

10 Analýza hlavních komponent (PCA)

Příklad 1. Datový soubor `cneck.txt` obsahuje antropometrické údaje mladých dospělých lidí (převážně studentů vysokých škol z Brna a Ostravy). U jedinců známe následující hodnoty: identifikační číslo pozorování (proměnná `id`), pohlaví (proměnná `sex`), tělesná hmotnost (proměnná `body.W`, v *kg*), tělesná výška (proměnná `body.H`, v *mm*), obvod pasu (proměnná `waist.C`, v *mm*), obvod boků (proměnná `hip.C`, v *mm*), obvod předloktí (proměnná `antb.C`, v *mm*) a obvod krku (proměnná `neck.C`, v *mm*). Analyzujte spojitě proměnné pomocí metody hlavních komponent:

1. Prozkoumejte závislost mezi spojitými veličinami pomocí dvourozměrného podového diagramu.
2. Vykreslete krabicové diagramy pro tyto proměnné.
3. Vypočítejte korelační matici. Otestujte hypotézu o úplné nezávislosti proměnných.
4. Proveďte analýzu hlavních komponent na základě korelační matice.
5. Zjistěte podíl variability a kumulativní podíl variability pro jednotlivé komponenty.
6. Kolik komponent by vybralo Kaiserovo kritérium? Kolik by jich bylo vybráno podle zploštění sutinového grafu? Kolik by jich bylo vybráno, pokud bychom požadovali vysvětlení alespoň 80 % variability? Při další práci se omezte na tento počet.
7. Podívejte se na korelace původních proměnných s těmito komponentami. Pokuste se komponenty vhodně interpretovat.
8. Vykreslete pozorování a proměnné v rovině prvních dvou komponent. Místo čísel označte pozorování pohlavím jedince (stačí značka `f` nebo `m` z proměnné `sex`).
9. Vypočítejte reprodukovanou korelační matici a reziduální korelační matici. Pozorujete vysoké nebo nízké reziduální hodnoty?