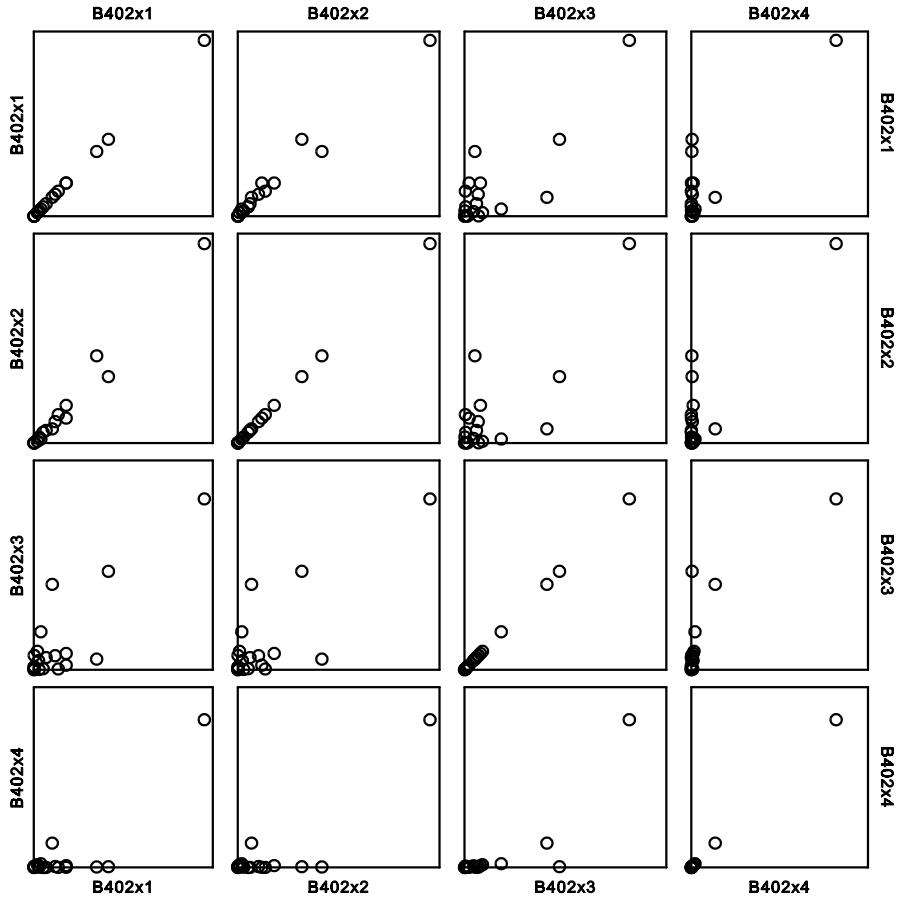
4.4 Exploratorní analýza struktury objektů (EDA)

Průzkumová analýza vícerozměrných dat je stejně jako u jednorozměrných dat založena na vyšetření grafických diagnostik. K tomuto účelu se využívá různých technik zobrazování vícerozměrných dat. Pro případ, kdy jsou jednotlivé sloupce matice $\boldsymbol{X}$ málo korelované postačují *rozptylové diagramy* pro jednotlivé kombinace složek vektoru $\boldsymbol{x}$ a pro nekorelované pak sloupce matice $\boldsymbol{X}$.



