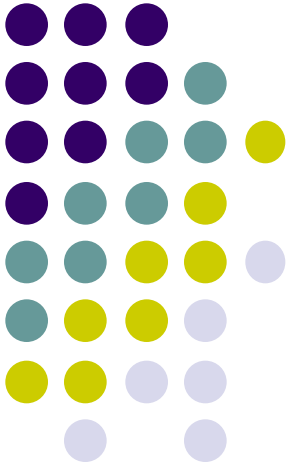
TÖBBVÁLTOZÓS ADATELEMZÉS

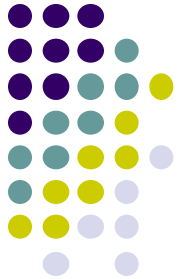


Főkomponens-elemzés

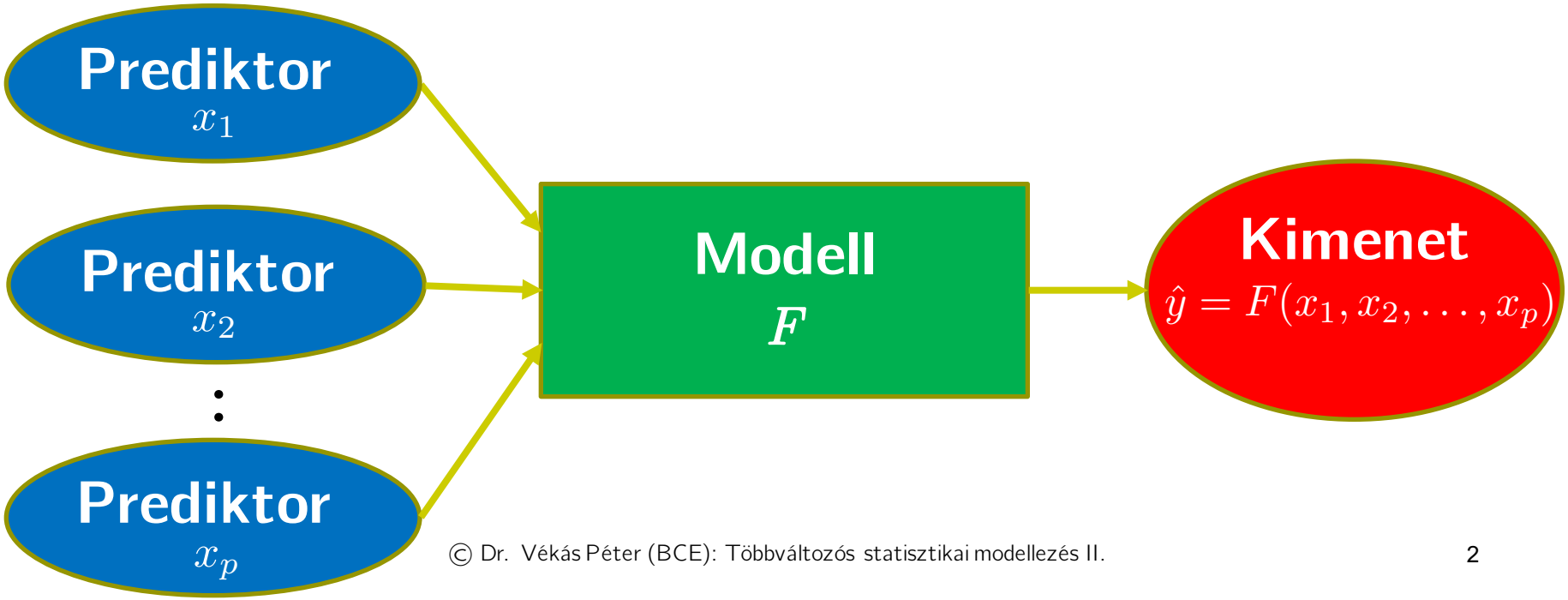
2020.11.23.

