Daugiamatės skalės (Multidimensional scaling)

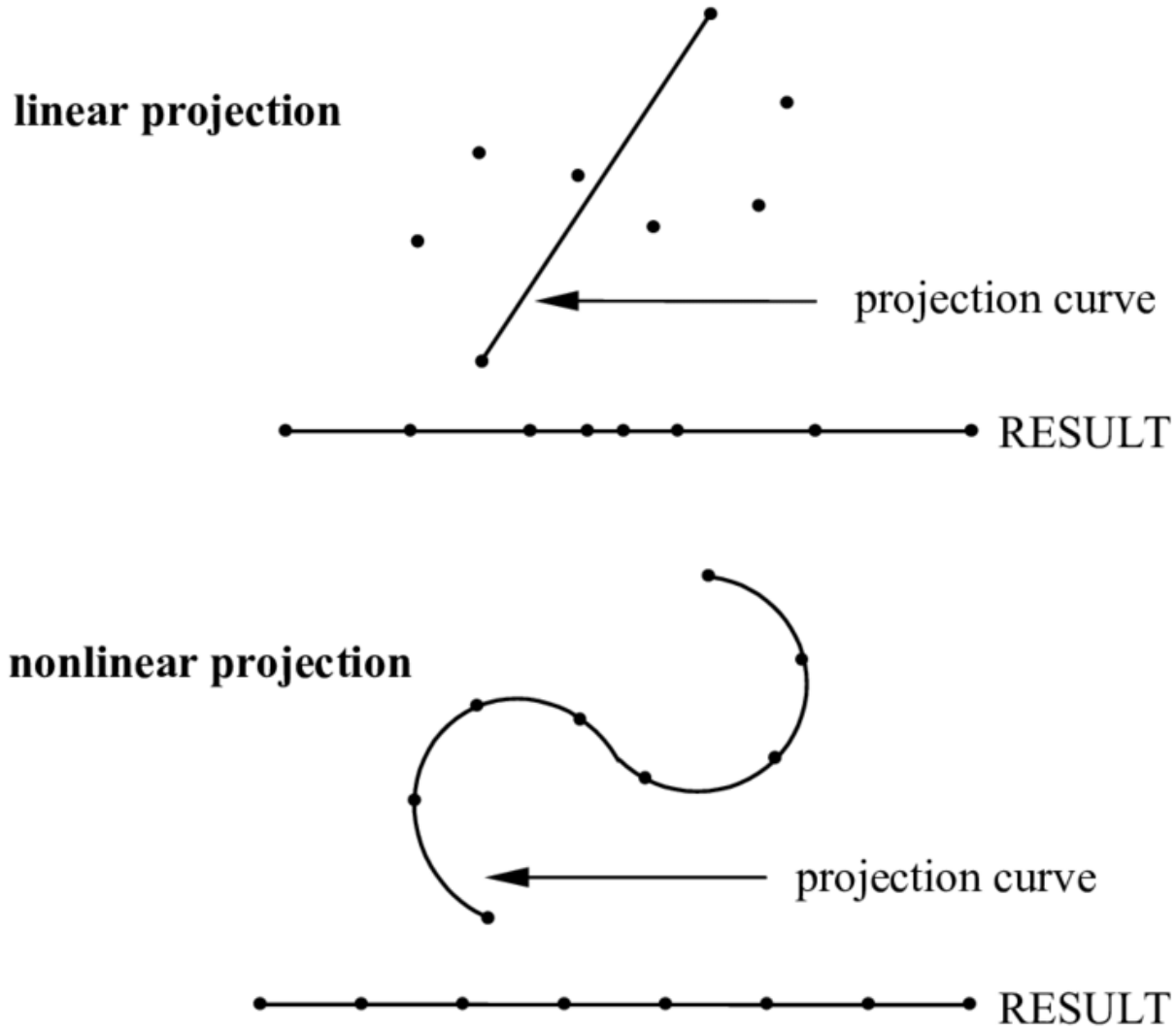
Matas Gaulia, Vainius Gataveckas, Dovydas Martinkus

Duomenų Mokslas 3 kursas 2 gr.

Vilnius, 2022

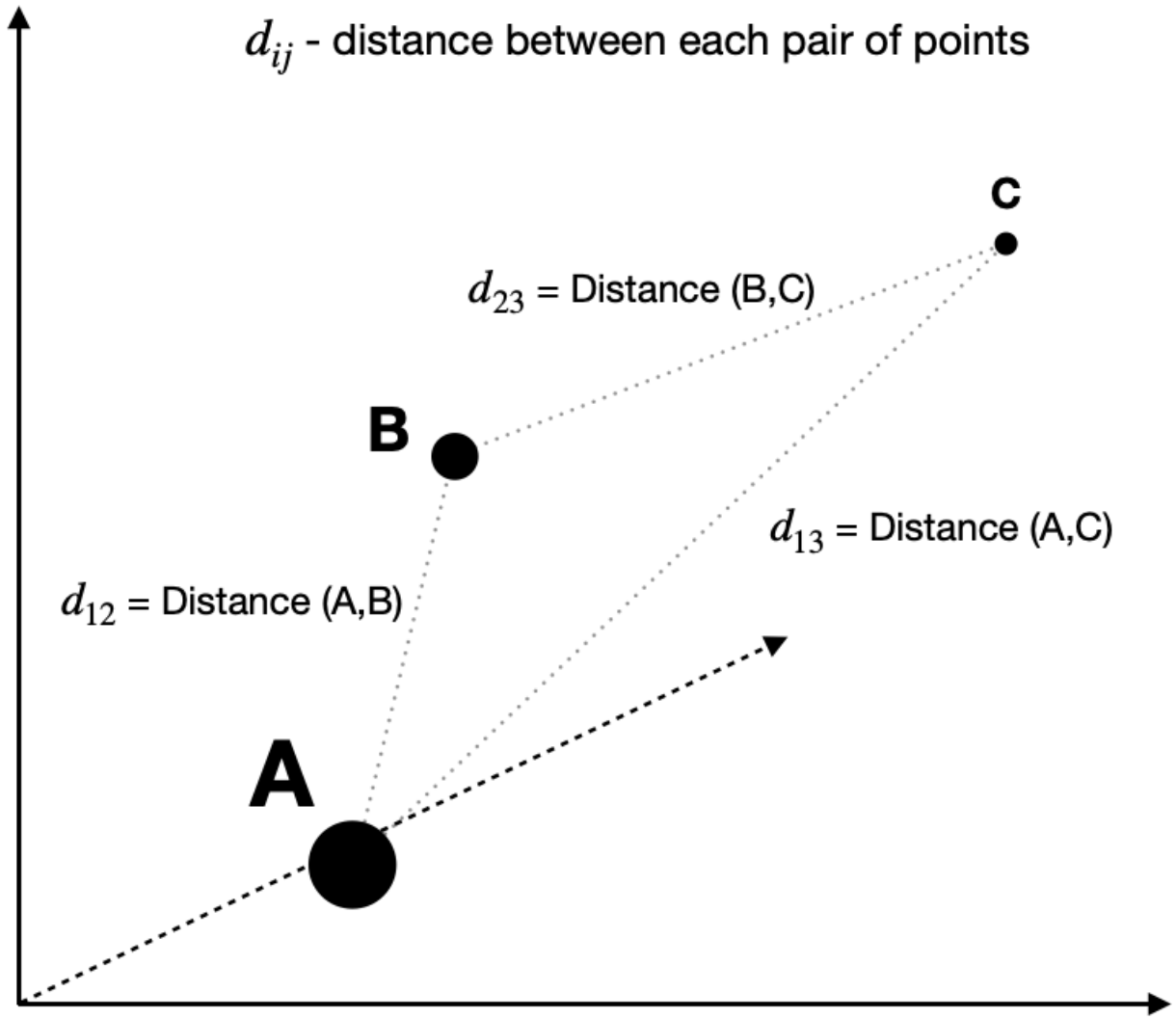
Tiesiniai ir netiesiniai dimensijos mažinimo metodai