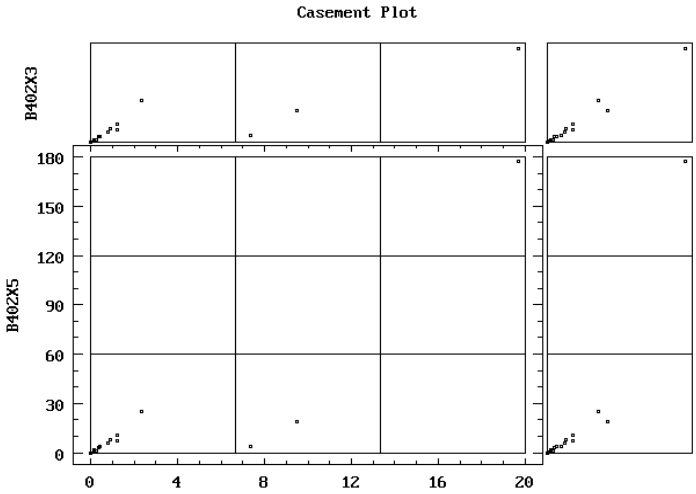
Obr. 4.1 Rozptylový diagram pro 20 objektů a 4 proměnné *B402X1, B402X2, B402X3, B402X4* nestandardizovaných dat B402, ***SCAN.*** *Je patrná podobnost objektů a vysoká korelovanost zejména prvních dvou proměnných. Jsou patrné i odlehlé objekty, představované body vzdálenými od ostatních*.

Rychlé posouzení podobnosti mezi jednotlivými objekty čili řádky datové matice usnadňují především *symbolové grafy.* Jednotlivé proměnné jsou v nich "kódovány" s ohledem na jejich konkrétní hodnoty do určitých geometrických tvarů, *symbolů*. Každému objektu $\boldsymbol{x}_i$ (např. autu) tak odpovídá jistý obrazec zvaný *symbol*. Vlastnosti dat se posuzují s ohledem na vizuální rozdíly mezi symboly. Tím lze v jednom grafu rozlišit *více proměnných* x_j , $j = 1, ..., m$. Prvním krokem před vlastním zobrazením do symbolů je obvykle standardizace. Mezi základní typy zobrazovaných symbolů patří *profily*, *polygony*, *tváře*, *křivky* a *stromy*.

Profily představují dvourozměrné zobrazení m-rozměrných objektů. Každý objekt $\boldsymbol{x}_i$ je charakterizován m proměnnými, zobrazenými zde vertikálními úsečkami. Jejich velikost je úměrná hodnotě odpovídající proměnné x_{ij} , $j = 1, ..., m$. Profil pak vzniká spojením koncových bodů těchto úseček. Je vhodné použít standardizované proměnné dle vzorce

$$x_{ij}^{(} \quad , \quad \frac{x_{ij}}{(\max_i {}^* x_{ij}{}^*)} \qquad ,$$

kde $\max^* x_{ij}{}^*$ je maximální hodnota absolutní velikosti proměnné $\boldsymbol{x}_j$ vektoru $\boldsymbol{x}_i^T$ přes všechny body, $i = 1, ..., n$. Profily jsou jednoduché a umožňují snadné určení rozdílů mezi jednotlivými objekty $\boldsymbol{x}_i$ a $\boldsymbol{x}_k$. Snadno lze takto identifikovat vybočující objekt.



Obr. 4.2 Korelační diagram (Casement Plot) pro 20 objektů a 4 proměnné *B402X1, B402X2, B402X3, B402X4* nestandardizovaných dat B402, ***STATGRAPHICS.*** *Je patrná vysoká korelovanost čtyř sledovaných proměnných. V pravém horním rohu jsou patrné odlehlé objekty*.

Polygony jsou vlastně profily v polárních souřadnicích, kdy každá proměnná objektu $\boldsymbol{x}_i^{\mathrm{T}}$, $i = 1, ..., n,$ odpovídá délce paprsku vycházejícího ze společného středu. Paprsky dělí kružnici ekvidistantně, proměnné jsou standardizovány do intervalu
[0, 1]. Mezi polygony patří *graf slunečních paprsků* a *hvězdicový graf.*

(a) **Graf slunečních paprsků** má tvar "sluníčka", které se skládá z paprsků, začínajících ve společném bodě, a úseček spojujících paprsky, které tak tvoří polygon. Zde každá proměnná x_{ij} objektu $\boldsymbol{x}_i^{\mathrm{T}}$ odpovídá délce paprsku vycházejícího ze středu sluníčka. Paprsky jsou rozmístěny ekvidistantně, ve stejných vzdálenostech na kružnici, a proto se provádí lineární transformace do intervalu [a, 1], kde a je zvolená spodní mez, obyčejně $a = 0$. Pro tuto transformaci platí, že