© Dr. Vékás Péter, e-mail: peter.vekas@uni-corvinus.hu
BCE Matematikai és Statisztikai Modellezés Intézet

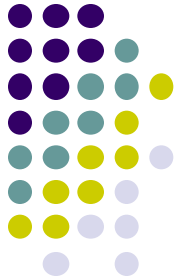
Gépi tanulási modell mint függvény ("varázsdoboz")



- A gépi tanulási modellek prediktorváltozókból (=inputok, független változókat) hoznak létre egy kimenetet (=output, függő változó).

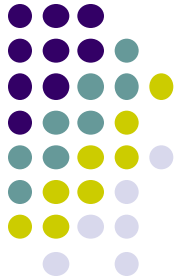


Gépi tanulási problémák típusai



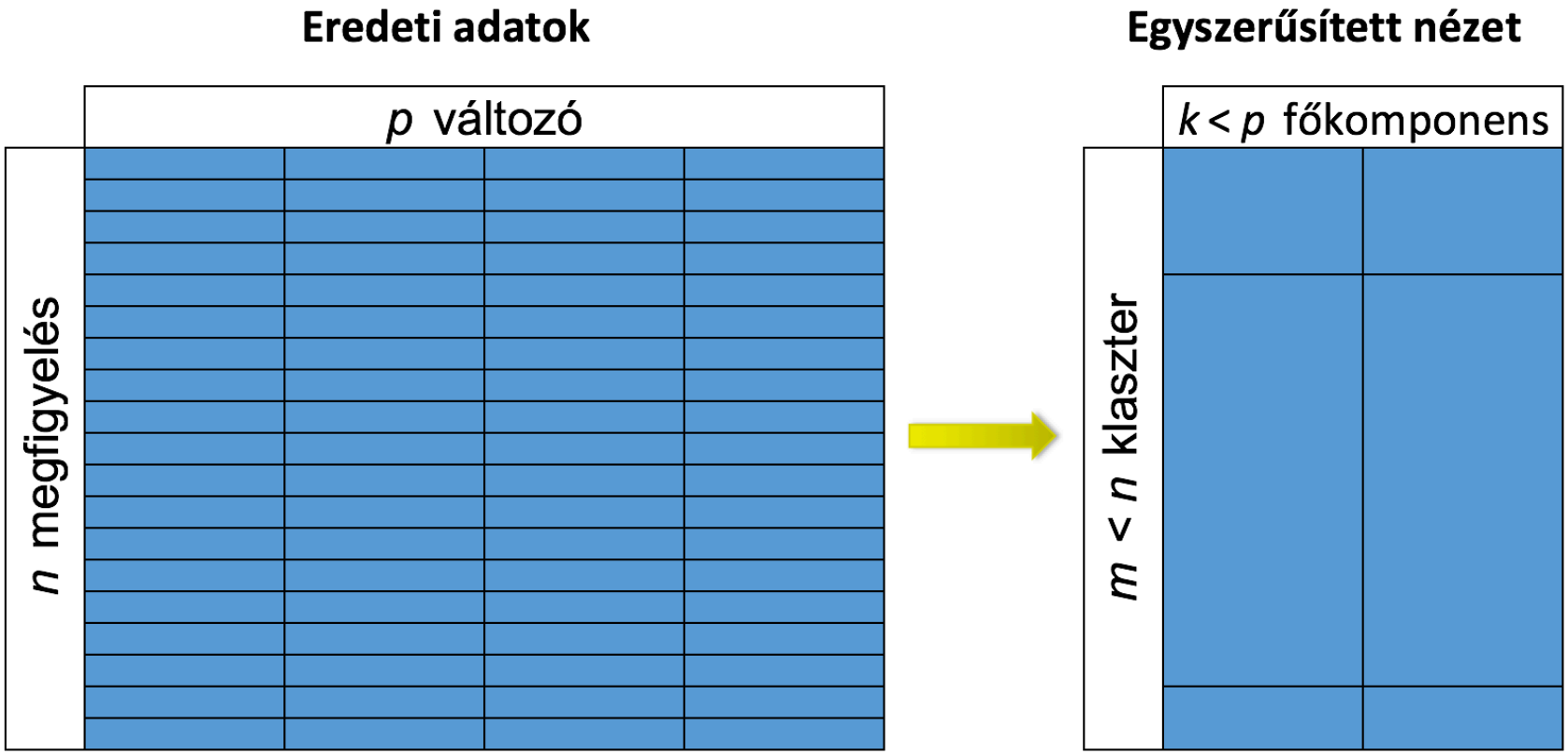
- **Felügyelt tanulás:**
  Van egy adatfájlunk, ahol ismert a "valódi" kimenet. A cél ennek a becslése (predikció).
  - Osztályozás (klasszifikáció): kategorikus kimenet.
  - Regresszió: numerikus kimenet.
- **Nem felügyelt tanulás:**
  A kimenet ismeretlen, "láthatatlan" változó.
  A cél az adatok struktúrájának megismerése.
  - Klaszterezés: kategorikus kimenet.
  - Dimenziócsökkentés: numerikus kimenet.

