Daugiamatės skalės (Multidimensional scaling)

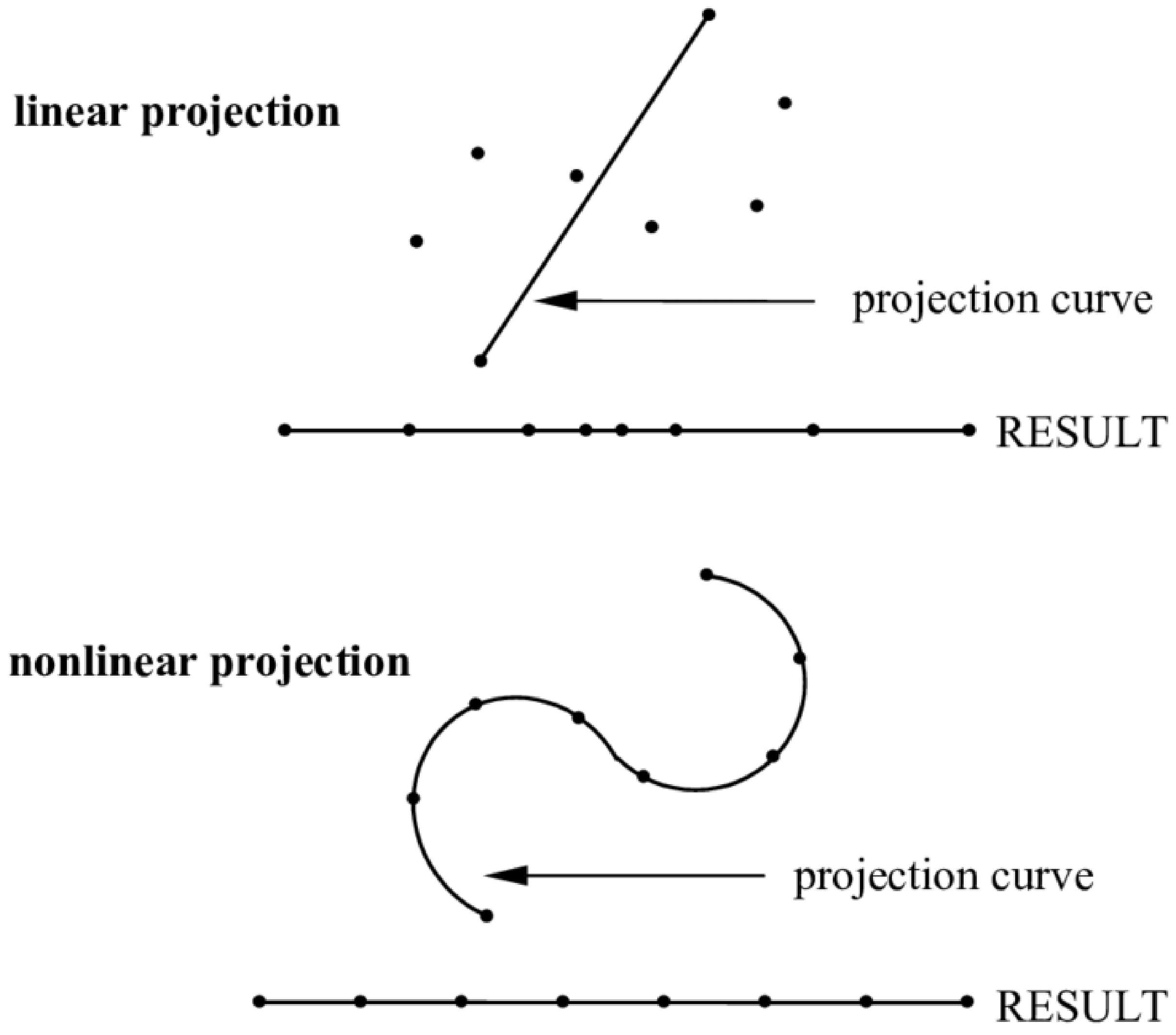
Matas Gaulia, Vainius Gataveckas, Dovydas Martinkus

Duomenų Mokslas 3 kursas 2 gr.

Vilnius, 2022

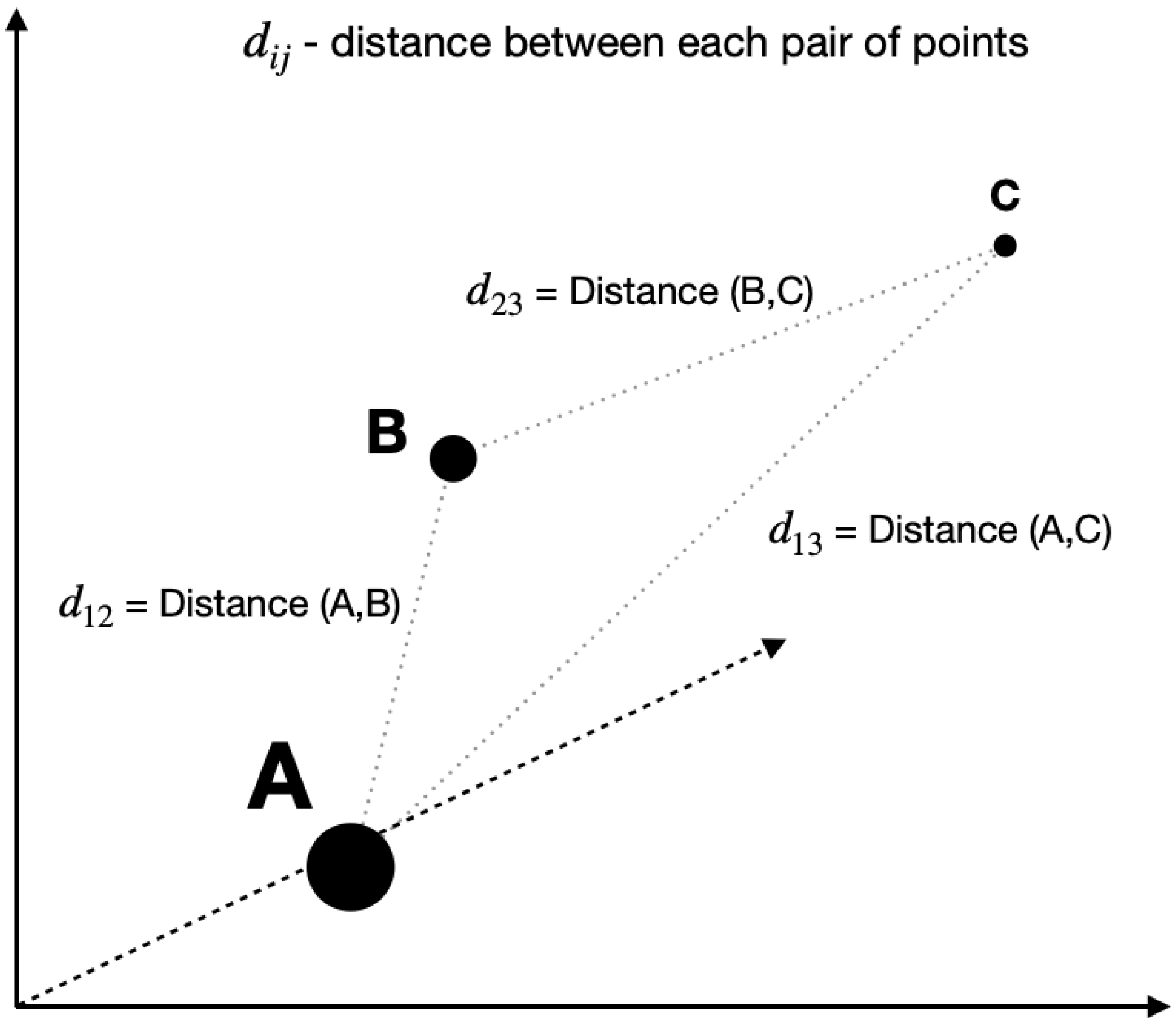
Tiesiniai ir netiesiniai dimensijos mažinimo metodai