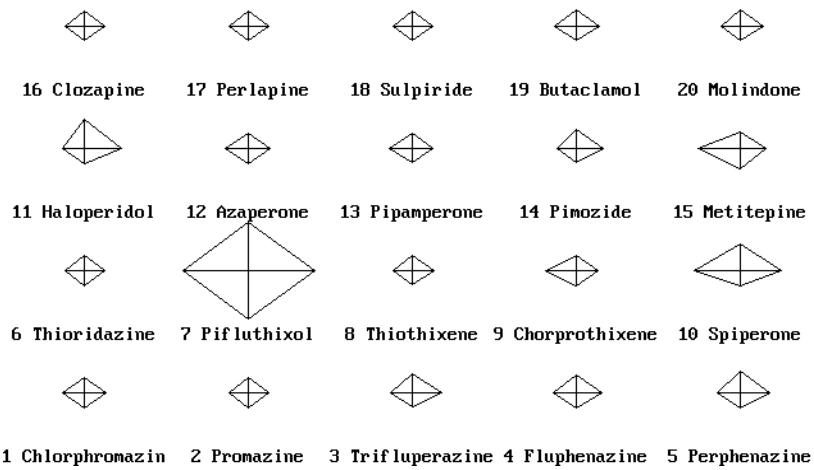
$$x_{ij}^{(} \quad ' \quad \frac{(1 \ \& \ a)\,(x_{ij} \ \& \ \min\limits_{i} x_{ij})}{\max\limits_{i} x_{ij} \ \& \ \min\limits_{i} x_{ij}} \ \% \ a \quad ,$$

kde $\min x_{ij}$ je minimální a $\max x_{ij}$ maximální hodnota j-té proměnné objektu $\boldsymbol{x}_i^{\mathrm{T}}$ přes všechny objekty $\boldsymbol{x}_i^{\mathrm{T}}$, $i = 1, ..., n$. K určení směrů jednotlivých paprsků se definuje jejich úhel α_j, pro který platí

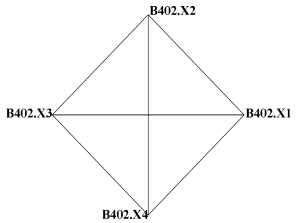
$$\alpha_j \quad ' \quad \frac{2\,\pi\,(j \ \& \ 1)}{m} \quad , \qquad j \ ' \ 1, \ ..., \ m \quad .$$

Za společný střed paprsků se obyčejně volí počátek souřadnic. Pokud má být maximální délka paprsků rovna R, je polygon pro objekt $\boldsymbol{x}_i^{\mathrm{T}}$ spojnicí m bodů p_{ij} o souřadnicích $p_{ij} \quad ' \quad (x_{ij}^{(} R \cos "_j \ , \ x_{ij}^{(} R \sin "_j)$. Aby vznikl uzavřený obrazec, spojují se ještě první a poslední bod p_{i1} a p_{im}. Vzájemné porovnání polygonů slouží k vizuálnímu posouzení podobnosti objektů. V případě velkého počtu proměnných, např. $m > 6$, bývá však výsledný obrázek polygonů nepřehledný.

(b) **Hvězdicový graf** vypadá na první pohled jako předchozí graf sluníčka. Sestává z paprsků, reprezentujících relativní hodnoty proměnných u jednotlivých objektů, které se pro každý objekt spojují v jednom centrálním bodě. Stejně směřující paprsky u různých objektů se liší svojí délkou. *Nejkratší paprsek* indikuje, že u objektu nabývá příslušná proměnná nejmenší hodnoty z celého výběru. Podobně *nejdelší paprsek* informuje o nejvyšší hodnotě příslušné proměnné. Délky ostatních paprsků se pohybují podle relativní velikosti hodnot proměnné u příslušného objektu mezi těmito dvěma krajními mezemi.