Gépi tanulási problémák típusai

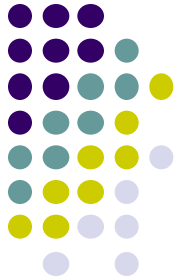


Kimenet / Modell	Felügyelt $\mathbf{x} \to y$	Nem felügyelt $\mathbf{x} \to ?$
Kategorikus 	Osztályozás	Klaszterezés
Numerikus 	Regresszió	Dimenzió-csökkentés

Az elemző a lényegre kíváncsi



Főkomponens-elemzés

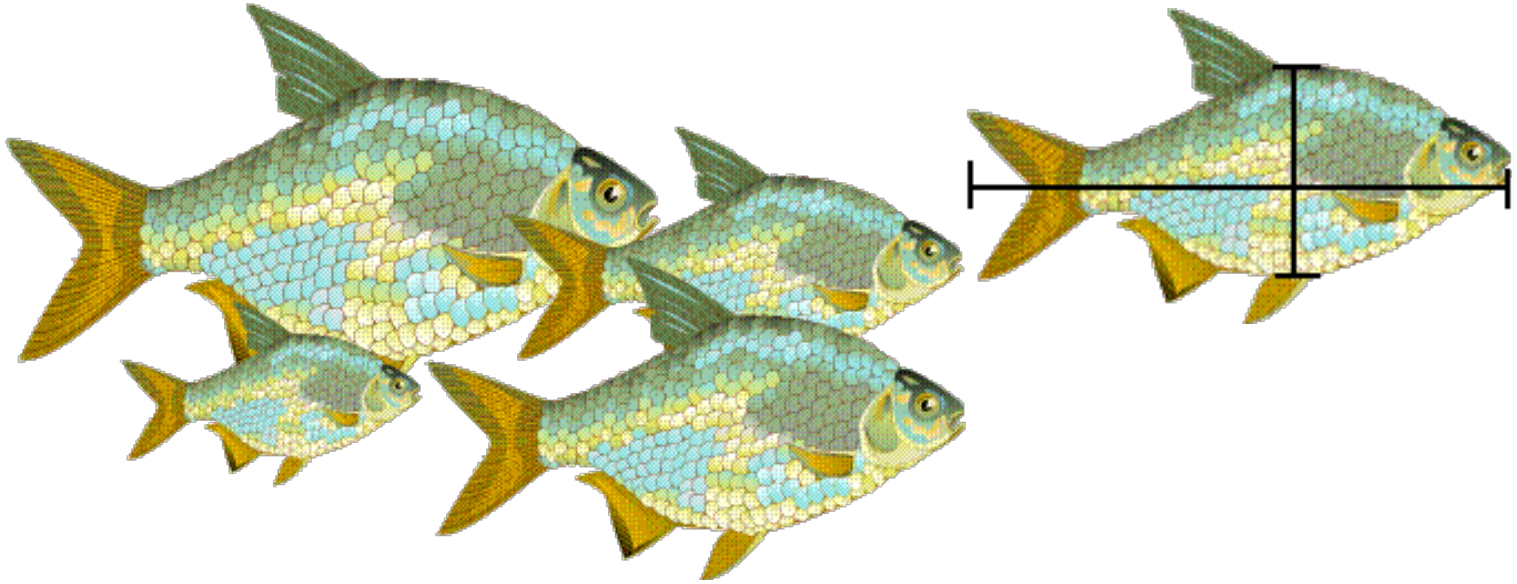


- Adott több egymással korreláló változó.
- Főkomponens-elemzés célja: egymással korreláló változók korrelálatlan főkomponensekbe tömörítése