- Tiesinės transformacijos: pasukimas, postūmis, atspindys, suspaudimas.
- Dimensijos mažinimas pagrįstas tiesinėmis transformacijomis neišlaiko netiesinių sąryšių tarp objektų.
- Daugiamatės skalės (angl. Multidimensional Scaling, toliau - MDS) yra netiesinis dimensijos mažinimo metodas.

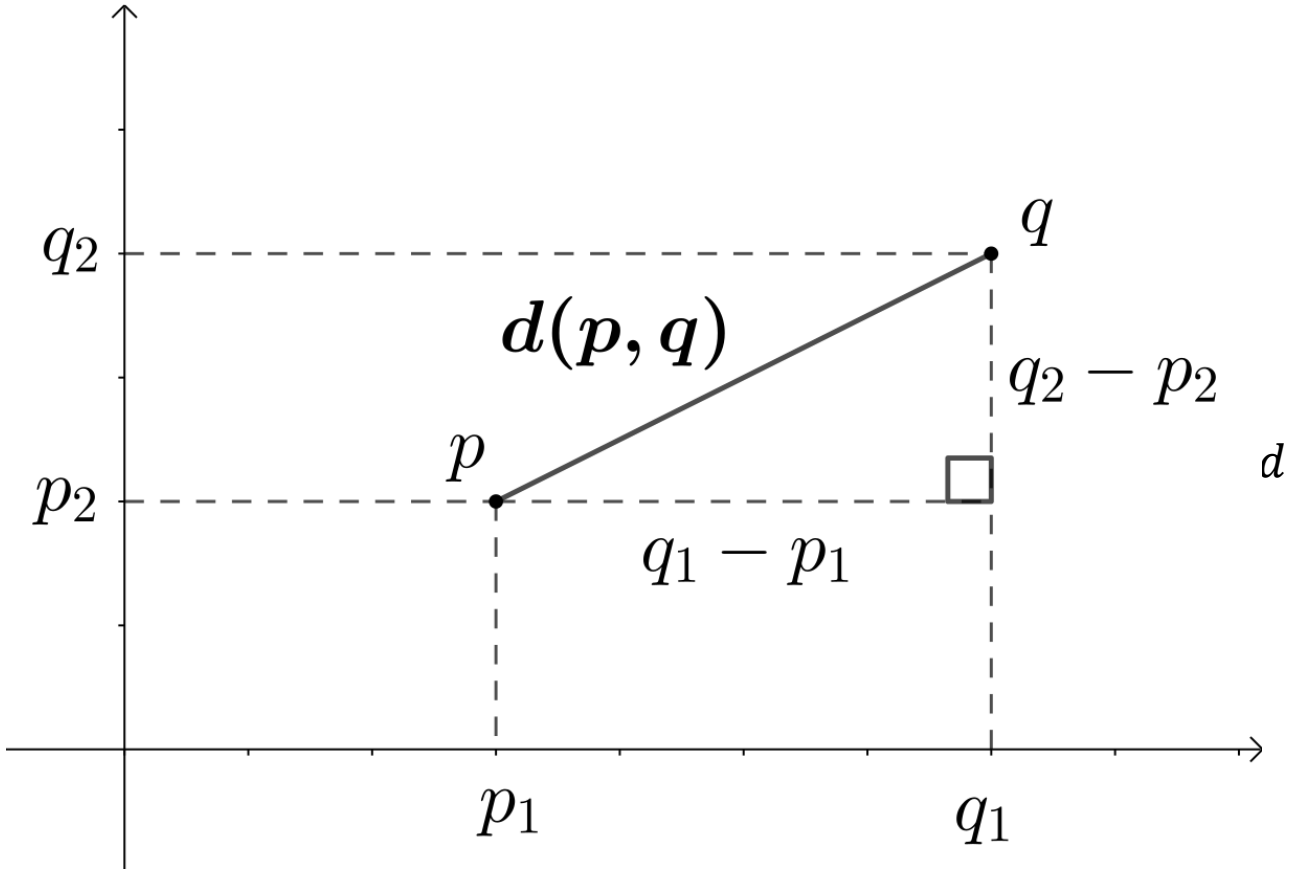

linear projection
projection curve
RESULT
nonlinear projection
projection curve
RESULT

Daugiamatės skalės

- Daugiamatės skalės kiekvieną objektą iš didesnės dimensijos transformuoja į iš anksto parinkto mažesnio dydžio dimensiją.
- Naudojant MDS ieškoma daugiamačių duomenų projekcijų mažesnės dimensijos erdvėje, siekiant išlaikyti atstumus tarp objektų.


d_{ij} - distance between each pair of points
C
d_{23} = Distance (B,C)
B
d_{13} = Distance (A,C)
d_{12} = Distance (A,B)
A

Atstumai



Įprastai naudojamas Euklidinis atstumas:

$$d(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

Kiti galimi atstumai

- $d_1(X_k, X_l) = \sum_{j=1}^{n} \left| x_{kj} - x_{lj} \right|$ Manheteno atstumas
- $d_{\infty(X_k, X_l)} = \max_j \left| x_{kj} - x_{lj} \right|$ Čebyševo atstumas
- $d_{(X_k, X_l)} = \sum_{i=1}^{n} \frac{|x_{kj} - x_{lj}|}{|x_{kl}| + |x_{lj}|}$ Kanberos atstumas

Nepanašumai

- Bendru atveju naudojami nepanašumai (angl. dissimilarities). Tai atstumai kuriems nebūtinai galioja trikampio taisyklė $d(x_i, x_j) \leq d(x_i, x_k) + d(x_k, x_j)$.

- Tokie matavimai dažnai naudojami sociologijoje, psichologijoje, kitose srityse. Pvz. kažkokie subjektyvūs įvertinimai.