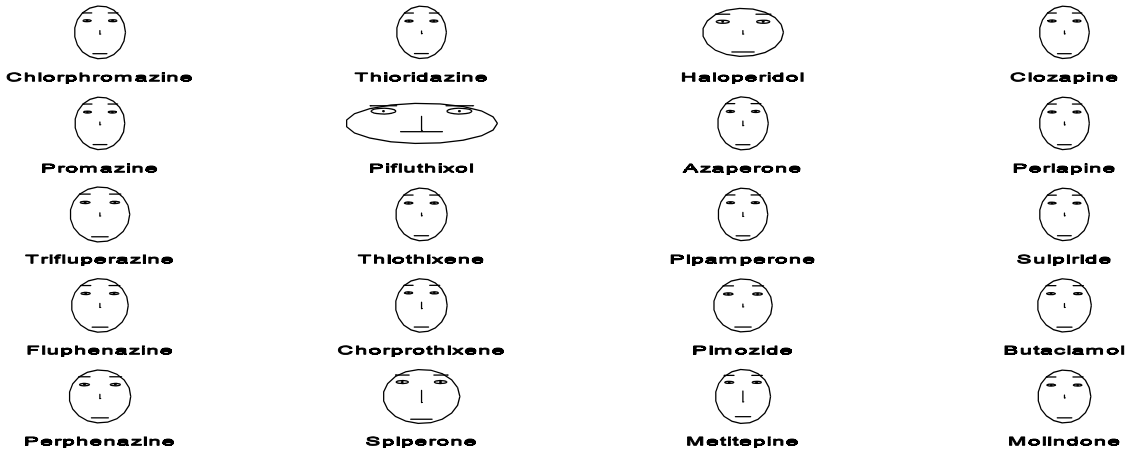


Obr. 4.3a Hvězdičkový graf (Stars Plot) pro 20 objektů a 4 proměnné *B402X1, B402X2, B402X3, B402X4* standardizovaných dat B402, ***STATGRAPHICS.***

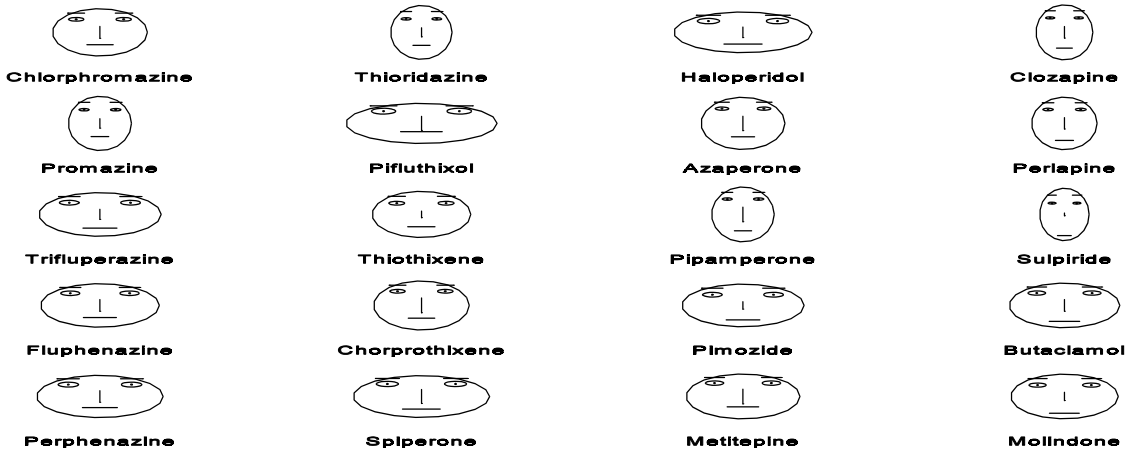


Obr. 4.3b Klíč ke hvězdičkovém graf pro 4 proměnné *B402X1, B402X2, B402X3, B402X4* standardizo-vaných dat B402, ***STATGRAPHICS.***

Tváře charakterizují každou proměnnou x_{ij} objektu $\boldsymbol{x}_i^{\mathrm{T}}$ nějakým znakem. Mezi znaky patří tvar tváře, délka nosu, velikost očí, tvar úst atp. Tvar tváře závisí na použitém pořadí proměnných, které ovlivňuje snadnost interpretace dat.



