանա հետևյալ տեսքը՝

$$d_{\mathrm{TV}}(\mu,\nu) = \frac{1}{2}\sum_{x\in\Omega} |\mu(x)-\nu(x)|.$$

Այսինքն՝ ստացվում է երկու բաշխումների միջև L^1-նորմի կեսը։

1.2.2 Հատկություններ

Նույնականության պայման

$$d_{\mathrm{TV}}(\mu,\nu) = 0,$$

այն և միայն այն դեպքում, երբ $\mu = \nu$։

Սիմետրիկություն

$$d_{\mathrm{TV}}(\mu,\nu) = d_{\mathrm{TV}}(\nu,\mu).$$

Սահմանափակ է

$$0 \le d_{\mathrm{TV}}(\mu,\nu) \le 1.$$

1.3 Հելլինգերի հեռավորություն

(Hellinger distance)

1.3.1 Սահմանում

P և Q-ն հավանականային բաշխումներ են, որոնք ունեն համապատասխան խտություններ p և q: Հելլինգերի հեռավորությունը P-ի և Q-ի միջև սահմանվում է հետևյալ կերպ՝

$$H(P,Q) = \left(\int (\sqrt{p(x)} - \sqrt{q(x)})^2 \, dx \right)^{1/2}.$$

Այսինքն, Հելլինգերի հեռավորությունը հանդիսանում է $\sqrt{p}$ և $\sqrt{q}$ ֆունկցիաների միջև L^2-նորմը:

Եթե Ω-ն հաշվելի տարածություն է, ապա Հելլինգերի հեռավորությունը կստանա հետևյալ տեսքը՝

$$H(P,Q) = \left(\sum_{\omega \in \Omega} \left(\sqrt{p(\omega)} - \sqrt{q(\omega)} \right)^2 \right)^{1/2}.$$

1.3.2 Հատկություններ

Ոչ բացասական և նույնականություն

$$H(P,Q) \geq 0, \quad H(P,Q) = 0 \iff P = Q.$$

Սիմետրիկություն

$$H(P,Q) = H(Q,P).$$

Սահմանափակ է

$$0 \leq H(P,Q) \leq \sqrt{2}.$$

1.4 Յենսեն–Շենոնի դիվերգենցիա

(Jensen–Shannon divergence)

1.4.1 Սահմանում

Յենսեն–Շենոնի դիվերգենցիան սահմանվում է՝

$$\mathrm{JSD}(P,Q) = \frac{1}{2} D_{\mathrm{KL}}(P \| M) + \frac{1}{2} D_{\mathrm{KL}}(Q \| M),$$

որտեղ

$$M = \frac{1}{2}(P + Q).$$

հանդիսանում է P և Q բաշխումների խառնուրդ բաշխումը, իսկ $D(\cdot\|\cdot)$-ն Կուլբակ–Լեյբլերի դիվերգենցիան է:

1.4.2 Հատկություն

$$0 \leq \mathrm{JSD}(P, Q) \leq \log 2$$

Ջենսեն–Շենոնի դիվերգենցիան միշտ վերջավոր է, նույնիսկ այն դեպքերում, երբ Կուլբակ–Լեյբլերի դիվերգենցիան կարող է լինել անվերջ, ինչն այն դարձնում է գործնական կիրառությունների համար առավել հարմար:

1.5 Վասերստեյնի հեռավորություն

(Wasserstein or Kantorovich metric)

1.5.1 Սահմանում

Դիտենք X և Y պատահական մեծություններն են՝ համապատասխանաբար P և Q բաշխումներով: Թող $\Gamma(P, Q)$-ն լինի X և Y պատահական մեծությունների բոլոր համատեղ բաշխումների բազմությունը, որոնց մարգինալ բաշխումները համապատասխանաբար P և Q-ն են:

Այս դեպքում Վասերստեյնի առաջին կարգի հեռավորությունը սահմանվում է հետևյալ կերպ.