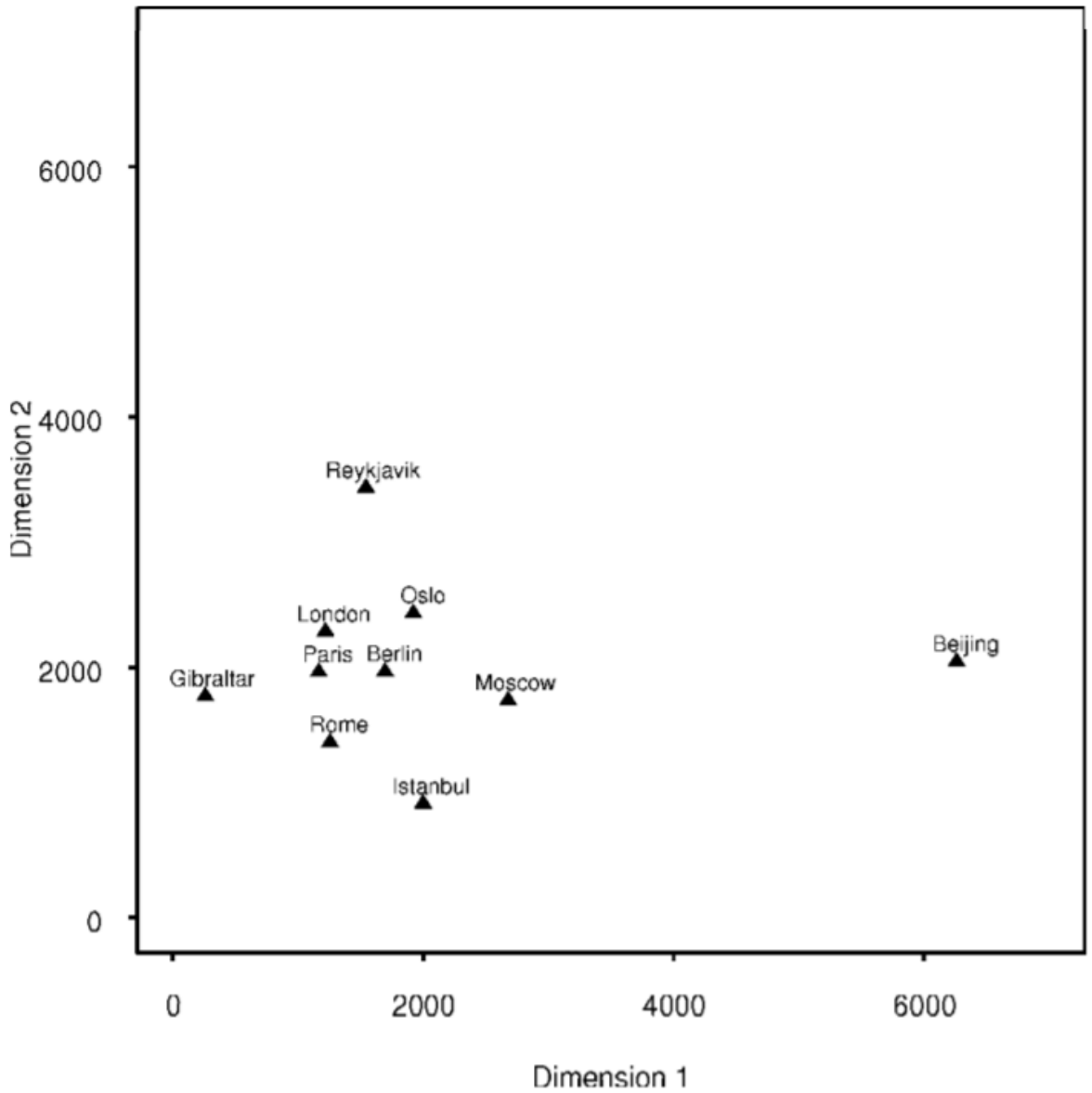
Nepanašumų matrica

- Tarkime duomenyse turime m objektų.
- $D_{ij} = d(X_i, X_j)$ – kaip nors apibrėžtas nepanašumas tarp i-tojo ir j-tojo objektų.

- Tada turime nepanašumų matricą (dissimilarity matrix):

$$D = \begin{pmatrix} \mathrm{D}_{11} & \mathrm{D}_{12} & \dots & \mathrm{D}_{1m} \\ \mathrm{D}_{21} & \mathrm{D}_{22} & \dots & \mathrm{D}_{2m} \\ \dots & \dots & & \dots \\ \mathrm{D}_{m1} & \mathrm{D}_{m2} & \dots & \mathrm{D}_{mm} \end{pmatrix}$$

	London	Berlin	Oslo	Moscow
London	–			
Berlin	570	–		
Oslo	710	520	–	
Moscow	1550	1000	1020	–
Paris	210	540	830	1540
Rome	890	730	1240	1470
Beijing	5050	4570	4360	3600
Istanbul	1550	1080	1520	1090
Gibraltar	1090	1450	1790	2410
Reykjavik	1170	1480	1080	2060



Disparities

- Praktikoje vietoje nepanašumų pradinėje dimensijoje D_{ij} apskaičiuojamos tam tikros transformacijos $\widehat{D_{ij}}$ vadinamos disparities. Jos atitinka “idealius” atstumus mažesnės dimensijos erdvėje.

- MDS siekia minimizuoti mažiausių kvadratų įterpimo funkciją (dažniausiai ji vadinama tiesiog Stress):

  $\sum_{i<j}\left(d_{ij} - \widehat{D_{ij}}\right)^2$, kur d_{ij} tuo metu turimas Euklidinis atstumas tarp objektų mažesnės dimensijos erdvėje.

- Šiuo atveju atsiminimui: mažoji raidė „d“ atitinka mažesnės erdvės dimensiją.

Optimizavimas

- Dimensijos mažinimas naudojant MDS yra optimizavimo procesas.

- Naudojamas iteratyvus algoritmas, kuris minimizuoja Stress funkciją.
- Pvz. scikit-learn naudojamas SMACOF algoritmas.

Metrikinė ir nemetrikinė MDS

MDS gali būti:

- Metrikinė (angl. metric)
- Nemetrikinė (angl. non-metric)

Metrikinė MDS

Metrikinėje MDS nepanašumų matrica gaunama iš metrikos (galioja trikampio nelygybė), todėl žemesnės dimensijos erdvėje siekiama, kad atstumai tarp taškų būtų kuo panašesni į atstumus pradinėje erdvėje.

Metrikiniu atveju naudojami disparities gavimo būdai:

- $\widehat{D_{ij}} = D_{ij}$
- $\widehat{D_{ij}} = bD_{ij}$ (ratio MDS)
- $\widehat{D_{ij}} = a + bD_{ij}$ (interval MDS)

Nemetrikinė MDS

- Nemetrikinėje versijoje MDS siekiama, kad atstumų tvarka mažesnės dimensijos erdvėje sutaptų su nepanašumų tvarka pradinėje erdvėje.