Obr. 4.4 Tváře nestandardizovaných dat B402 pro 20 objektů a 4 proměnné *B402X1, B402X2, B402X3, B402X4,* ***S-Plus.*** Lze nalézt řadu vzájemně podobných tváří, ukazujících na podobnost objektů. Tvář Pifluthixolu se jeví silně odlišná od ostatních.



Obr. 4.5 Tváře zlogaritmovaných dat B402 pro 20 objektů a 4 proměnné *B402X1, B402X2, B402X3, B402X4*, ***S-Plus.*** *Logaritmováním dat se rozdíly mezi objekty poněkud setřou a odlehlé objekty nejsou při porovnání podobnosti tak výrazné odlišné. Tvář Pifluthixolu se však stále jeví odlišná od ostatních.*

Křivky využívají transformace každého objektu $\boldsymbol{x}_i^{\mathbf{T}}$ do spojité křivky, která je lineární kombinací všech jeho proměnných. Andrews[12] volí pro vyjádření křivky f_i odpovídajícího objektu $\boldsymbol{x}_i^{\mathbf{T}}$ konečnou Fourierovu řadu

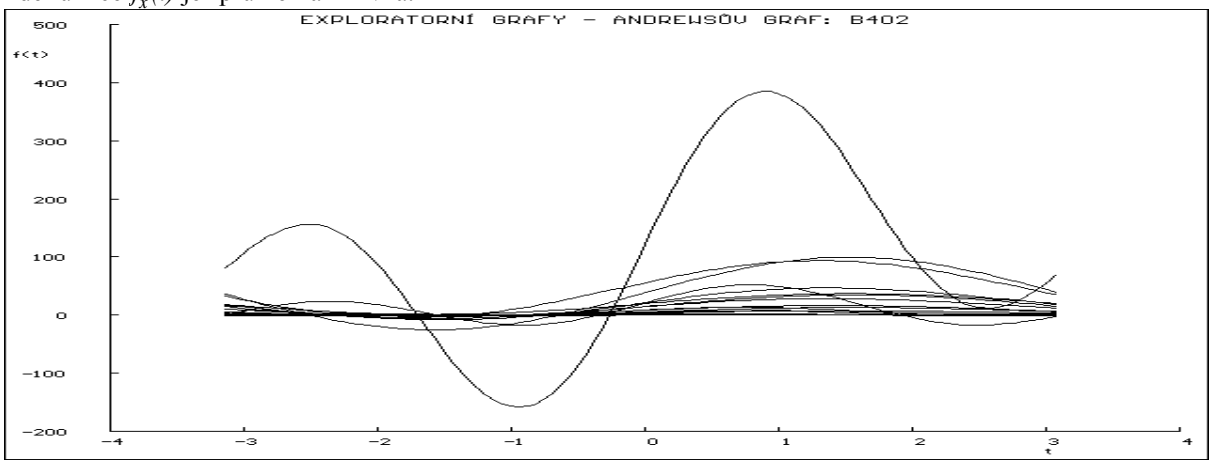
$$f_{\boldsymbol{x}_i}(t) \;=\; f_i \;=\; \frac{x_{i1}}{\sqrt{2}} \;+\; x_{i2}\sin(t) \;+\; x_{i3}\cos(t) \;+\; x_{i4}\sin(2\,t) \;+\; x_{i5}\cos(2\,t) \;+\; \ldots$$

Křivky f_i, $i = 1, \ldots, n$, se vynášejí jako funkce proměnné t v intervalu $-\pi \# t \# \pi$. Funkce f_i mají řadu výhodných vlastností:

a) Funkce f_i zachovávají *průměr*. To znamená, že pokud je $\bar{\boldsymbol{x}}$ průměrem z celkového počtu n vícerozměrných dat $\boldsymbol{x}_i$, je funkce rovna

$$f_{\bar{x}}(t) \;=\; \frac{1}{n}\sum_{i=1}^{n} f_{\boldsymbol{x}_i}(t),$$

kde funkce $f_{\bar{x}}(t)$ je "průměrná" křivka.



