4.7 Vícerozměrné škálování MDS

Vícerozměrné škálování (**M**ulti**D**imensional **S**caling, MDS) je technika vytvoření diagramu relativního umístění objektů v rovině dvojrozměrného grafu na základě dat vzdáleností mezi objekty, tzv. *matice proximity* (blízkosti). Diagram může obsahovat jeden, dva, tři a zřídka i více rozměrů, dimenzí. Technika vyčíslí metrické klasické (CMDS) nebo nemetrické (NNMDS) řešení a vychází buď přímo z experimentálních hodnot $\boldsymbol{X}$, z korelační matice $\boldsymbol{R}$ nebo z matice podobností $\boldsymbol{S}$ či vzdáleností $\boldsymbol{D}$. Vzdálenost mezi oběma objekty je Eukleidovská, počítaná na základě Pythagorovy věty, $d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$, kde m je počet proměnných a x_{ik} jsou data i-tého řádku a k-tého sloupce. I když vynášíme vzdálenosti do dvojrozměrného grafu, může být d_{ij} vyčísleno na základě většího počtu proměnných $m \geq 2$. Matice vzdáleností je potom trojúhelníková a zajímá nás jenom její horní část. S růstem objektů však roste i počet dimenzí, takže pro *tři* objekty je to *dvoj*-rozměrná rovina, pro *čtyři* objekty pak *troj*-rozměrný prostor atd.

Kritérium maximální věrohodnosti. Jak těsně prokládá model vzdáleností daná experimentální data se hodnotí *testem těsnosti proložení* s využitím statistického kritéria *stress*, založeného na rozdílu mezi skutečnou vzdáleností d_{ij} a modelem predikovanou hodnotou $\hat{d}_{ij}$,

$$stress = \sqrt{\frac{\sum_{k=1}^{m} (d_{ij} - \hat{d}_{ij})^2}{\sum_{k=1}^{m} d_{ij}^2}},$$

kde $\hat{d}_{ij}$ je predikovaná vydálenost, založená na MDS modelu. Predikovaná hodnota závisí především na počtu užitých dimenzí a algoritmu, a to metrickém či nemetrickém. Je-li *stress* číslo nízké, blízké nule, jeví se MDS proložení jako nejlepší.

Počet dimenzí. Důležitým úkolem v MDS je určení počtu dimenzí v MDS modelu. Každá dimenze zde představuje latentní proměnnou. Cílem MDS je udržet počet dimenzí na co možná nejmenší hodnotě. Obvykle volí uživatel dvoj- maximálně trojrozměrný prostor. Vychází-li vyšší počet dimenzí, není MDS technika k analýze dotyčných dat vhodná. Počet dimenzí se volí na základě co nejmenší hodnoty kritéria *stress*. Někteří autoři si pomáhají indexovým grafem relativní velikosti vlastních čísel, která jsou vyčíslována pro rostoucí počet dimenzí, tzv. *grafem úpatí*. Postup a inter-pretace jsou pak stejné jako u metod PCA nebo FA.

Vstupní data. Data mohou být trojího typu, mohou obsahovat (1) vzdálenosti mezi objekty $\boldsymbol{D}$, (2) podobnost mezi objekty $\boldsymbol{S}$ nebo (3) hodnoty proměnných (sloupce) pro jednotlivé objekty (řádky) $\boldsymbol{X}$.

Vzdálenost (disimilarita) d_{ij} představující vzdálenost mezi objekty, může být měřena přímo, jako např. vzdálenost dvou měst. MDS užívá vzdálenost v datech přímo a matice vzdáleností $\boldsymbol{D}$ je symetrická.

Podobnost (similarita) s_{ij} vyjadřuje, jak blízko se nacházejí dva objekty. MDS umožňuje načíst míry podobnosti pro každý pár objektů. Matice podobností $\boldsymbol{S}$ je opět symetrická. Podobnost lze konvertovat do veličiny vzdálenosti vzorcem

$$d_{ij} = \sqrt{s_{ii} + s_{jj} - 2s_{ij}},$$

kde d_{ij} představuje vzdálenost a s_{ij} podobnost.

Hodnoty x_{ij} proměnných pro jednotlivé objekty představují spíše standardní míry. Z nich se vypočte nejprve korelační matice $\boldsymbol{R}$ a potom matice Eukleidovských či Mahalanobisových vzdáleností $\boldsymbol{D}$.