- Matematiškai tai reiškia, kad jeigu $D_{ij} < D_{jk}$ originalios dimensijos erdvėje, tai $d_{ij} < d_{jk}$ mažesnės dimensijos erdvėje

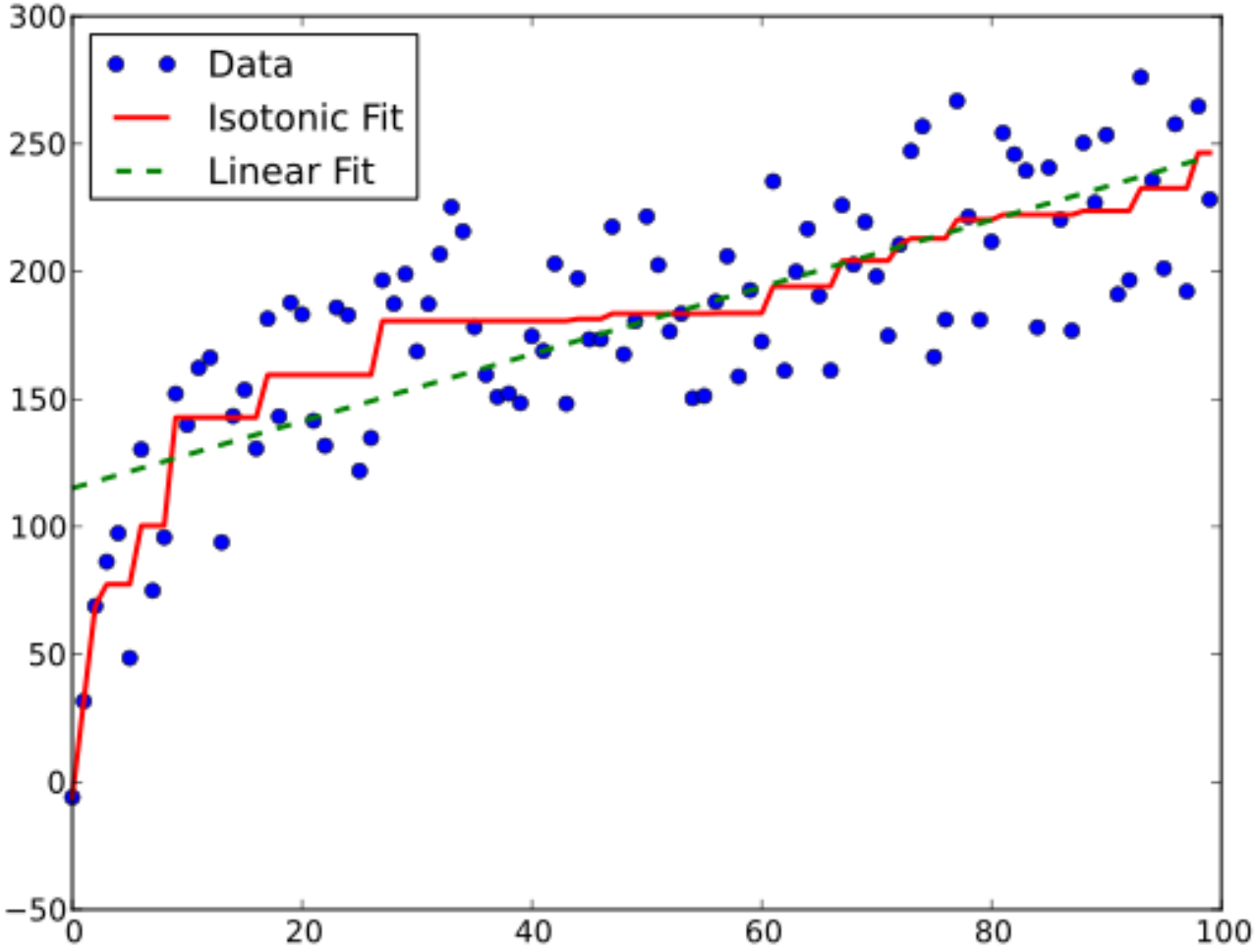
Monotoninė regresija

- Paprastas algoritmas užtikrinti šį sąryšį yra monotoninė regresija.
- Monotoninėje regresijoje regresijos kreivė yra nemažėjanti arba nedidėjanti.

- Atliekama monotoninė regresija su prediktoriumi D_{ij} ir atsaku d_{ij}.
- Tada taškai ant monotoninės regresijos kreivės (fitted values) yra $\widehat{D_{ij}}$.

Pvz. $\widehat{D_{ij}}$ taškai yra ant raudonos spalvos linijos.

Siekiama, kad kitoje iteracijoje monotoninės regresijos kreivė būtų labiau tolygiai „laiptuota“.



Bendra MDS schema

- Bendra MDS schema atrodo taip:

    1. Pradinis taškų išsidėstymas.
    2. Apskaičiuojamas Stress.
    3. Stress minimizavimas tam tikru algoritmu.
    4. 2 ir 3 žingsnio kartojimas iki konvergavimo.