$$W_1(P, Q) = \inf_{\gamma \in \Gamma(P,Q)} \int_{\mathbb{R} \times \mathbb{R}} |x - y| \, d\gamma(x, y).$$

Chapter 2

Ստատիստիկական հեռավորությունների կիրառությունները

Ստատիստիկական հեռավորությունները լայնորեն կիրառվում են ինչպես տեսական ստատիստիկայում, այնպես էլ կիրառական ոլորտներում, մասնավորապես՝ մեքենայական ուսուցման մեջ: Ստորև ներկայացվում են այս աշխատանքում դիտարկված հեռավորությունների հիմնական կիրառությունները:

2.1 Կուլբակ-Լեյբլերի դիվերգենցիա

Կուլբակ-Լեյբլերի դիվերգենցիան հանդիսանում է ինֆորմացիոն տեսության հիմնարար գործիք և լայնորեն կիրառվում է ինչպես ստատիստիկայում, այնպես էլ մեքենայական ուսուցման մեջ:

Կիրառությունները.

- **Մոդելի համապատասխանության գնահատում (Goodness-of-fit)** Կուլբակ-Լեյբլերի դիվերգենցիան չափում է, թե որքան տեղեկատվությունն է կորցվում, երբ իրական բաշխումը մոտարկվում է տեսական մոդելով: Այդ պատճառով այն հաճախ օգտագործվում է մոդելի որակը գնահատելու համար:

- **Bayesian inference և Variational մեթոդներ** Կուլբակ-Լեյբլերի դիվերգենցիան հանդիսանում է Variational Inference-ի և Variational Autoencoder-ների (VAE) հիմնական օպտիմալացման չափանիշը:

- **Ինֆորմացիոն շահույթ (Information gain)** Կուլբակ-Լեյբլերի

դիվերգենցիան օգտագործվում է գնահատելու համար, թե որքան նոր տեղեկատվություն է ստացվում մի բաշխումից մյուսին անցնելիս:

2.2 Յենսեն-Շենոնի դիվերգենցիա

Յենսեն-Շենոնի դիվերգենցիան Կուլբակ-Լեյբլերի դիվերգենցիայի սիմետրիկ և հարթեցված տարբերակն է, ինչը այն դարձնում է ավելի կայուն և գործնական:

Կիրառությունները.

- **Generative Adversarial Networks (GAN)** Դասական GAN-երի տեսական հիմքում ընկած է հենց Յենսեն-Շենոնի դիվերգենցիան, որը չափում է իրական և գեներացված տվյալների բաշխումների տարբերությունը:
- **Բաշխումների նմանության չափում** Քանի որ Յենսեն-Շենոնի դիվերգենցիան սիմետրիկ է և միշտ վերջավոր, այն հաճախ օգտագործվում է երկու բաշխումների նմանությունը գնահատելու համար՝ NLP-ում, տեքստերի և թեմատիկ մոդելների համեմատության մեջ:
- **Կայուն այլընտրանք Կուլբակ-Լեյբլերիին** Այն կիրառվում է այն իրավիճակներում, որտեղ Կուլբակ-Լեյբլերի դիվերգենցիան կարող է լինել անվերջ կամ թվային անկայուն:

2.3 Վասերստեյն հեռավորություն

Վասերստեյնի հեռավորությունը հիմնված է օպտիմալ տրանսպորտի տեսության վրա և էապես տարբերվում է Կուլբակ-Լեյբլերի և Յենսեն-Շենոնի դիվերգենցիաներից:

Կիրառությունները.