- **Főkomponensek:** a sztenderdizált változók *lineáris kombinációiként* állnak elő
- Az eredetinél kevesebb számú változót tartunk meg (-> **dimenzió-csökkentés)**, úgy hogy az adott dimenziószám mellett legyen maximális a magyarázott variancia.

Grafikus magyarázat

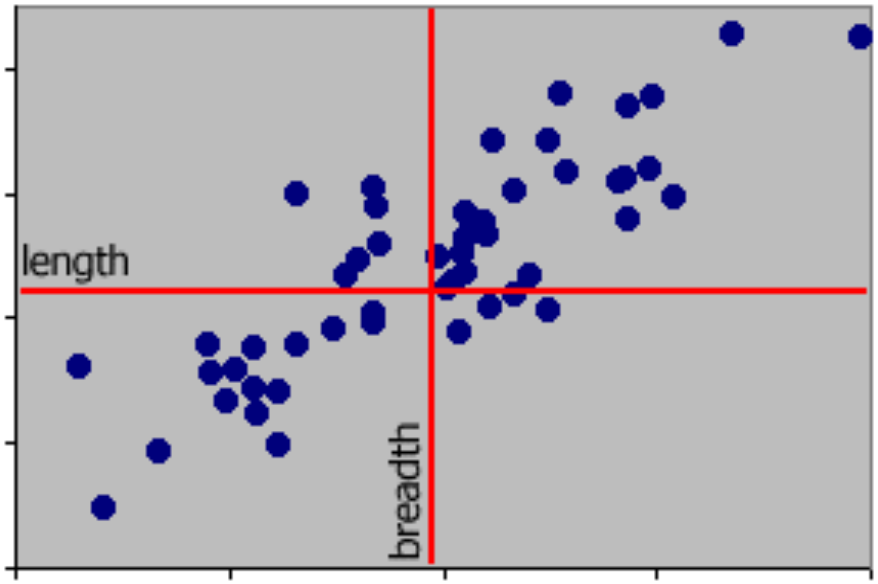
- Forrás: http://www.cmbi.ru.nl/edu/bioinf4/prac-microarray/stats/PCA%20graphical%20explanation.htm

- Adott: halak hosszúságai és magasságai (két változó).