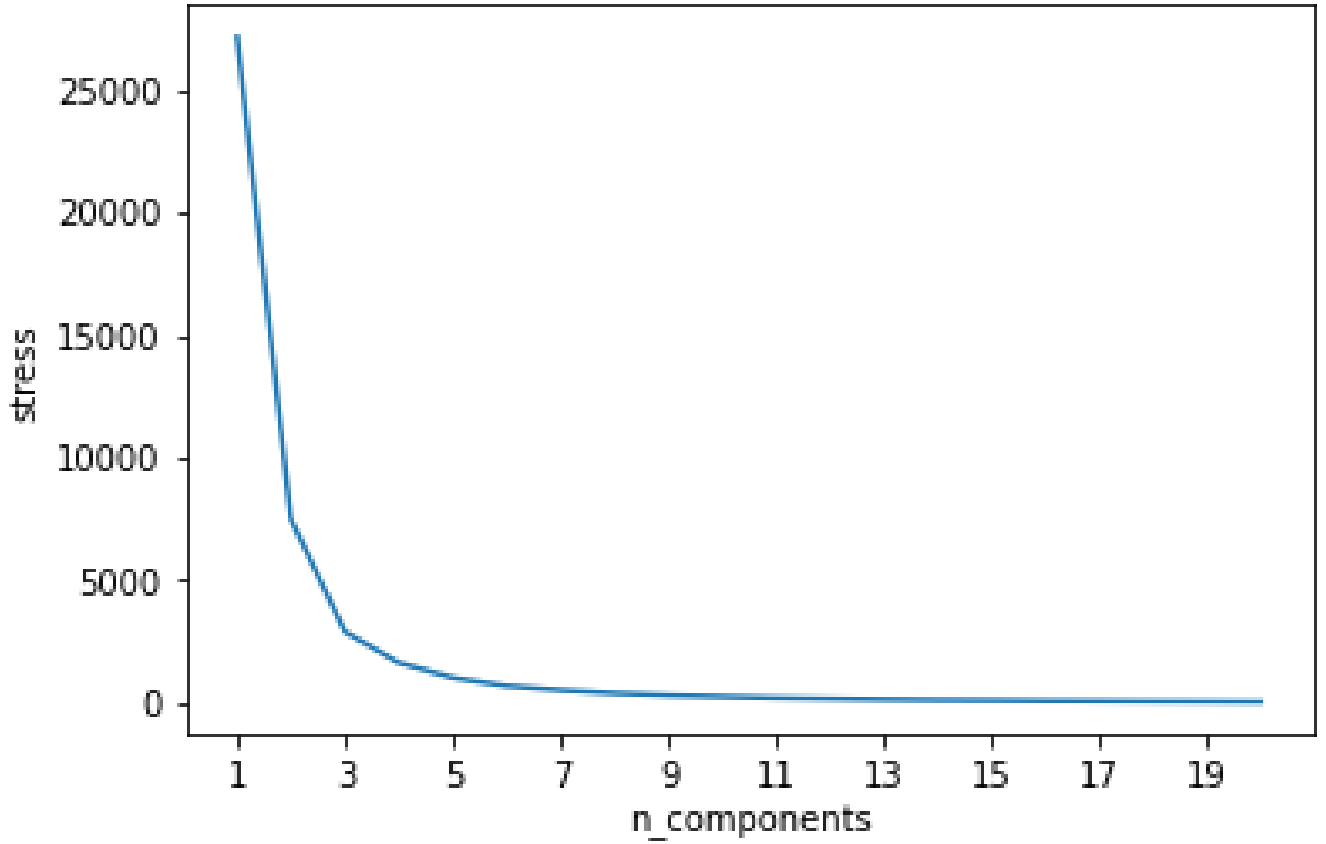
- Kiekvieną kartą prieš apskaičiuojant Stress iš naujo apskaičiuojamos disparities (nemetrikiniu atveju pakartotinai atliekama monotoninė regresija).

- Iš praktinės pusės tai reiškia kad:
  - Reikia pasirinkti iteracijų skaičių.
  - Jeigu pradinės ieškomų vektorių reikšmės atsitiktinės, tai kiekvieną kartą gali būti randamas kitas sprendimas.
  - Algoritmas gali užstrigti lokaliame minimume (siekiant to išvengti algoritmas paleidžiamas kelis kartus ir pasirenkamas geriausias spendimas).
  - Pasirinkama, kada deklaruojamas konvergavimas.

MDS tinkamumo įvertinimas

- Norima dimensija turi būti parenkama iš anksto.
- Natūralu, kad Stress reikmė didėja kuo labiau mažinama dimensija.
- Įprastai MDS dimensijų skaičius randamas ieškant mažiausios dimensijos, kuri vis dar turi pakankamai mažas Stress reikšmes.

Scree plot ieškoma alkūnės taško (angl. elbow point)



Standartizuotas Stress

- Grynos Stress reikšmes nėra informatyvios (pvz. gaunamos didesnės tiesiog papildžius duomenų aibę).
- Informatyvesnis Kruskal‘s Stress arba kitaip Stress-1 (pavadinimas, o ne reiškinys).

$$Stress - 1 = \sqrt{\frac{d_{ij} - \widehat{D_{ij}}}{\sum d_{ij}^2}}$$

- Reikšmės nuo 0 iki 1, todėl galima kalbėti apie goodness-of-fit nykščio taisyklę:

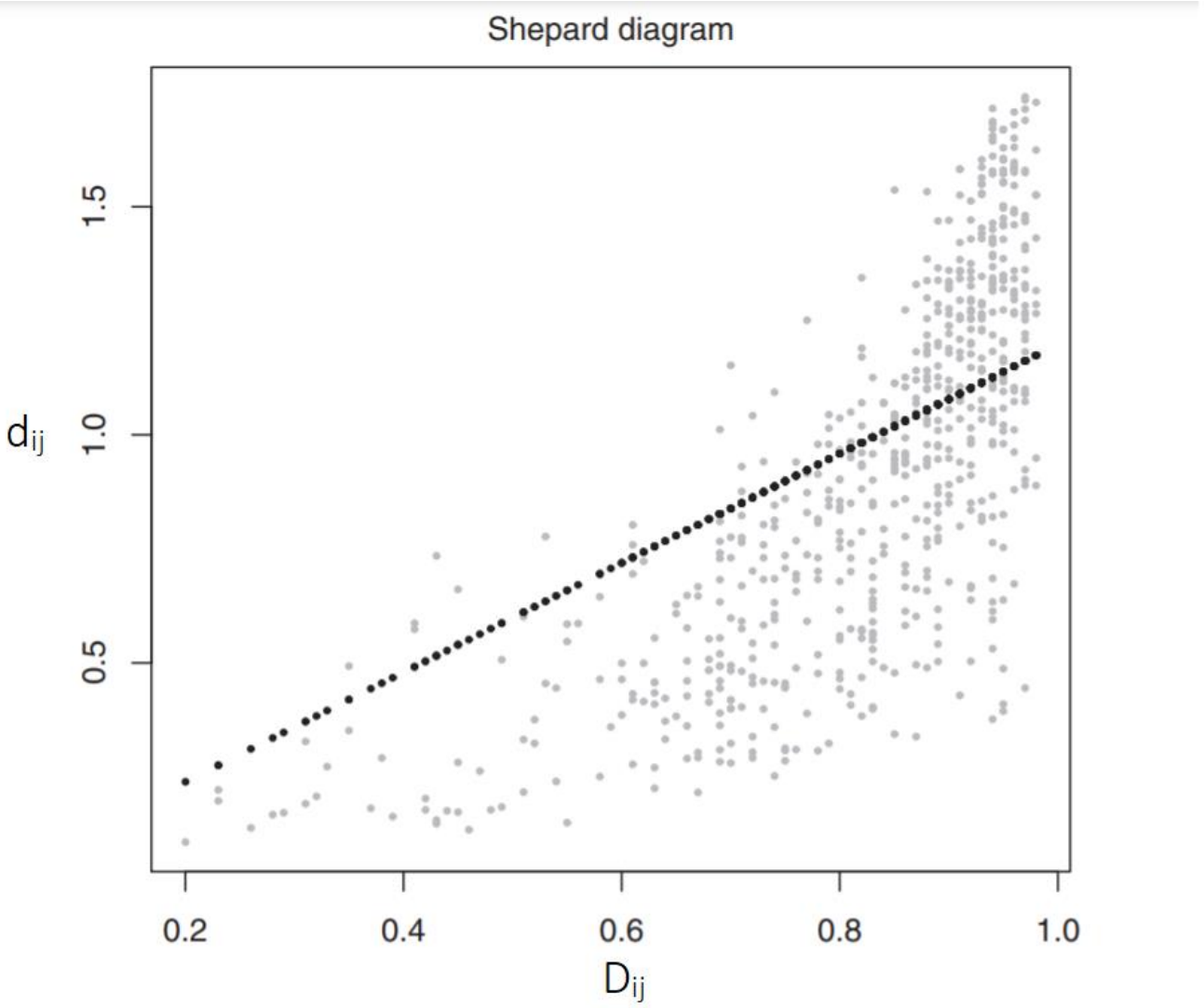
0.200	Blogas
0.100	Vidutinis
0.050	Geras
0.025	Puikus
0.000	Tobulas

Diagnostiniai grafikai pagrįsti d_{ij}, D_{ij}, $\widehat{D_{ij}}$ tarpusavio ryšio vaizdavimu.