- 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

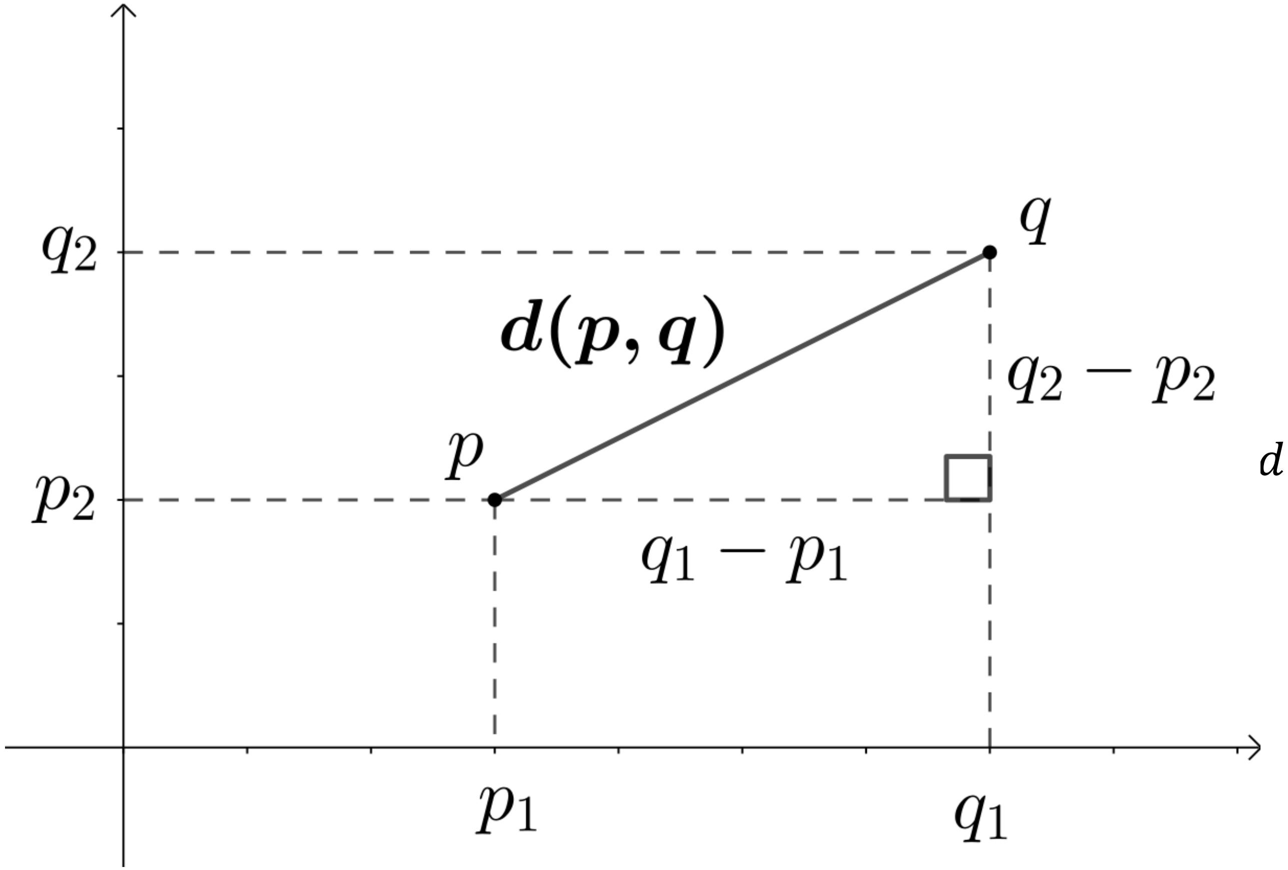
Daugiamatės skalės

- Daugiamatės skalės kiekvieną objektą iš didesnės dimensijos duomenų erdvės transformuoja į iš anksto parinkto dydžio mažesnės dimensijos erdvę (vadinama vaizdo erdve).
- Naudojant MDS ieškoma daugiamačių duomenų projekcijų vaizdo 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



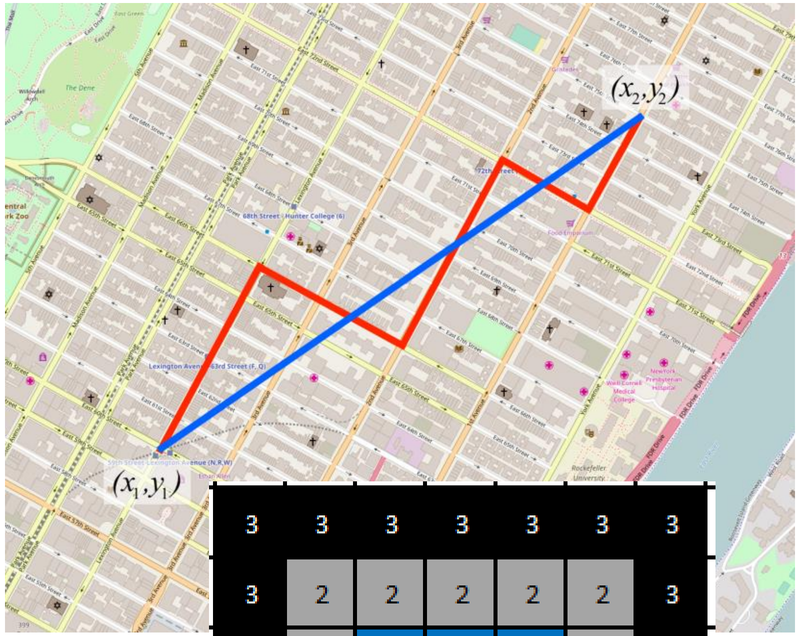
Įprastai naudojamas Euklidinis atstumas:

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

Kiti galimi atstumai

Pvz:

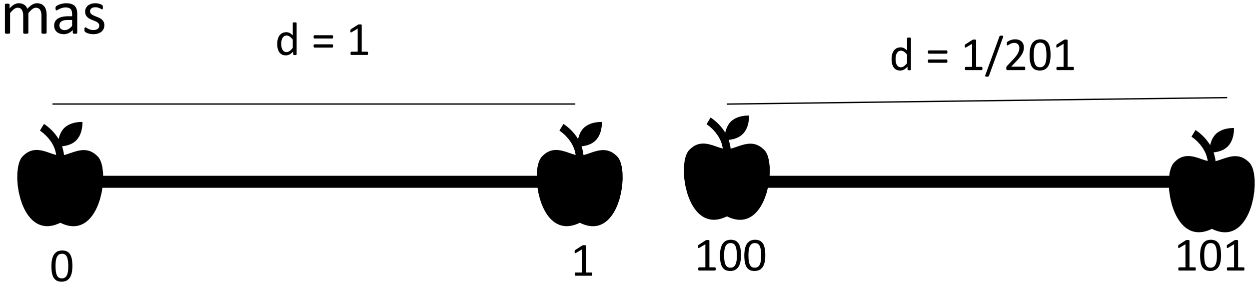
- $d_1(X_k, X_l) = \sum_{j=1}^{n} \left| x_{kj} - x_{lj} \right|$ Manheteno atstumas



- $d_{\infty(X_k,X_l)} = \mathrm{max_j} \left| \mathrm{x}_{\mathrm{k}j} - \mathrm{x}_{\mathrm{lj}} \right|$ Čebyševo atstumas

3	3	3	3	3	3	3
3	2	2	2	2	2	3
3	2	1	1	1	2	3
3	2	1	0	1	2	3
3	2	1	1	1	2	3
3	2	2	2	2	2	3
3	3	3	3	3	3	3

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



Nepanašumai

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

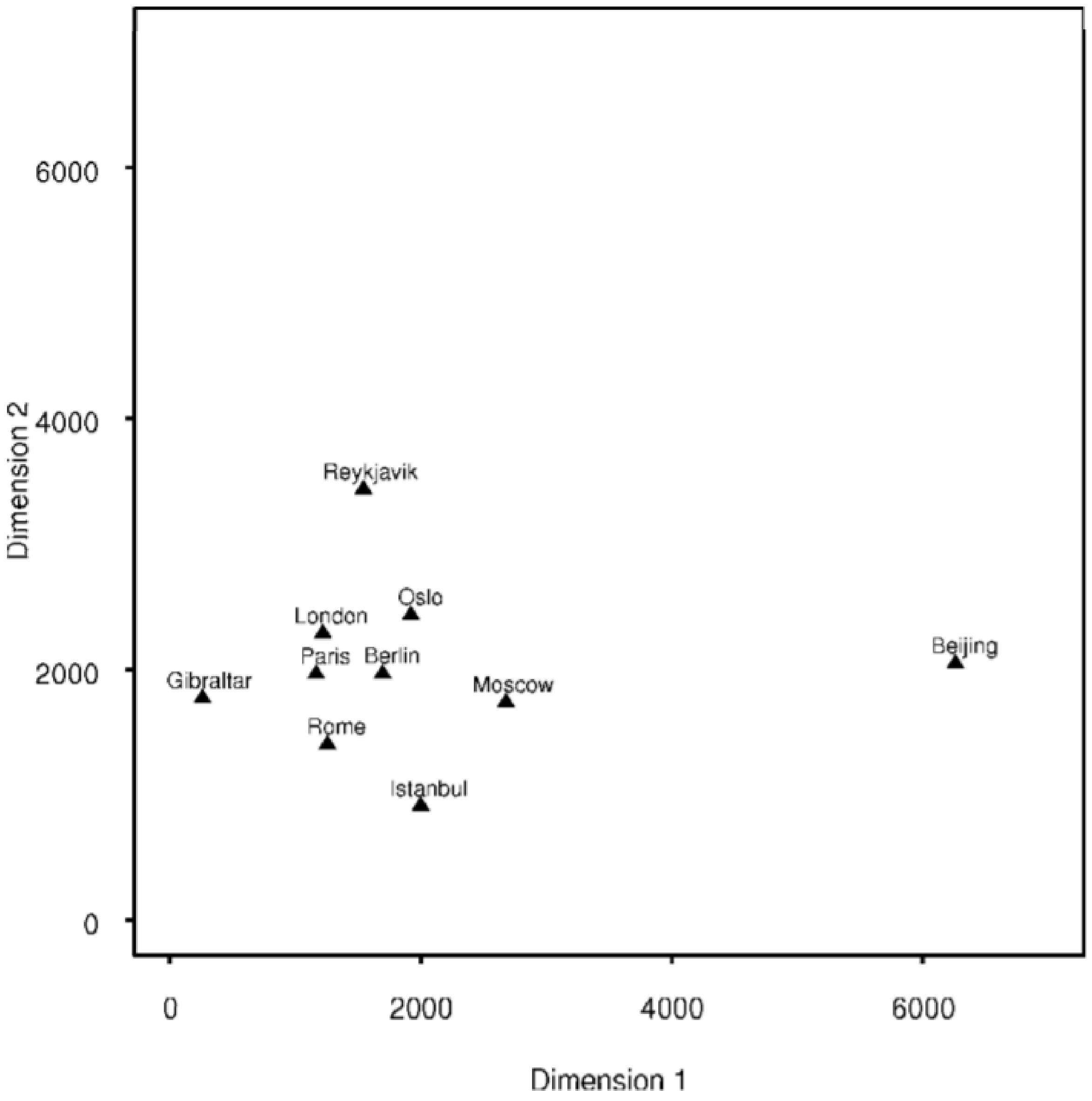
- Tokie matavimai dažnai naudojami sociologijoje, psichologijoje, kitose srityse. Pvz. kažkokie subjektyvūs įvertinimai.

