

Inspiration: PCA-metoden og mælkeproduktion i Uruguay

Baggrund

I 2014 offentliggjorde forskere fra Universidad de la República i Uruguay et studie af 24 mælkeproducerende gårde i den sydlige del af landet. Målet var at undersøge, hvordan forskellige driftsformer påvirker mælkens *carbon footprint* – altså hvor meget CO₂ der udledes pr. kg produceret mælk.

Forskernes data indeholdt mange variable: mælkeydelse pr. ko, antal køer pr. hektar, andel af kraftfoder, kvælstofudledning osv. For at forstå sammenhængene mellem de mange tal brugte de en statistisk metode kaldet **Principal Component Analysis (PCA)**¹.

Idéen bag PCA

Når vi måler mange forskellige størrelser på én gang, kan det være svært at danne sig et overblik. Hver gård i undersøgelsen kan beskrives som et punkt i et højdimensionelt rum, hvor hver akse svarer til én variabel (f.eks. én akse for mælkeydelse pr. ko, én for fodermængde osv.).

PCA søger at finde nye akser (kaldet *hovedkomponenter*), som forklarer mest mulig af variationen i data. Disse nye akser er lineære kombinationer af de oprindelige variable, dvs. at de dannes ved at "veje" de gamle akser sammen.

En intuitiv forklaring

Forestil dig, at man har målt to variable for hver gård: (1) mælkeydelse pr. ko og (2) fodermængde pr. ko. Hvis man tegner punkterne i et koordinatsystem, vil de måske ligge langs en skrå ret linje. PCA finder netop denne retning som *første hovedkomponent* — den retning hvor punkterne varierer mest.

Den anden hovedkomponent står vinkelret på den første og beskriver variationen, som ikke blev forklaret af den første. På den måde reducerer man et komplekst datasæt til få, letforståelige dimensioner.

¹Data stammer fra: https://www.researchgate.net/publication/269989556_Practices_to_Reduce_Milk_Carbon_Footprint_on_Grazing_Dairy_Farms_in_Southern_Uruguay_Case_Studies (Practices to Reduce Milk Carbon Footprint on Grazing Dairy Farms in Southern Uruguay: Case Studies)

PCA i undersøgelsen

I den uruguayanske undersøgelse blev otte variable valgt: mælkeydelse, mælk pr. hektar, antal køer pr. hektar, foderforbrug m.m. Forskerne fandt, at to hovedkomponenter kunne forklare hele 86% af variationen mellem gårdene:

- Den første komponent (72%) var forbundet med mælkeydelse og fodereffektivitet.
- Den anden (14%) beskrev forskelle i arealanvendelse og kvælstofudledning.

Ved at se på et plot af gårdene i disse to nye akser kunne forskerne se mønstre: nogle gårde udnyttede foder og areal langt mere effektivt end andre.

PCA og lineær algebra

Selv om man ikke behøver kunne udføre beregningerne i hånden, bygger PCA på centrale ideer fra lineær algebra:

- Vektorer repræsenterer observationer (gårde) i et rum.
- Retninger, hvor data varierer mest, findes ved at beregne *egenvektorer* til en såkaldt kovariansmatrix.
- Disse retninger bliver til de nye koordinataksler (hovedkomponenter).

I praksis lader man computeren gøre arbejdet, men bag metoden gemmer sig et geometrisk billede af vektorer, vinkler og projektioner.

Opgave / Refleksion

- a) Forklar med dine egne ord, hvorfor det kan være nyttigt at beskrive hver gård som et punkt i et koordinatsystem.
- b) Tegn et tænkt datasæt med to variable (f.eks. mælkeydelse og fodermængde). Marker, hvordan PCA's første akse kunne se ud.
- c) Hvorfor tror du, at PCA kan være et nyttigt redskab i arbejdet med bæredygtig landbrugsproduktion?
- d) Overvej: Hvis du kun måtte vælge to variable til at beskrive gårdenes klimaaftryk, hvilke ville du vælge – og hvorfor?

Perspektiv

PCA bruges i mange felter: fra økonomi og biologi til billedgenkendelse og kunstig intelligens. Når man reducerer et datasæt fra mange til få dimensioner, får man et overblik over de vigtigste mønstre. Metoden viser, hvordan matematiske idéer om vektorer og retninger kan omsættes til praktiske værktøjer i moderne dataanalyse.