Tarp jų dažniausiai naudojamas Shepard diagram.

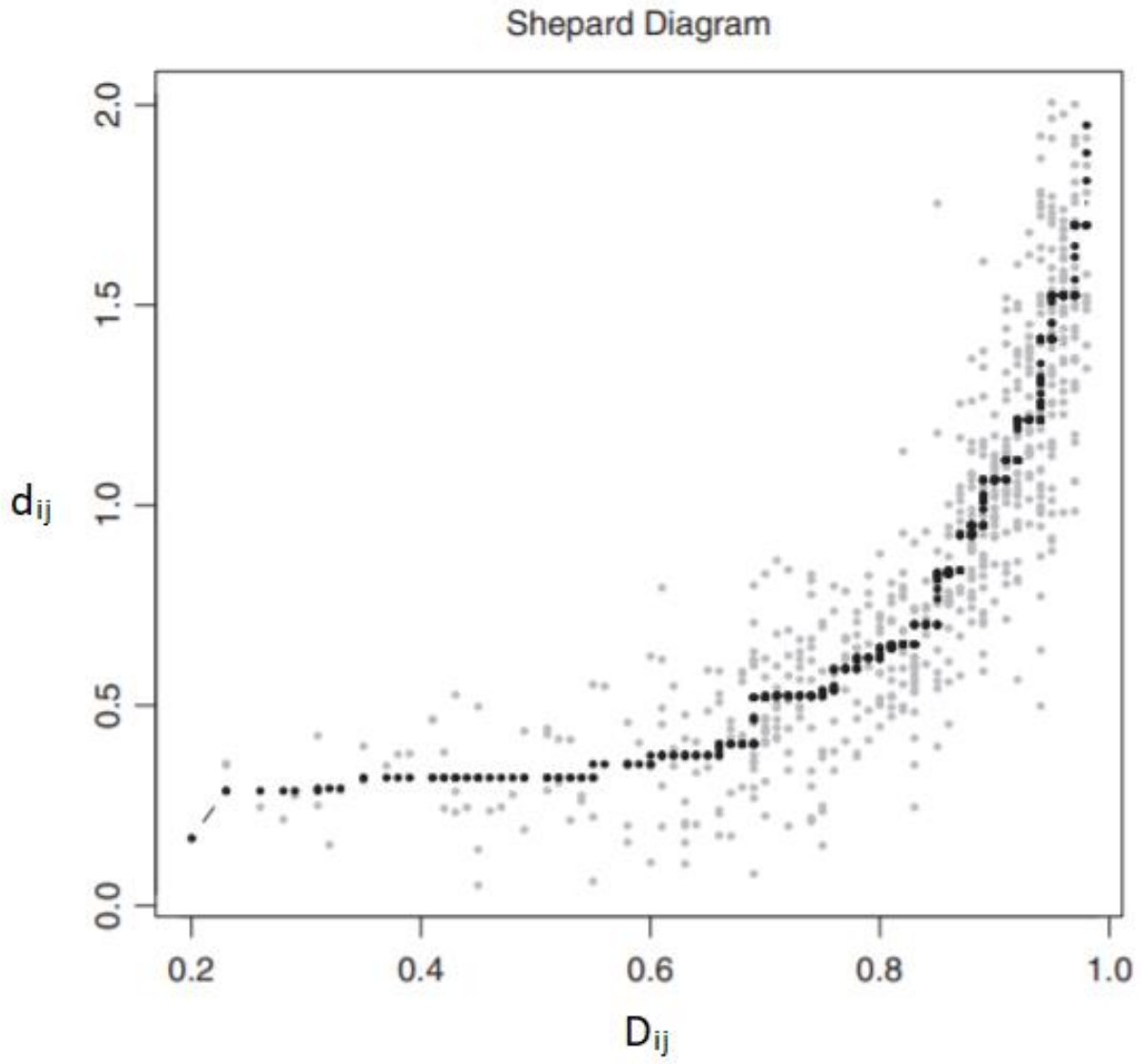
Metrikinės MDS atveju lyginama su tiesinės regresijos tiese.

Galima ieškoti, kokie taškai labiausiai nutolę nuo tiesės (netiksliai atvaizduojamas atstumas tarp dviejų objektų mažesnėje dimensijoje).