Nepanašumų matrica

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



Įterpimo funkcijos

Atvaizdavimo kokybė matuojama įterpimo funkcija, kuria naudojantis lyginamas objektų nepanašumas su atstumu tarp juos atvaizduojančių taškų.

Dažniausiai naudojama:

$S(Y) = \sum_{i<j} w_{ij}\left(d_{ij} - D_{ij}\right)^2$ mažiausių kvadratų įterpimo funkcija (Stress)

$d_{ij} = d\left(Y_i, Y_j\right)$ - tuo metu turimas Euklidinis atstumas tarp objektų Y vaizdo erdvėje.

D_{ij} - nepanašumas duomenų erdvėje.

w_{ij} - neneigiami simetriški svoriai

Įtempimo funkcijos gali būti:

- $\sum_{i<j} w_{ij}\left(d_{ij}^2 - \mathrm{D}_{\mathrm{ij}}^2\right)^2$
- $\sum_{i<j} w_{ij}\left|d_{ij} - D_{ij}\right|$
- ...

Nuo įterpimo funkcijos priklauso gaunamas rezultatas vaizdo erdvėje.

Bendru atveju d_{ij} gali būti ir kitoks negu Euklidinis atstumas.

Disparities

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

Tada mažiausių kvadratų įterpimo funkcijos (Stress) atveju gaunama:

$$\sum_{i<j} w_{ij}\left(d_{ij} - \widehat{D_{ij}}\right)^2$$

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 vaizdo erdvėje siekiama, kad atstumai tarp taškų būtų kuo panašesni į nepanašumus duomenų 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 tik kad atstumų tvarka vaizdo erdvėje sutaptų su nepanašumų tvarka duomenų erdvėje.

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

Optimizavimas

- Dimensijos mažinimas naudojant MDS yra optimizavimo procesas.

- Naudojamas iteratyvus algoritmas, kuris minimizuoja tam tikrą įterpimo funkciją.
- Pvz. scikit-learn naudojamas SMACOF algoritmas.

Pradinė konfigūracija

- MDS gautas rezultatas priklauso ne tik nuo įterpimo funkcijos, bet ir nuo pradinės taškų konfigūracijos vaizdo erdvėje.

Egzistuoja du dažniausiai naudojami pradinės konfigūracijos būdai:

- Taškai išdėliojami atsitiktinai.
- Naudojami klasikinio MDS (Torgeson‘s MDS) metodo, kuris sprendimą gauna analiziškai, tačiau yra mažiau lankstus už skaitinius MDS, gauti rezultatai.

Prieš tai minėti faktai apie įterpimo funkcijos minimizavimą ir pradinės konfigūracijos pasirinkimą reiškia kad:

- Jeigu pradinė konfigūracija yra atsitiktinė, tai kiekvieną kartą randamas kitoks sprendimas.
- Algoritmas gali užstrigti lokaliame įterpimo funkcijos minimume. Siekiant to išvengti algoritmas paleidžiamas kelis kartus su kitokiomis pradinėmis reikšmėmis ir pasirenkamas geriausias spendimas.
- Reikia pasirinkti maksimalų leidžiamą iteracijų skaičių arba sąlygą, kada deklaruojamas konvergavimas.

Bendra MDS schema

- Apibendrinus, bendra MDS schema atrodo taip:

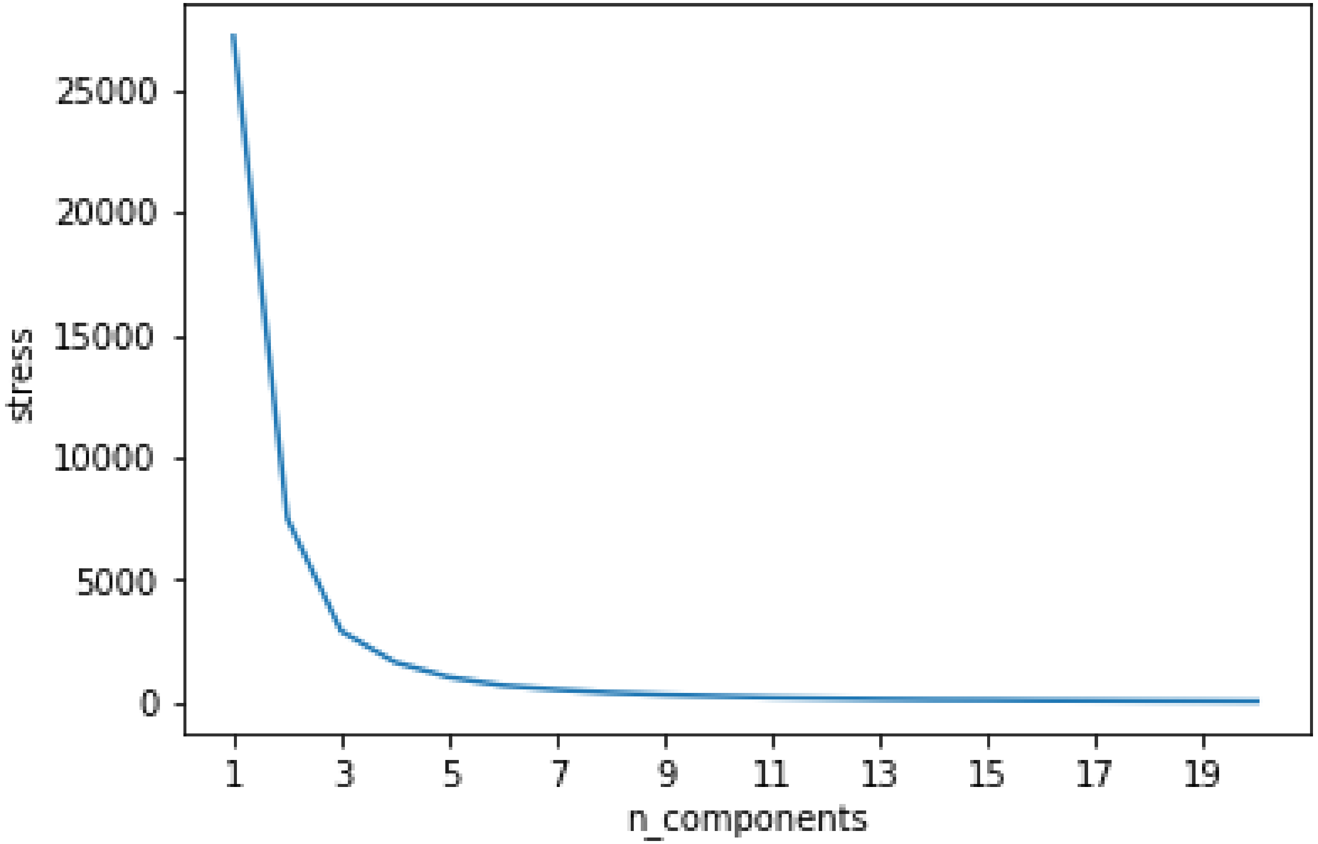
    1. Pradinis taškų išsidėstymas.
    2. Apskaičiuojamas įterpimo funkcijos reikšmė*.
    3. Įterpimo funkcijos minimizavimas tam tikru algoritmu.
    4. 2 ir 3 žingsnio kartojimas iki konvergavimo.