- **Optimal Transport տեսություն** Վասերստեյնի հեռավորությունը հանդիսանում է օպտիմալ տրանսպորտի հիմնական չափանիշը՝ չափելով բաշխումների միջև նվազագույն տեղափոխման ծախսը:
- **Wasserstein GAN (WGAN)** Վասերստեյնի հեռավորության կիրառումը WGAN-երում ապահովում է ավելի կայուն ուսուցում և թույլ է տալիս խուսափել mode collapse երևույթից:

- **Երկրաչափական համեմատություն** Քանի որ Վասերստեյնի հեռավորությունը հաշվի է առնում արժեքների իրական թվային հեռավորությունը, այն լայնորեն կիրառվում է պատկերների, սիգնալների և շարունակական տվյալների բաշխումների համեմատության մեջ:

2.4 Ընդհանուր վարիացիոն հեռավորություն

Ընդհանուր վարիացիոն հեռավորությունը հանդիսանում է բաշխումների տարբերության ուժեղ չափանիշ:

Կիրառությունները.

- **Տեսական ստատիստիկա և սահմաններ** Ընդհանուր վարիացիոն հեռավորությունը լայնորեն կիրառվում է տեսական արդյունքների և անհավասարությունների ձևակերպման մեջ:
- **Կոնվերգենցիայի ուսումնասիրություն** Այն օգտագործվում է Մարկովյան շղթաների և պատահական պրոցեսների կոնվերգենցիան ուսումնասիրելիս:
- **Goodness-of-fit թեստեր** Ընդհանուր վարիացիոն հեռավորությունը կիրառվում է պարզ և ինտուիտիվ համեմատություններում՝ բաշխումների տարբերությունը գնահատելու համար:

2.5 Հելլինգերի հեռավորություն

Հելլինգերի հեռավորությունը համարվում է կայուն և մաթեմատիկորեն հարմար չափանիշ:

Կիրառությունները.

- **Robust statistics** Այն ավելի կայուն է աղմուկի և արտահայտված պոչերի նկատմամբ, քան Կուլբակ-Լեյբլերի դիվերգենցիան:
- **Պարամետրերի գնահատում** Օգտագործվում է գնահատիչների որակը և կայունությունը վերլուծելիս:
- **Տեսական կապեր** Հելլինգերի հեռավորությունը հաճախ օգտագործվում է այլ հեռավորությունների համար սահմաններ և հարաբերություններ ստանալու նպատակով:

Chapter 3

Համեմատություն հեռավորությունների միջև և հիմնական անհավասարություններ

3.1 Պինսկերի անհավասարությունը

(Pinsker's inequality)

Pinsker-ի անհավասարությունը կապում է Կուլբակ–Լեյբլերի դիվերգենցիան և Ընդհանուր վարիացիոն հեռավորությունը՝

$$d_{\mathrm{TV}}(P,Q) \le \sqrt{\frac{1}{2}\, D_{\mathrm{KL}}(P\|Q)}.$$

Pinsker-ի անհավասարությունը ցույց է տալիս, որ Կուլբակ–Լեյբլերի դիվերգենցիան ավելի ուժեղ չափում է, քան Ընդհանուր վարիացիոնը։

Այսինքն՝ եթե $D_{\mathrm{KL}}(P\|Q)$ փոքր է, ապա P և Q բաշխումները պարտադիր մոտ են նաև Ընդհանուր վարիացիոն իմաստով, բայց հակառակը պարտադիր չէ՝ Ընդհանուր վարիացիոն հեռավորությունը կարող է փոքր լինել, իսկ Կուլբակ–Լեյբլերի դիվերգենցիան՝ մեծ կամ նույնիսկ անվերջ։

3.2 Ընդհանուր վարիացիոն և Հելլինգերի հեռավորությունների կապը

Ընդհանուր վարիացիոն և Հելլինգերի հեռավորությունների միջև գործում են հետևյալ անհավասարությունները.

$$d_{\mathrm{TV}}(P,Q) \;\le\; H(P,Q) \;\le\; \sqrt{2\, d_{\mathrm{TV}}(P,Q)}.$$

Ընդհանուր վարիացիոն հեռավորությունը չափում է առավելագույն տարբերությունը երկու բաշխումների միջև՝ դիտարկելով բոլոր հնարավոր իրադարձությունները։