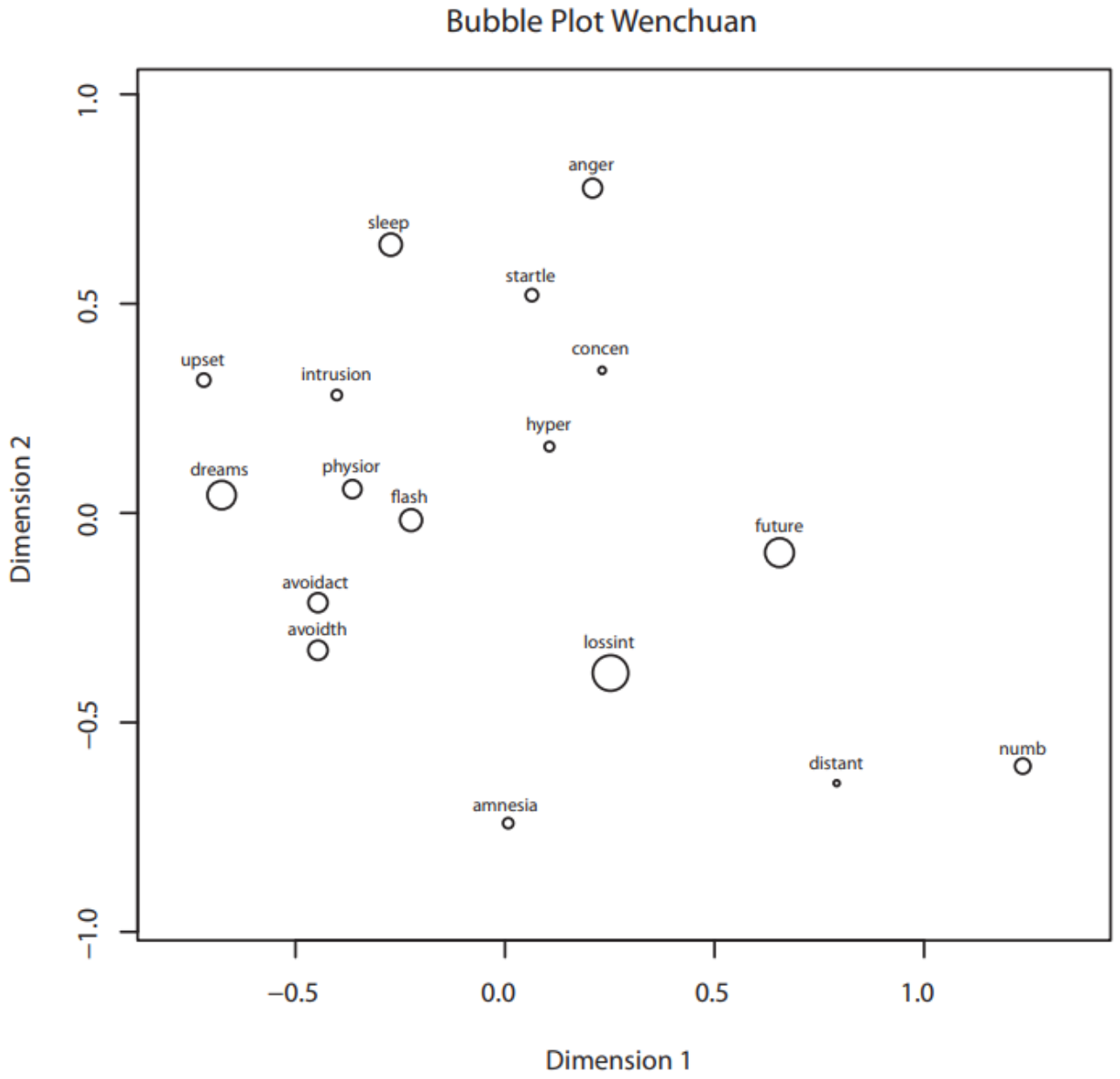


Nemetrikinės MDS atveju vaizduojama monotoninės regresijos kreivė.

Šiame pavyzdyje matoma, kad nemetrikinė MDS geriau tinka šiems duomenims.



- Bendresnis būdas ieškoti blogai atvaizduojamų taškų yra stress per point.
- Kiekvienam objektui apskaičiuojama kokia dalis Stress gaunama dėl jo.
- Pvz. sklaidos diagramoje didesniais taškai vaizduojami objektai daugiau prisideda prie Stress.



MDS interpretacija

- Priešingai negu naudojant PCA, ašys nėra reikšmingos, nes MDS rezultatai pagrįsti vien tik atstumais tarp objektų.

Atstumai nekinta:

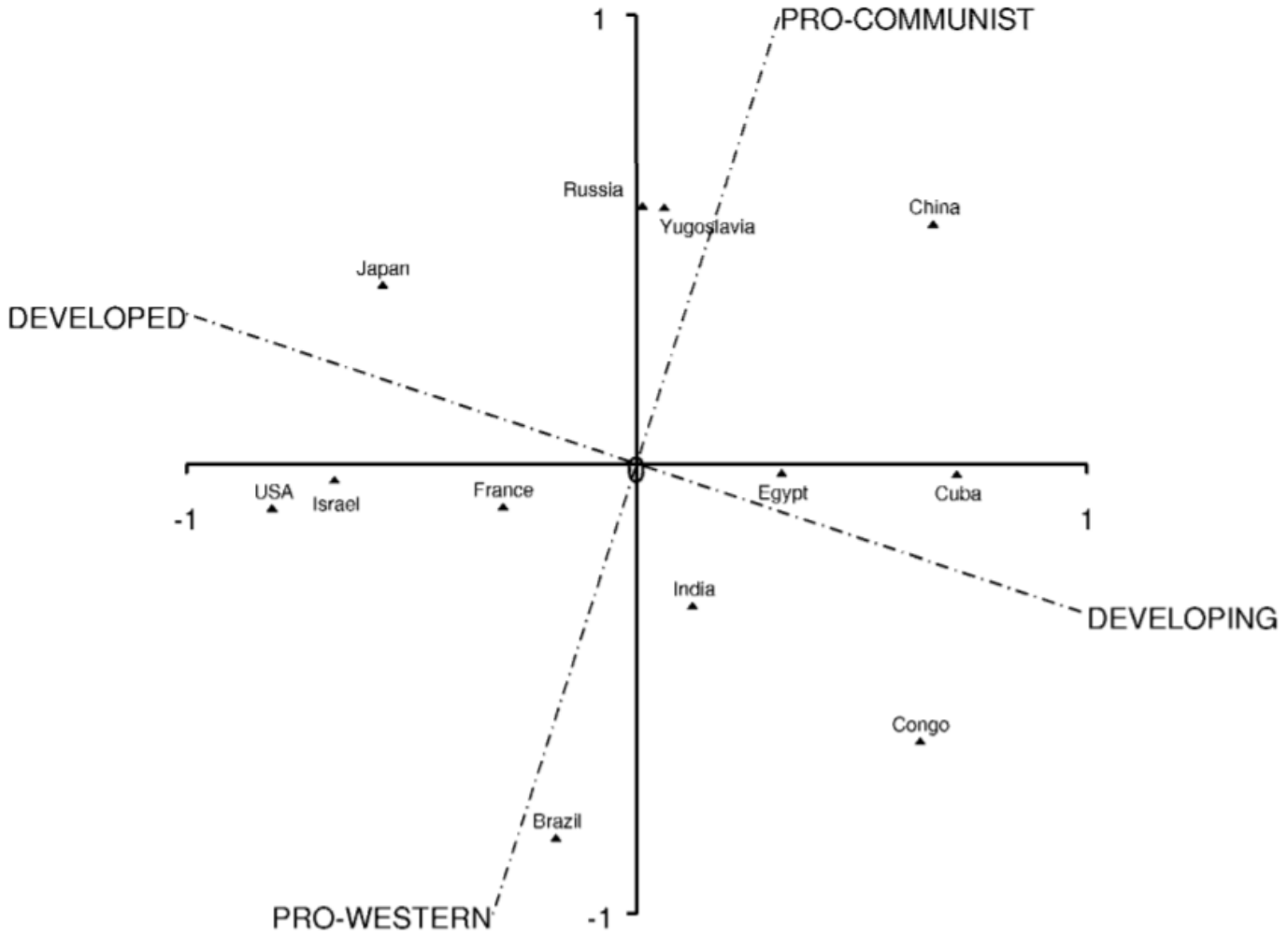
- Prie vienos koordinatės pridedant konstantą visiems objektams (paslinkus)
- Pasukant ašis
- Paimant atspindį kurios nors ašies atžvilgiu

Todėl peržiūrint MDS gauta rezultatą gali tekti ieškoti „prasmingiausių“ ašių.