* Kiekvieną kartą prieš apskaičiuojant įterpimo funkcijos reikšmes iš naujo turi būti apskaičiuojamos disparities (nemetrikiniu atveju pakartotinai atliekama monotoninė regresija).

Vaizdo erdvės dimensijos parinkimas

- Norima dimensija turi būti parenkama iš anksto.
- Natūralu, kad įterpimo funkcijos reikšmė didėja kuo labiau mažinama dimensija.
- Galimas dimensijos dydžio vaizdo erdvėje parinkimo būdas yra ieškant mažiausios dimensijos, kuri vis dar turi pakankamai mažas įterpimo funkcijos reikšmes.

Scree plot ieškoma alkūnės taško (angl. elbow point)



MDS interpretacija

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

Atstumai nekinta:

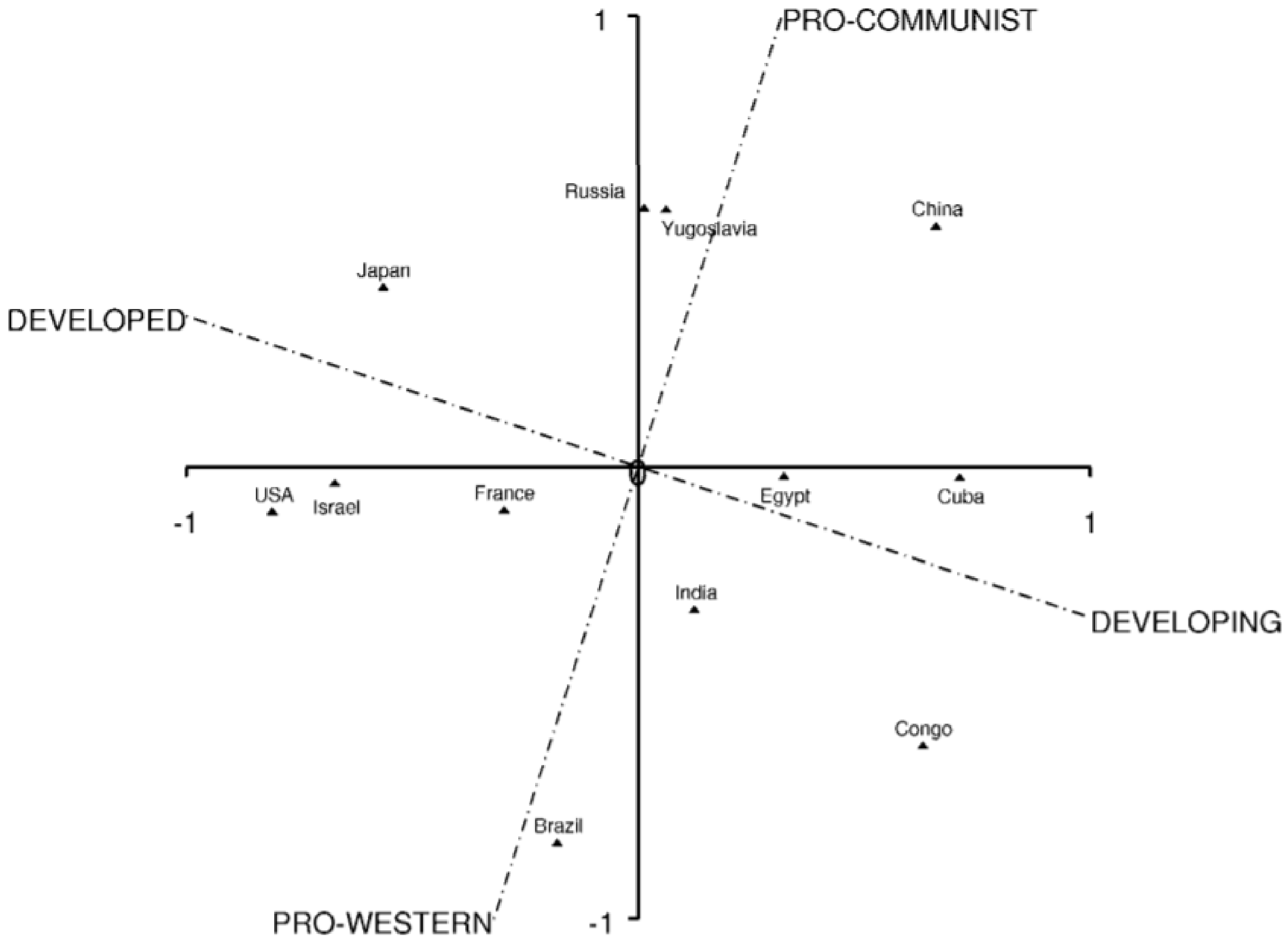
- Prie vienos koordinatės pridedant konstantą visiems objektams (paslinkus)
- Pasukant ašis
- Paimant atspindį kurios nors ašies atžvilgiu

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.



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 įterpimo funkcijos optimizavimu susijusios problemos:
  - Pradinis duomenų išdėstymas daro įtaką galutiniam rezultatui.
  - Gali būti nerastas optimalus sprendimas.
  - Pridėjus naujų stebėjimų duomenų konfigūracija turi būti randama iš naujo.