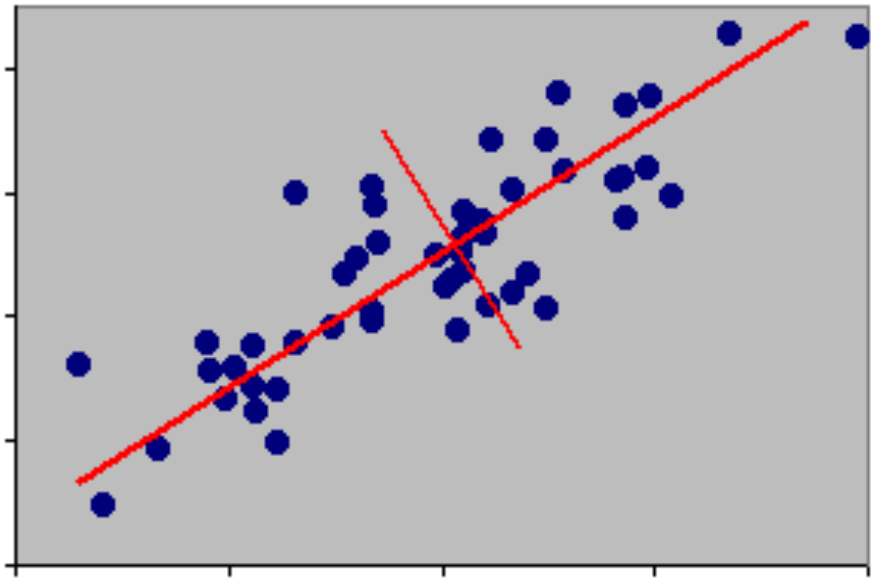


Áttérés új derékszögű koordinátarendszerre

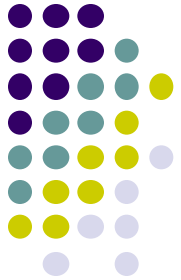
Sztenderdizált változók (eredeti tengelyek):



Főkomponens-tengelyek:

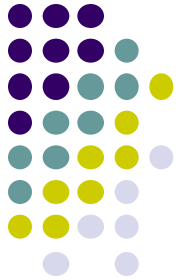


A főkomponensek előállítása



- Az előbbi ábra alapján úgy tűnik, a halak hosszúsága és magassága erősen összefügg (korreláló változók).
- Az eredeti koordinátarendszer tengelyei helyett megkeressük azt az új tengelyt (origón átmenő egyenest), ami mentén a legnagyobb az értékek varianciája (ez a leginformatívabb lehetséges új tengely).