Klasická metrická metoda MDS. Je dána matice vzdáleností $\boldsymbol{D}$, která vystihuje meziobjektové vzdálenosti objektů $\boldsymbol{X}$ v prostoru spíše nižšího rozměru dle vzorce

$$d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}.$$

Jednotlivé kroky klasické MDS jsou následující:

1. Z $\boldsymbol{D}$ se vypočte $\boldsymbol{A} = \{-0.5\, d_{ij}^2\}$.
2. Z $\boldsymbol{A}$ se vypočte $\boldsymbol{B} = \{a_{ij} - a_{i.} - a_{.j} + a_{..}\}$, kde $a_{i.}$ je průměr všech a_{ij} přes j.
3. Nalezne se m největších vlastních čísel $\lambda_1 > \lambda_2 > ... > \lambda_m$ matice $\boldsymbol{B}$ a odpovídající vlastní vektory $\boldsymbol{L} = \boldsymbol{L}_{(1)}, \boldsymbol{L}_{(2)}, ..., \boldsymbol{L}_{(m)}$, které jsou normovány, takže $\boldsymbol{L}^{\mathrm{T}}_{(i)}\, \boldsymbol{L}_{(i)} = \lambda_i$. Předpokládáme, že m je voleno tak, že vlastní hodnoty jsou relativně velké a kladné.
4. Souřadnicemi objektů jsou řádky matice $\boldsymbol{L}$.

Klasické řešení je optimalizováno metodou nejmenších čtverců: přímé řešení $\boldsymbol{L}$ minimalizuje sumu čtverců vzdáleností mezi skutečnými prvky matice $\boldsymbol{D}$, tj. d_{ij} a predikcemi $\hat{d}_{ij}$, založenými na $\boldsymbol{L}$. Předpokládejme, že experimentální hodnoty vzdálenosti d_{ij} jsou zatíženy náhodnou chybou ε_{ij} dle vzorce $d_{ij} = \delta_{ij} + \varepsilon_{ij}$, kde ε_{ij} představuje kombinaci náhodných chyb z měření, distorze vzdáleností, když MDS model zcela neodpovídá konfiguraci navržených m vzdáleností. Navrhněme model závislosti mezi vzdáleností dvou objektů vztahem $\hat{d}_{ij} = \beta_0 + \beta_1 * _{ij} + \varepsilon_{ij}$ a potom nalezením nejlepších odhadů b_0 pro β_0 a b_1 pro β_1 obdržíme odhad vypočtené vzdálenosti $\hat{d}_{ij} = b_0 + b_1 * _{ij}$. Optimalizační procedura vychází z účelové funkce

$$U = \sum_{i<j}^{n} (d_{ij} - \hat{d}_{ij})^2 \;.\; \min.$$

Aby byla zajištěna úplná invariantnost vůči transformaci proměnných, užívá se modifikovaná účelová funkce U_{mod} dle vztahu

$$U_{\mathrm{mod}} = \frac{\sum_{i<j}^{n} (d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j}^{n} d_{ij}^2}$$

a především její druhá odmocnina, zvaná *stress* $= \sqrt{U_{\mathrm{mod}}}$. Je proto výhodné hledat optimální počet dimenzí, která se vezmou k vyčíslení predikce MDS vzdálenosti $\hat{d}_{ij}$ pomocí minimální hodnoty veličiny *stress*. Pro *stress* < 0.05 je těsnost proložení ještě přijatelná a pro *stress* < 0.01 je těsnost proložení výtečná.

Nemetrická MDS. V dosavadnín postupu se předpokládalo, že vzdálenosti jsou vyčísleny metricky. Jsou však situace, kdy jedna hodnota nevystihuje dostatečně skutečnost: např. při porovnávání barev na stupnici může být jedna barva zářivější než druhá, a tento fakt však nikterak neovlivní polohu barvy na stupnici. Predikované vzdálenosti $\hat{d}_{ij}$ jsou vyčíslovány *monotónní regresí:* experimentální vzdálenosti jsou uspořádány vzestupně do řady

$$d_{i1,j1} \le d_{i2,j2} \le ... \le d_{iN,jN}, \text{ kde } N = n(n - 1)/2$$

a $\hat{d}_{ij}$ jsou odhadovány tak, aby splnily podmínku *slabé monotonicity (WM)*

$$\hat{d}_{i1,j1} \le \hat{d}_{i2,j2} \le \ldots \le \hat{d}_{iN,jN}, \text{ nebo}$$

nebo podmínku *silné monotonicity (SM)*

$$\hat{d}_{i1,j1} < \hat{d}_{i2,j2} < \ldots < \hat{d}_{iN,jN}.$$

Prvním krokem k získání počátečních odhadů predikovaných vzdáleností $\hat{d}_{ij}$ bývá vždy metrické vyčíslení. Pak následuje nemetrický přístup monotónní regrese. Indexový graf úpatí veličiny *stress* je užitečnou pomůckou i u nemetrické metody. Hledá se jednak zlom na tomto grafu a jednak se vyšetřuje, kdy veličina *stress* nabyde hodnot menších než 0.05, resp. 0.01. Takový index, čili počet dimenzí, se pak jeví jako optimální. Obdobně, jako metrická metoda CMDS, ústí i nemetrická NNMDS ve vícerozměrnou škálovací mapu, na které se sleduje roztřídění vyšetřovaných objektů.