Obr. 4.6 Andrewsův graf křivek dat pro 20 objektů a 4 proměnné *B402X1, B402X2, B402X3, B402X4* nestandardizovaných dat B402, ***S-Plus***, *Graf kaz je na značno podobnost celé řady objektů. Jeden objekt je však výrazně odlišný od ostatních, jde o odlehlý objekt*.

b) Funkce f_i zachovávají *vzdálenosti*. To znamená, že celková vzdálenost mezi křivkami f_i a f_j, definovaná jako integrální kvadratická odchylka, odpovídá vzdálenosti mezi objekty $\boldsymbol{x}_i^{\mathbf{T}}$ a $\boldsymbol{x}_j^{\mathbf{T}}$. Blízké křivky ukazují na nepříliš vzdálené objekty.

c) Pro zvolenou hodnotu t_0 je funkce $f_{\boldsymbol{x}_i}(t_0)$ projekcí objektu $\boldsymbol{x}_i$ na vektor $\boldsymbol{p}_0$ o složkách

$$\boldsymbol{p}_0 \;=\; \left(\frac{1}{\sqrt{2}},\ \sin(t_0),\ \cos(t_0),\ \sin(2t_0),\ \cos(2t_0),\ \ldots\right).$$

Tato projekce do jednoho bodu umožňuje odhalení vybočujících objektů či skupin objektů, které mohou být ve více dimenzích špatně identifikovatelné. Křivka $f_{\boldsymbol{x}_i}(t)$ je složena ze všech projekcí na daném intervalu hodnot t.

d) Funkce f_i zachovávají *rozptyl*. To znamená, že pokud jsou proměnné x_j objektu $\boldsymbol{x}_i^{\mathbf{T}}$ nekorelované náhodné veličiny se stejným rozptylem σ^2, je

$$D(f_i) \;=\; \sigma^2\,(0.5 \;+\; \sin^2(t) \;+\; \cos^2(t) \;+\; \sin^2(2\,t) \;+\; \cos^2(2\,t) \;+\; \ldots).$$

Pro liché m je $D(f_i) = 0.5\ \sigma^2$ m a pro sudé m je $0.5\ \sigma^2\ (m - 1) < D(f_i) < 0.5\ \sigma^2\ (m + 1)$. Rozptyl funkce f_i je téměř konstantní v celém rozmezí veličiny t.

V praktických úlohách je běžné, že jednotlivé proměnné jsou silně korelované a mají nestejné rozptyly. Pak je výhodné převést objekty původních dat $\boldsymbol{x}_i$ na objekty $\boldsymbol{y}_i$, kde y_{ij} odpovídá transformaci do j-té hlavní komponenty. Veličiny y_{ij} jsou již nekorelované. Snadno lze provést i jejich standardizaci tak, aby měly konstantní rozptyly. Nevýhodou křivek je to, že jejich tvar závisí na pořadí složek. Na druhé straně lze pomocí křivek snadno indikovat vybočující objekty nebo skupiny objektů a konstruovat i konfidenční křivky. Pro větší počty objektů ($n > 10$) dochází ke splývání křivek, což ztěžuje jejich interpretaci. Pak je možné vynášet pouze zvolené podskupiny objektů.

Stromy čili **dendrogramy** jsou vhodné pro případy, kdy je počet proměnných m objektu $\boldsymbol{x}_i^{\mathbf{T}}$ veliký. Jednotlivé složky $\boldsymbol{x}_j$ představují délku větví schematického stromu. Jeho struktura větví vzniká na základě předběžného hierarchického shlukování proměnných (viz *shluková analýza*). Předběžná shluková analýza se dá použít také při výběru pořadí složek objektu $\boldsymbol{x}$ při konstrukci ostatních symbolových grafů