Հելլինգերի հեռավորությունը համեմատում է $\sqrt{p}$ և $\sqrt{q}$ ֆունկցիաները, ինչի արդյունքում այն հաճախ ավելի թվային կայուն չափում է։

Chapter 4

Թվային փորձեր և արդյունքներ

Տվյալների նկարագրություն

Վերցնենք

$$X_1, \ldots, X_{5000}$$

i.i.d. պատահական փոփոխականներ, որոնք գեներացվել են նորմալ բաշխումից՝

$$X_i \sim \mathcal{N}(\mu, \sigma^2),$$

որտեղ տվյալ փորձի համար ընտրված է

$$\mu = 0, \qquad \sigma^2 = 1.$$

Սինթետիկ տվյալների օգտագործումը թույլ է տալիս ուսումնասիրել վիճակագրական հեռավորությունների վարքը, երբ իրական բաշխումը հայտնի է:

Էմպիրիկ բաշխման կառուցում

Դիտարկված նմուշի(sample)

$$X_1, \ldots, X_{5000}$$

հիման վրա կառուցվում է էմպիրիկ բաշխումը, որը նկարագրում է տվյալների փաստացի բաշխումը՝ հիմնված դիտարկված արժեքների հաճախականությունների վրա:

Էմպիրիկ բաշխումը չի ենթադրում որևէ պարամետրային տեսք և կապված չէ որևէ կոնկրետ տեսական բաշխման հետ: Այն արտացոլում է միայն այն, ինչ իրականում ցույց են տալիս տվյալները:

Տեսական բաշխման մոտարկում տվյալներից

Տվյալների համար ընտրվում է տեսական բաշխումների ընտանիք, մասնավորապես՝ նորմալ բաշխումը:

Բաշխման պարամետրերը գնահատվում են տվյալներից՝

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2.$$

Բաշխումների միջև հեռավորության չափում

Այսպիսով, ստացվում են երկու տարբեր բաշխումներ.

- էմպիրիկ բաշխում՝ կառուցված անմիջապես տվյալներից,
- տեսական բաշխում՝ ընտրված բաշխումը գնահատված պարամետրերով:

Նպատակն է չափել այս երկու բաշխումների միջև հեռավորությունը: Կիրառվում են հետևյալ հեռավորությունները՝ ընդհանուր վարիացիոն հեռավորություն, Հելլինգերի հեռավորություն, Կուլբակ–Լեյբլերի դիվերգենցիա, Յենսեն–Շենոնի դիվերգենցիա և Վասերստեյնի հեռավորություն:

Արդյունքներ

Հեռավորությունների արժեքները ամբողջ նմուշի համար

$n = 5000$ նմուշի դեպքում հաշվարկվել են էմպիրիկ և տեսական բաշխումների միջև հեռավորությունների արժեքները: Ստացված արդյունքները ներկայացված են Աղյուսակ 4.1-ում:

Հեռավորություն	Արժեք
Total Variation	0.032667
Hellinger	0.030744
Jensen–Shannon	0.000945
KL divergence	0.003770
Wasserstein	0.022250

Table 4.1: Բաշխումների միջև հեռավորությունները ($n = 5000$)

Բոլոր դիտարկված հեռավորությունները փոքր են, ինչը ցույց է տալիս, որ էմպիրիկ բաշխումը մոտ է տվյալներից ստացված տեսական նորմալ բաշխմանը:

Bootstrap արդյունքներ (կայունություն)

Հեռավորությունների արդյունքների կայունությունը գնահատելու համար կիրառվել է ոչ պարամետրիկ bootstrap մեթոդ: Յուրաքանչյուր հեռավորության համար հաշվարկվել են միջին արժեքները և 95% վստահության միջակայքները: Արդյունքները ներկայացված են Աղյուսակ 4.2-ում:

Հեռավորություն	Median	CI
Total Variation	0.042900	(0.032330, 0.053937)
Hellinger	0.038334	(0.029891, 0.048465)
Jensen–Shannon	0.001468	(0.000893, 0.002346)
KL divergence	0.005827	(0.003545, 0.009462)
Wasserstein	0.022250	(0.015160, 0.035541)

Table 4.2: Bootstrap 95% վստահության միջակայքներ

Ստացված վստահության միջակայքերը հարաբերականորեն մոտ են, ինչը ցույց է տալիս, որ հեռավորությունների գնահատումները կայուն են և էականորեն կախված չեն տվյալների փոքր փոփոխությունից:

Նմուշի չափի ազդեցությունը (գրաֆիկներ)

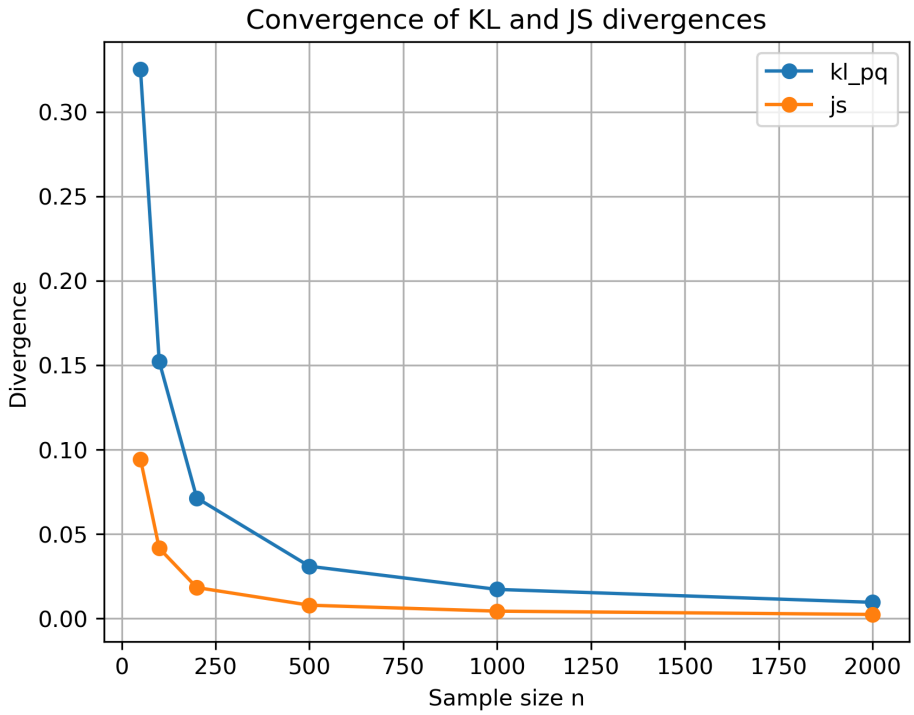

Figure 4.1: Կուլբակ–Լեյբլերի և Յենսեն–Շենոնի դիվերգենցիաների փոփոխությունը նմուշի չափի մեծացման հետ

Գրաֆիկ 4.1-ում ներկայացված է Կուլբակ–Լեյբլերի և Յենսեն–Շենոնի դիվերգենցիաների փոփոխությունը նմուշի չափի մեծացման հետ: Նկատենք, որ նմուշի չափի մեծացման դեպքում երկու դիվերգենցիաներն էլ արագ նվազում են:

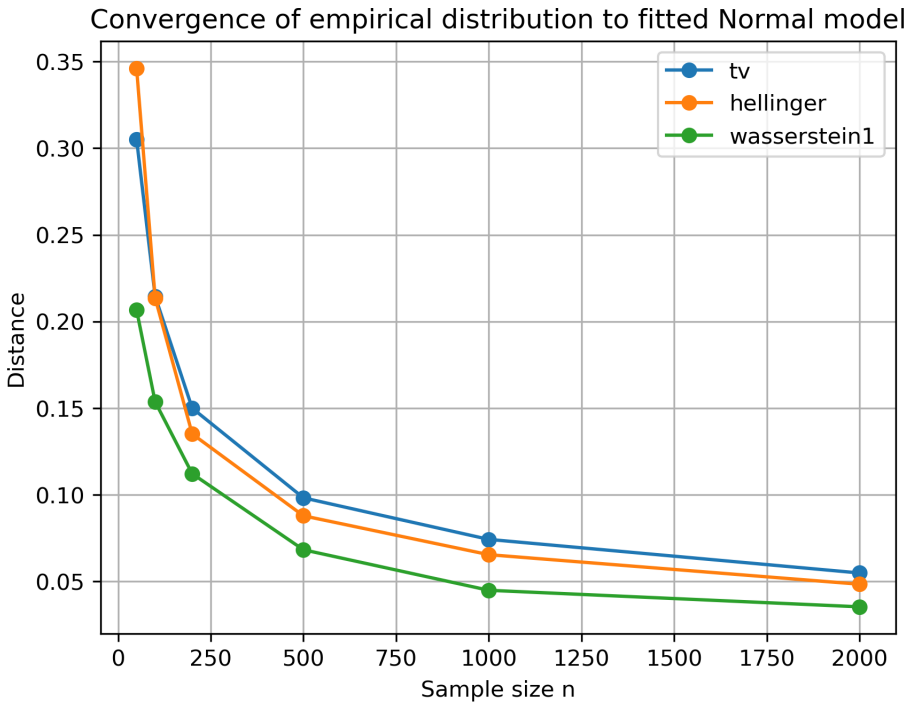

Figure 4.2: Ընդհանուր վարիացիոն, Հելլինգերի և Վասսերստեյնի հեռավորությունների վարքը տարբեր նմուշի չափերի դեպքում

Գրաֆիկ 4.2-ում ներկայացված է Ընդհանուր վարիացիոն, Հելլինգերի և Վասսերստեյնի հեռավորությունների վարքը տարբեր նմուշի չափերի դեպքում: Բոլոր երեք հեռավորությունները նվազում են նմուշի չափի մեծացման հետ:

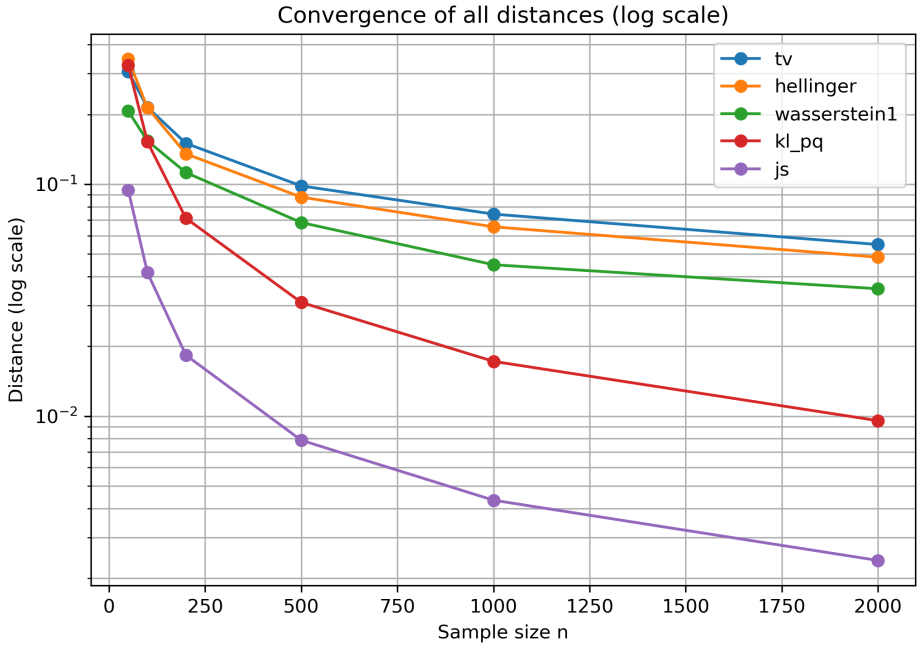

Figure 4.3: Բոլոր դիտարկված հեռավորությունները լոգարիթմական մասշտաբով

Գրաֆիկ 4.3-ում ներկայացված են բոլոր դիտարկված հեռավորությունները լոգարիթմական մասշտաբով, ինչը թույլ է տալիս համեմատել դրանց մեծությունները և նվազման արագությունները նույն գրաֆիկում:

Եզրակացություն

Այս աշխատանքում իրականացվել է տվյալների հիման վրա կառուցված էմպիրիկ բաշխման և նույն տվյալներից ստացված տեսական բաշխման համեմատություն՝ կիրառելով տարբեր ստատիստիկական հեռավորություններ: Քննարկվել են մի շարք հեռավորություններ և դիվերգենցիաներ, ներկայացվել են դրանց սահմանումները, հիմնական հատկությունները և կիրառությունները:

Թվային փորձի արդյունքները ցույց են տալիս, որ նմուշի չափի մեծացման հետ էմպիրիկ բաշխումը աստիճանաբար մոտենում է տվյալներից ստացված տեսական բաշխմանը: Սա արտահայտվում է բոլոր դիտարկված հեռավորություններ արժեքների նվազմամբ, ինչը համապատասխանում է ստատիստիկայի տեսական ակնկալիքներին:

Միևնույն ժամանակ նկատվում է, որ տարբեր հեռավորությունները տարբեր կերպ են արձագանքում նմուշի չափի փոփոխությանը: Մասնավորապես, Կուլբակ–Լեյբլերի և Յենսեն–Շենոնի դիվերգենցիաները նվազում են ավելի արագ, ինչը վկայում է դրանց բարձր զգայունության մասին բաշխումների տեղային տարբերությունների նկատմամբ:

Այս արդյունքները ցույց են տալիս, որ ստատիստիկական հեռավորությունների ընտրությունը կարևոր է և կախված է կոնկրետ խնդրից: Տարբեր հեռավորություններ ընդգծում են բաշխումների տարբեր հատկություններ և դրանց համատեղ կիրառումը թույլ է տալիս ավելի ամբողջական պատկերացում կազմել մոդելի համապատասխանության և բաշխումների միջև տարբերությունների մասին:

Գրականության ցանկ

[GS02] Alison L. Gibbs and Francis E. Su. “On Choosing and Bounding Probability Metrics”. In: *International Statistical Review* (2002).

[Was04] Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2004.

[CL20] Yuhang Cai and Lek-Heng Lim. “Distances between probability distributions of different dimensions”. In: *Journal of Multivariate Analysis* (2020).

[Wik24a] Wikipedia. *Gibbs’ inequality*. 2024. URL: https://en.wikipedia.org/wiki/Gibbs%27_inequality.

[Wik24b] Wikipedia. *Hellinger distance*. 2024. URL: https://en.wikipedia.org/wiki/Hellinger_distance.

[Wik24c] Wikipedia. *Jensen–Shannon divergence*. 2024. URL: https://en.wikipedia.org/wiki/Jensen%E2%80%93Shannon_divergence.

[Wik24d] Wikipedia. *Kullback–Leibler divergence*. 2024. URL: https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence.

[Wik24e] Wikipedia. *Total variation distance of probability measures*. 2024. URL: https://en.wikipedia.org/wiki/Total_variation_distance_of_probability_measures.

[Wik24f] Wikipedia. *Wasserstein metric*. 2024. URL: https://en.wikipedia.org/wiki/Wasserstein_metric.