Papildoma informacija

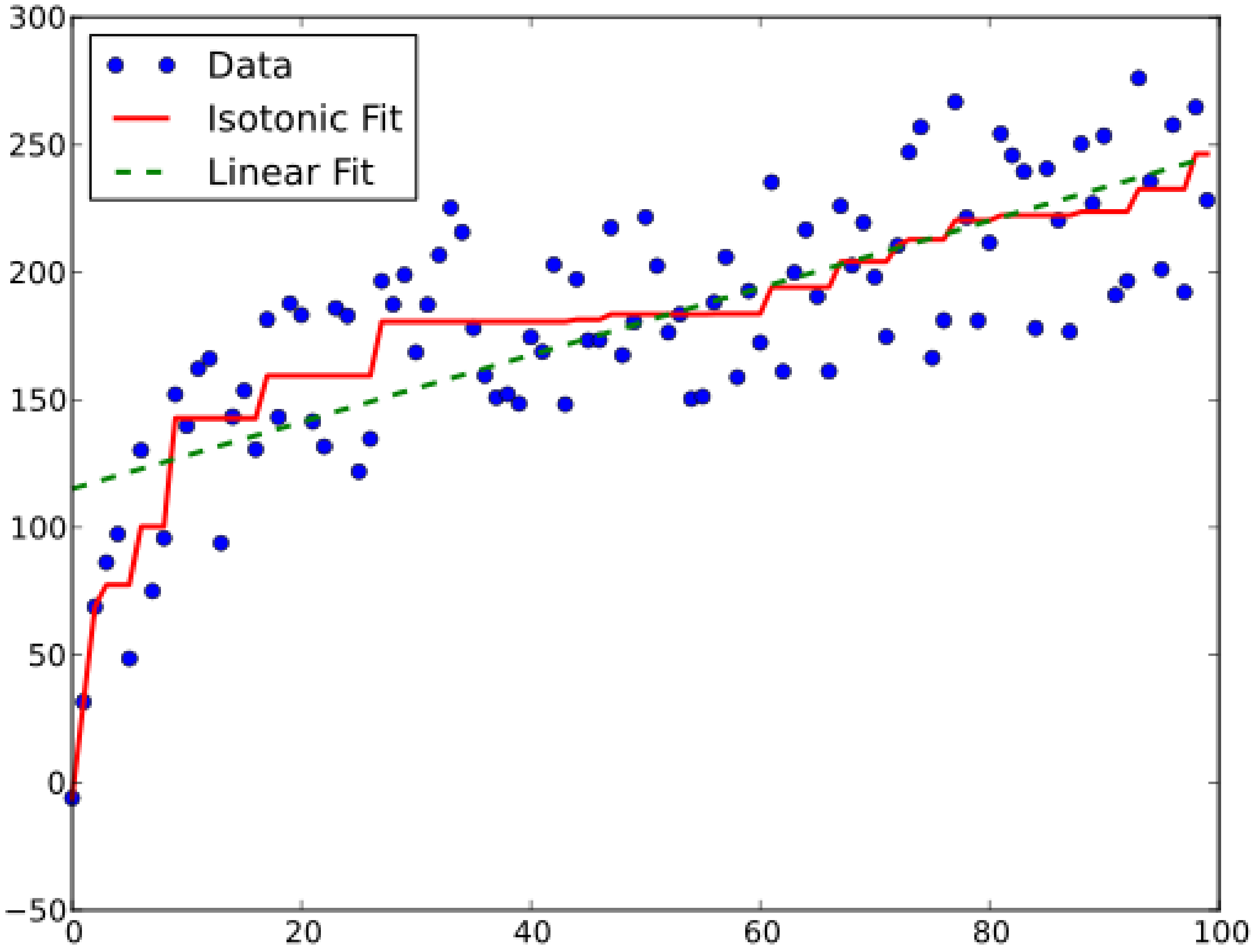
Monotoninė regresija

- Paprastas algoritmas užtikrinti sąryšį, reikalingą nemetrikinei MDS, 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.

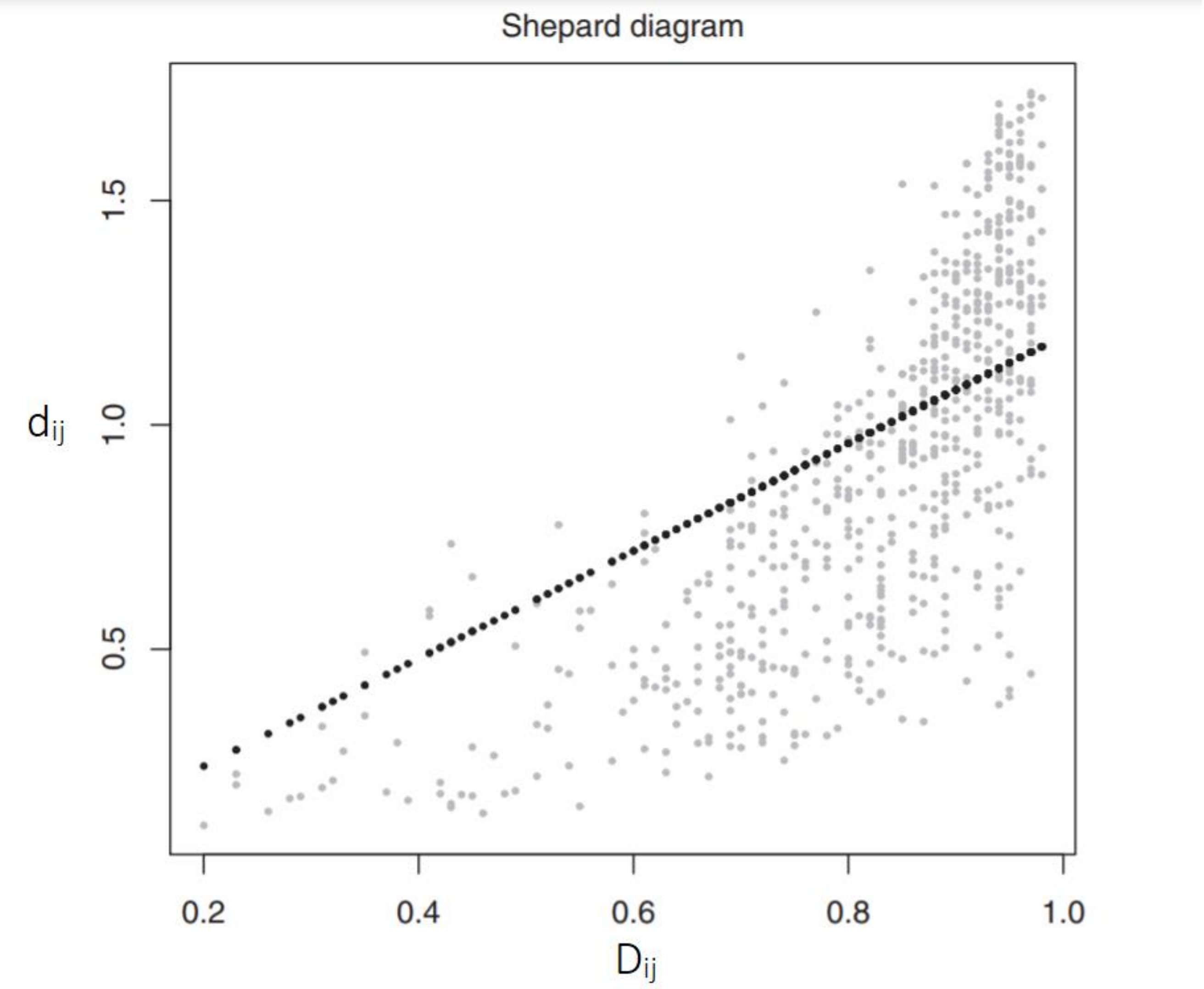
Siekiama, kad kitoje iteracijoje monotoninės regresijos kreivė būtų labiau tolygiai „laiptuota“.