Pvz. respondentai vertino šalis pagal jų panašumą.

Gautoje sklaidos diagramoje pridedamos prasminės ašys.

Kai kurios MDS implementacijos automatiškai panaudoja PCA perorientuoti ašis.



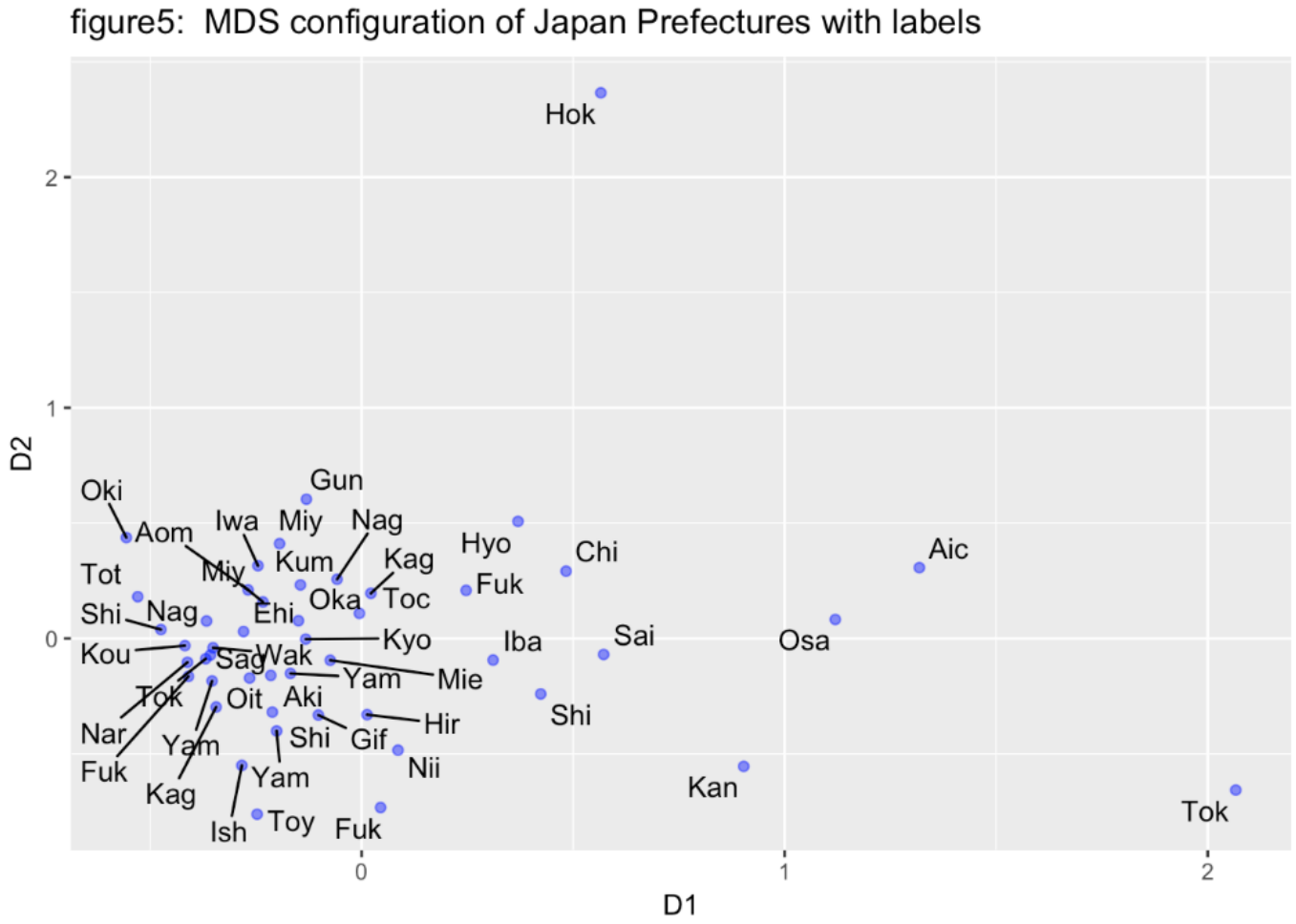
Dideli atstumai ir išskirtys

- MDS gauti atstumai tarp objektų yra kažkiek iškreipta jų tarpusavio santykio reprezentacija (jeigu Stress nelygus 0).
- Didesnės Stress reikšmės reiškia ši reprezentacija yra labiau iškreipta.

- Gautus didelius atstumus tarp objektų galima interpretuoti kaip „teisingus“:
- Jeigu atstumai tarp kažkurių objektų didelės dimensijos erdvėje dideli, o gautoje – maži (arba atvirkščiai), tai stipriai padidintų Stress reikšmę, vadinasi optimizacijos procesas „labiau“ stengiasi teisingai atvaizduoti šiuos atstumus.

Pvz. sumažinę dimensiją naudodami MDS pastebime, kad Aiči, Hokaido, Tokijo prefektūros yra išsiskiriančios iš kitų.

PCA stipriai paveikiamas išskirčių, MDS šiuo atveju jas randa.



Privalumai

- Netiesinė transformacija, kuri siekia išsaugoti duomenų topologiją.
- MDS nėra stipriai veikiama išskirčių kaip PCA, gali būti naudojama siekiant jas aptikti.
- Vienas iš paprasčiausių netiesinių dimensijos mažinimo metodų.

Trūkumai

- Gautos dimensijos neturi aiškios interpretacijos.
- Sunkiau parinkti dimensijų kiekį (PCA galima parinkti naudojant paaiškintą variaciją).
- Su optimizavimu susijusios problemos (nebėra tokios svarbios padidėjus skaičiavimo galingumams):
  - Pradinis duomenų išdėstymas daro įtaką galutiniam rezultatui.
  - Gali būti nerastas optimalus sprendimas.
  - Pridėjus naujų stebėjimų duomenų konfiguracija turi būti randama iš naujo.

Paruošta pagal:

- http://web.vu.lt/mii/j.zilinskas/DzemydaKurasovaZilinskasDDVM.pdf
- https://www.bristol.ac.uk/media-library/sites/cmm/migrated/documents/chapter3.pdf
- https://www.researchgate.net/publication/280717361_Shepard_Diagram
- https://www.researchgate.net/publication/309617943_Goodness-of-Fit_Assessment_in_Multidimensional_Scaling_and_Unfolding
- https://rstudio-pubs-static.s3.amazonaws.com/246348_b31bca1e4be04bb395825dc6a00de364.html#3_why_mds_advantage_and_disadvantages_of_mds

Ačiū už dėmesį