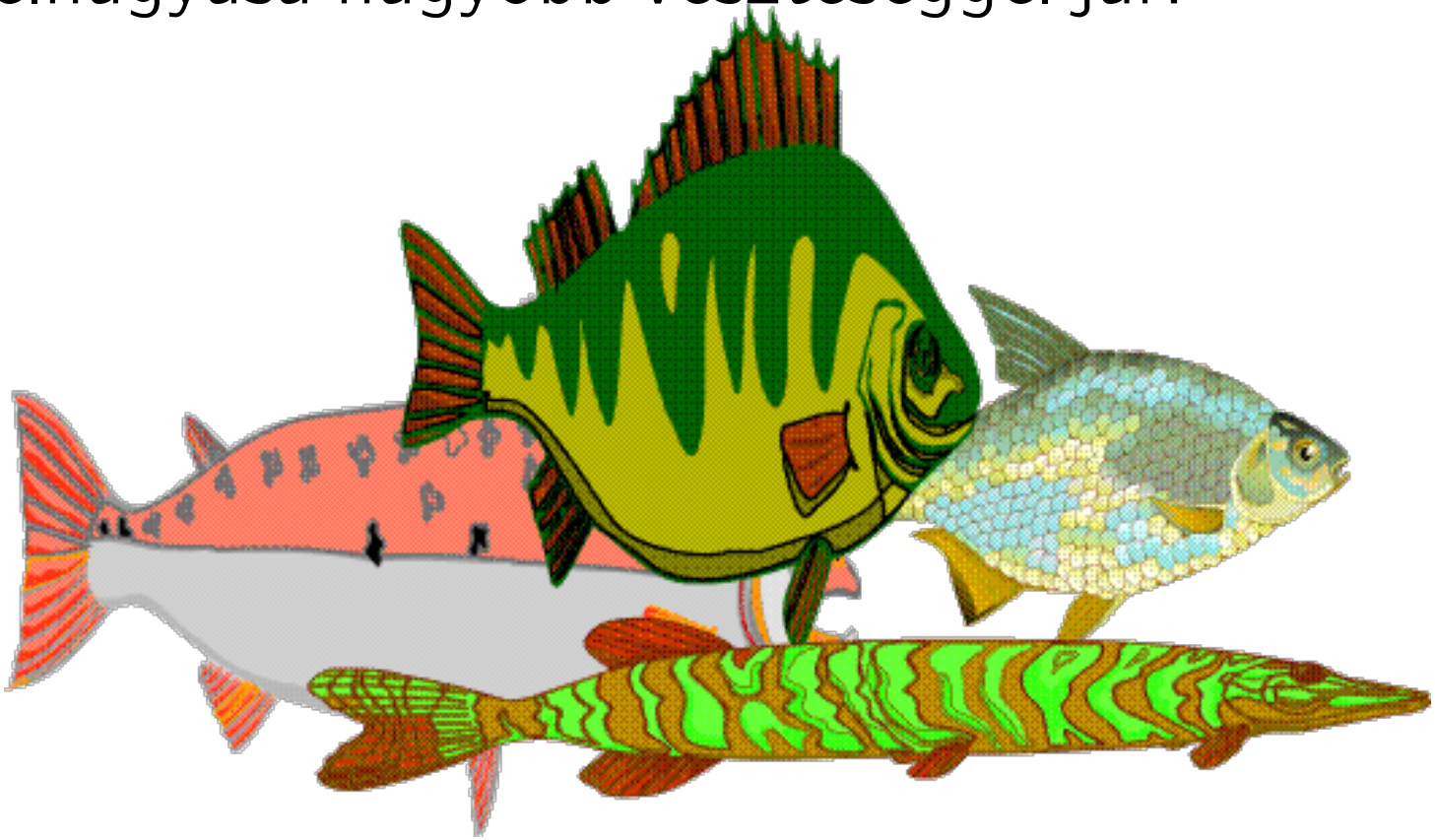
A főkomponensek előállítása



- Mivel új *derékszögű* koordinátarendszert keresünk, ezért a másik új tengelyt az erre merőleges, origón átmenő egyenes adja meg.
- A hosszúság és magasság varianciájának jó része egy új látens változóba („a hal mérete”) tömöríthető. A másik tengelyen („a hal tömzsisége”) mért variancia elhanyagolható, ezt elhagyva egy dimenziót kapunk.

Fontos, hogy erősen korreláló változók legyenek!

- Gyenge korreláció esetén a második dimenzió elhagyása nagyobb veszteséggel jár:



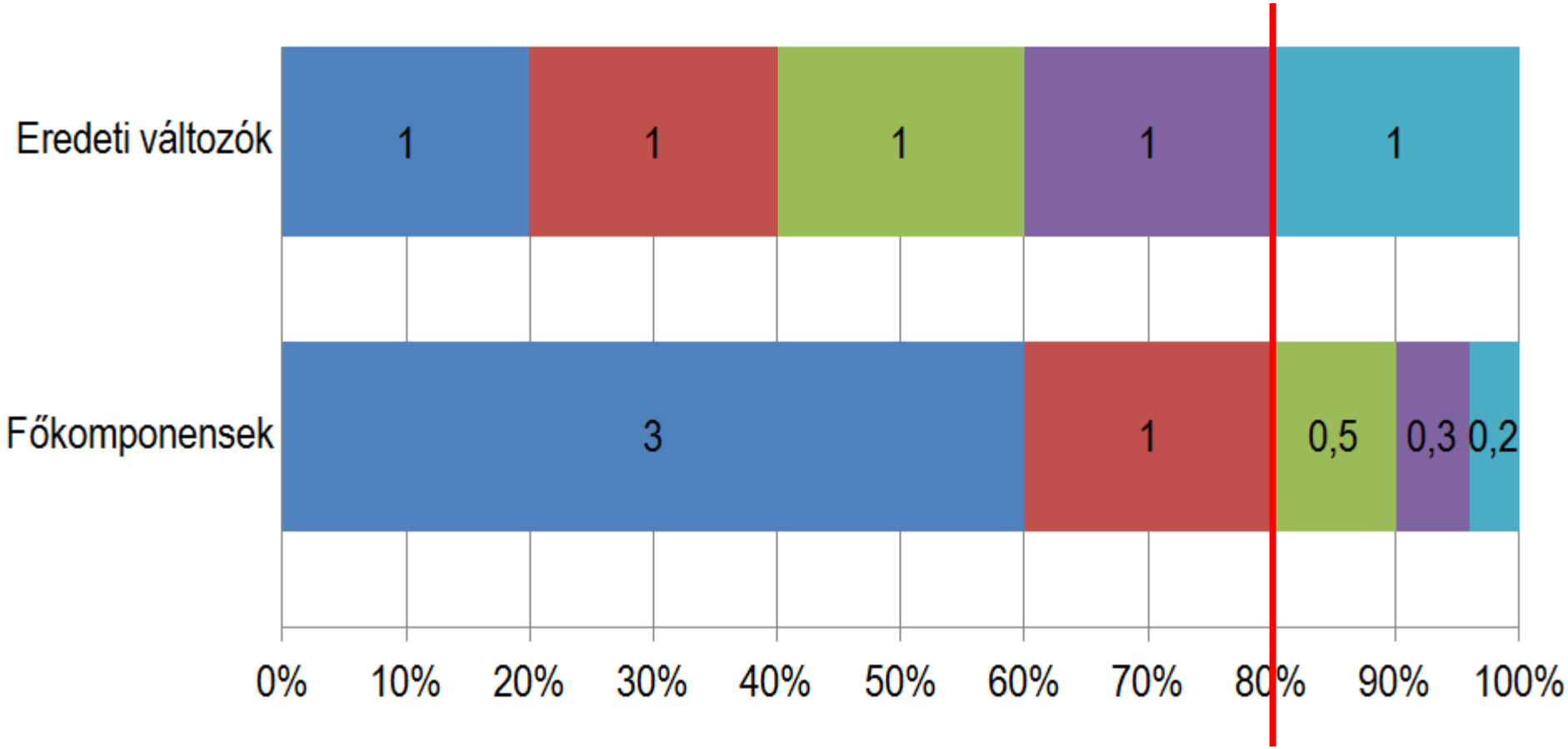
Matematikai háttér

- A p változó korrelációs mátrixának előállítjuk a p darab sajátértékét. Ezek összege mindig p-vel azonos.
- A legnagyobb sajátérték adja meg a legfontosabb főkomponens varianciáját, a második legnagyobb sajátérték a második legfontosabb főkomponensét, stb.

Matematikai háttér

- A legfontosabb főkomponens úgy áll elő, hogy a legnagyobb sajátértékhez tartozó, egységnyi normájú sajátvektorban szereplő együtthatókkal vesszük a sztenderdizált változók lineáris kombinációját.
- A második legfontosabb főkomponenshez a második legnagyobb sajátértékhez tartozó, egységnyi normájú sajátvektorban szereplő együtthatókat használjuk, stb.