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

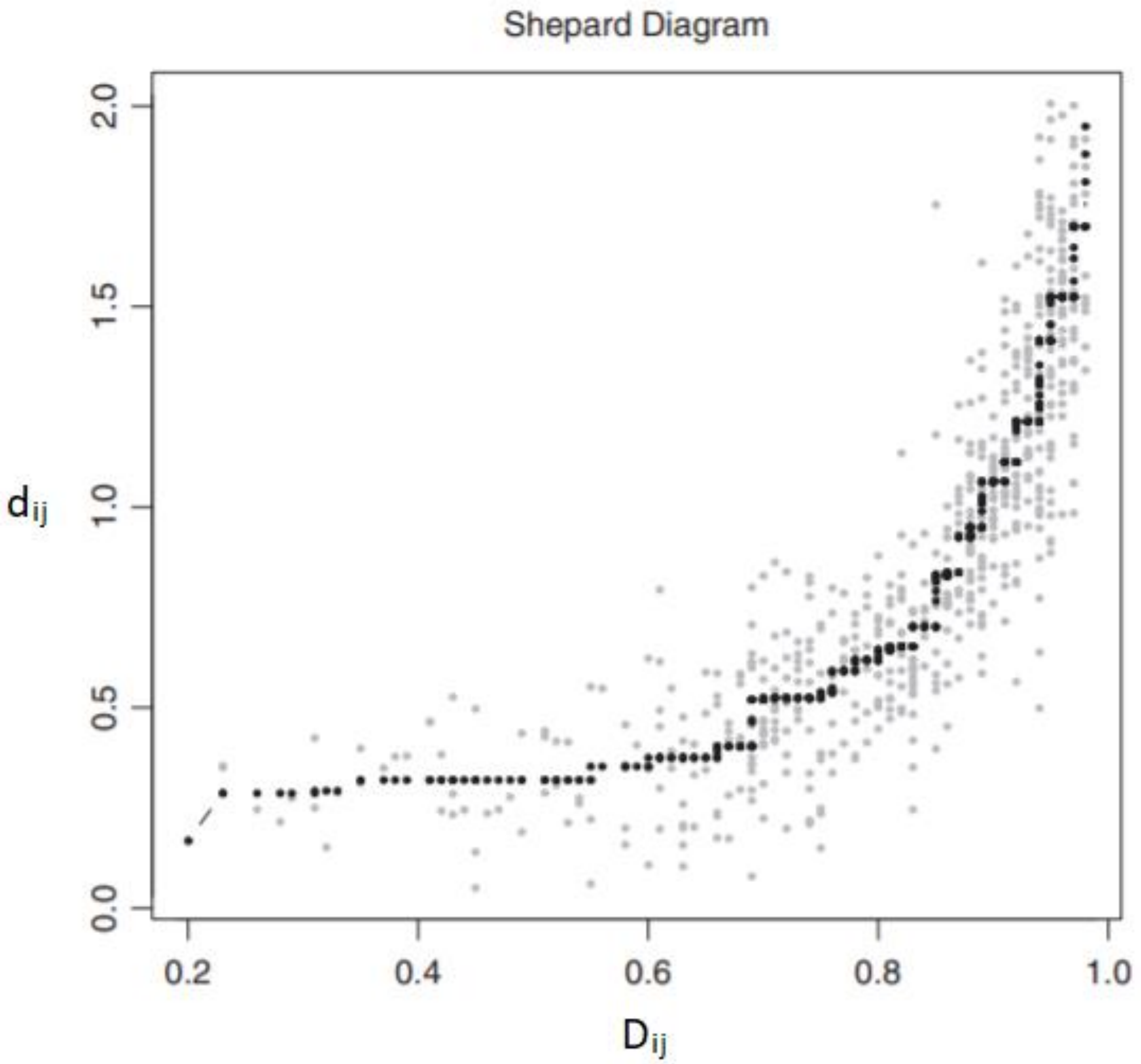
Tarp jų dažniausiai naudojama Shepard diagrama, y ašyje vaizduojanti atstumus vaizdo erdvėje, x ašyje – atstumus duomenų erdvėje tarp tų pačių objektų.

Metrikinės MDS atveju taškų pasiskirstymas lyginamas su tiesinės regresijos tiese.