Példa: 5 változó → 2 főkomponens, a megőrzött variancia 80%

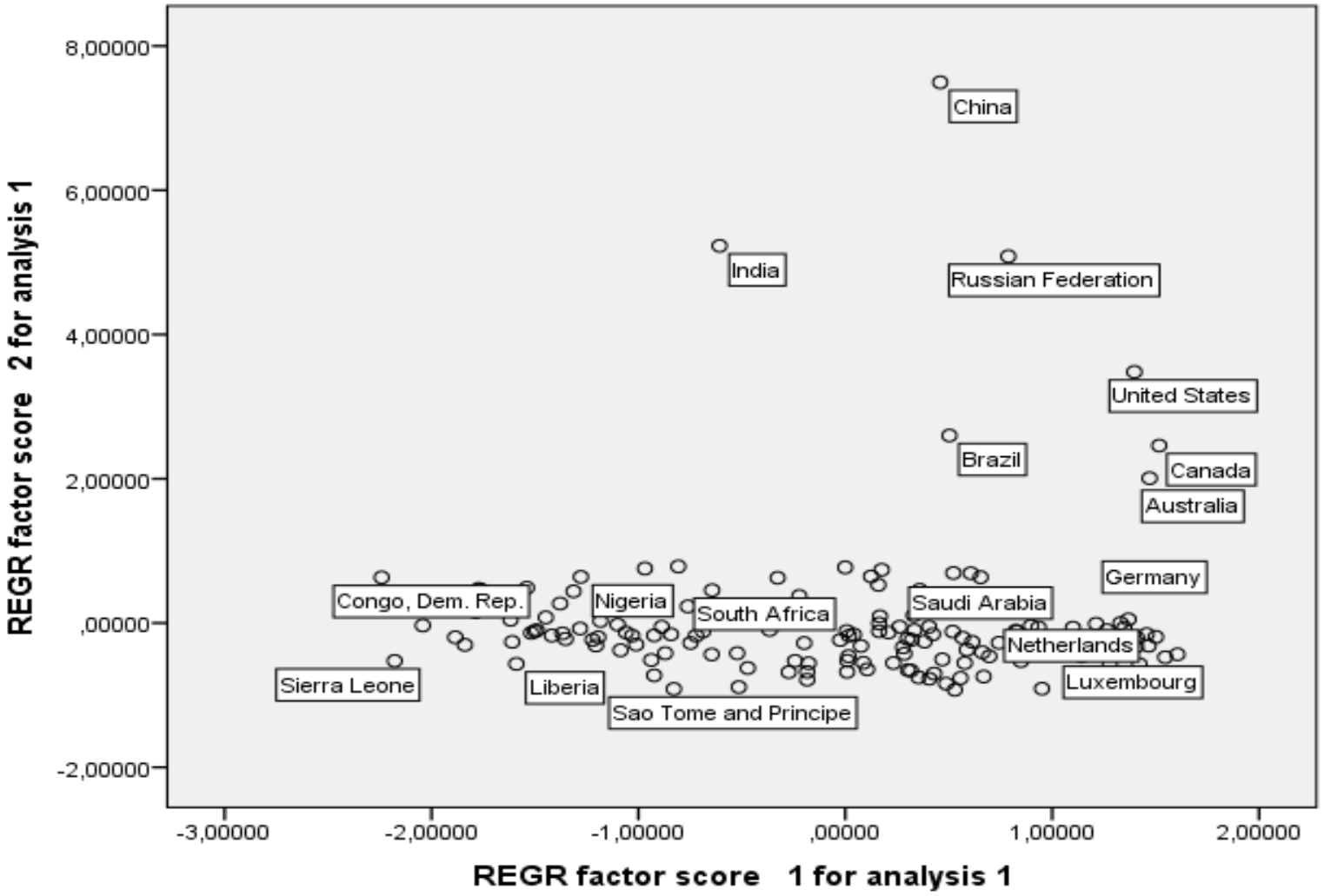


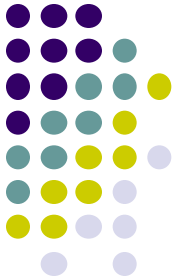
Főkomponensek értelmezése

- **Loadingok:**

  Az együtthatók amelyekkel kell venni a sztenderdizált változók lineáris kombinációját a főkomponensek előállításához.
- **Főkomponensek-változók közti korrelációk:**
  Ezek alapján is látható a főkomponensek tényleges jelentése.
- **Kommunalitás:**
  Az adott változó dimenziócsökkentés után megőrzött %-os információtartalma.
- **Tiszta főkomponens-struktúra:**

Példa: országok fejlettsége és mérete (főkomponensek)





Köszönöm a figyelmet!