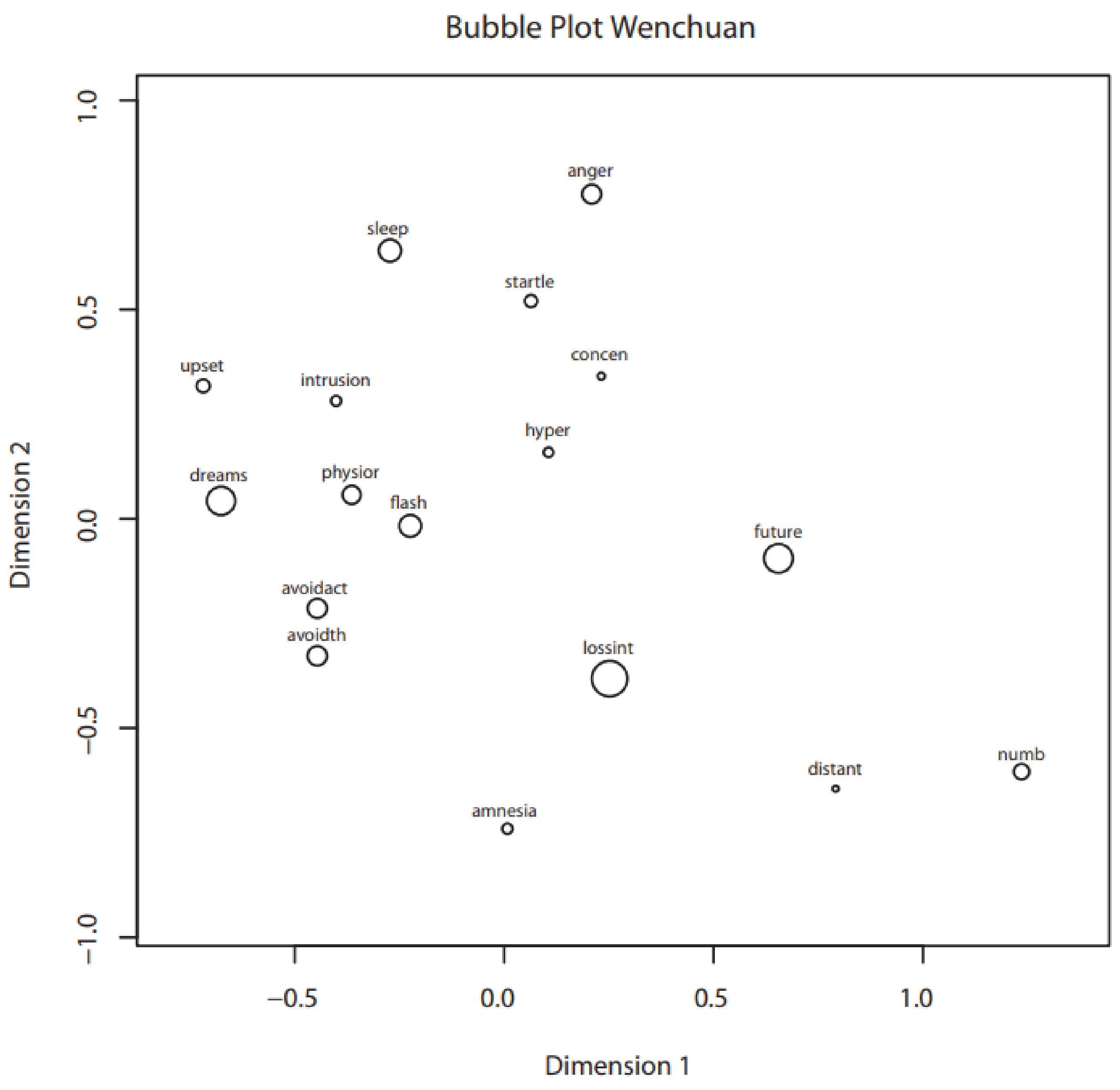


Nemetrikinės MDS atveju pasiskirstymas lyginamas su monotoninės regresijos kreive.

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



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



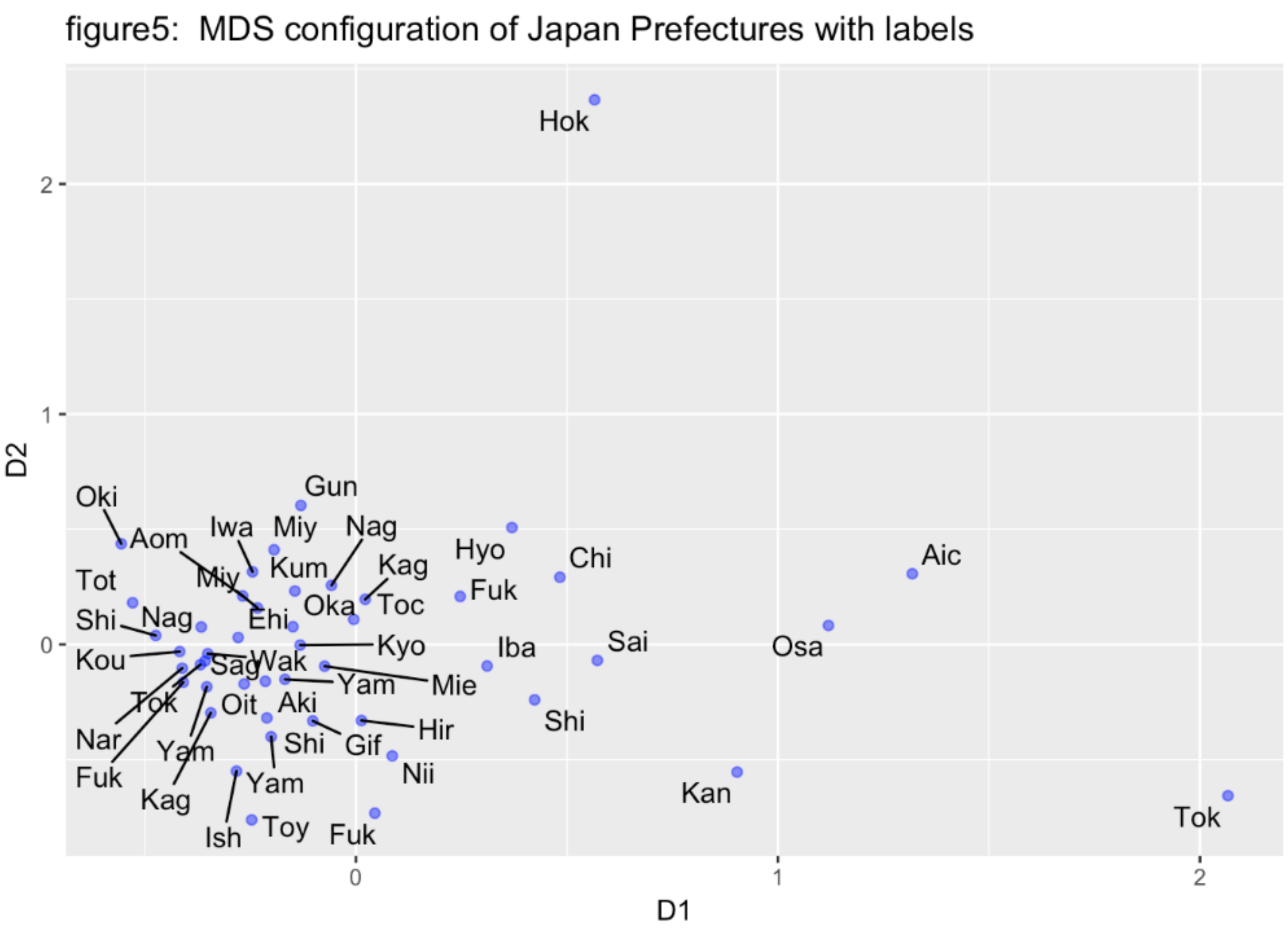
Dideli atstumai ir išskirtys

- MDS gauti atstumai tarp objektų vaizdo erdvėje visada yra kažkiek iškreipta jų tarpusavio santykio reprezentacija (jeigu įterpimo funkcijos reikšmė nelygi 0).
- Didesnės įterpimo funkcijos reikšmės reiškia, kad ši reprezentacija yra labiau iškreipta.

- Tačiau 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 įterpimo funkcijos reikšmę, vadinasi optimizacijos algoritmas „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.



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://www.researchgate.net/publication/305303417_The_Choice_of_Initial_Configurations_in_Multidimensional_Scaling_Local_Minima_Fit_and_Interpretability

Ačiū už dėmesį