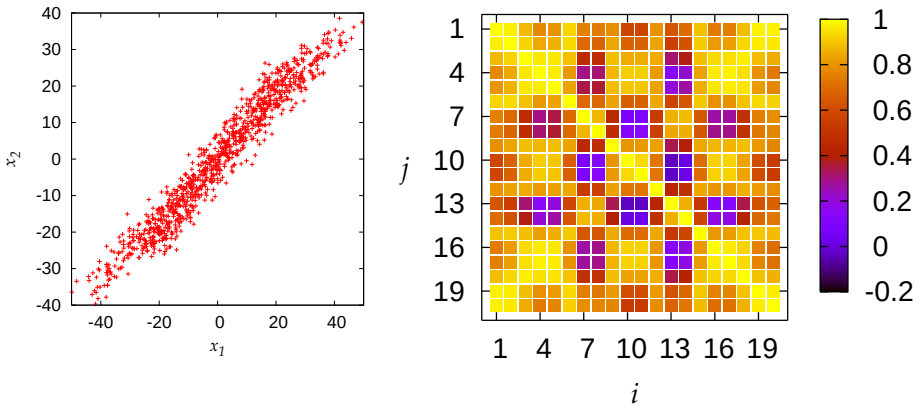
Zadání

Cílem tohoto cvičení je ukázat Vám, jak se používá PCA (Principal Component Analysis) k analýze a dimenzionální redukci dat.



Obrázek 1: Na levé straně jsou vyneseny do grafu první dva sloupce dat, vpravo je matice korelačních koeficientů $C_{ij}/sqrtC_{ii}C_{jj}$.

V souboru `d1k.dat` je po řádcích zapsáno 1000 dvacetirozměrných náhodných vektorů. Je zjevné (viz obrázek), že mezi jednotlivými komponentami vektorů existují silné **lineární** korelace. To znamená, že v datech je pravděpodobně ukryta nějaká jednodušší struktura. Prvním krokem k jejímu odhalení je nalézt takovou transformaci dat, která zachová všechnu obsaženou informaci a zároveň odstraní lineární korelace. Z nekonečně mnoha způsobů jak toho dosáhnout je PCA nejjednodušší a nejelegantnější, protože transformace kterou používá je lineární a navíc ortogonální.

Vaším prvním úkolem je nalézt výše popsanou transformaci. Označme řádky souboru d1k.dat $\vec{x}_{(k)}$, kde $k = 1..n = 1000$. Lineární korelace v datech jsou vyjádřeny kovarianční maticí

$$C = \frac{1}{n}\sum_{k=1}^{n}(\vec{x}_{(k)} - \vec{x}_{\mathrm{m}})(\vec{x}_{(k)} - \vec{x}_{\mathrm{m}})^{\mathrm{T}}, \tag{1}$$

kde $\vec{x}_{\mathrm{m}}$ je aritmetický průměr vektorů $\vec{x}_{(k)}$,

$$\vec{x}_{\mathrm{m}} = \frac{1}{n}\sum_{k=1}^{n}\vec{x}_{(k)}.$$

Matice C je symetrická a pozitivně (semi)definitní, takže její vlastní vektory tvoří ortogonální bázi ve které je C diagonální. Vpraxi můžete buď nejprve spočítat matici C a použít na ni některou z mnoha numerických metod pro výpočet vlastního systému čtvercové symetrické matice, nebo můžete aplikovat metodu SVD přímo na data.

Označme X matici, která má v řádcích zapsané vektory $\vec{x}_{(k)} - \vec{x}_{\mathrm{m}}$. Pomocí maticového násobení můžeme rovnici (1) přepsat do tvaru

$$C = \frac{1}{n}XX^{\mathrm{T}}. \tag{2}$$

Metoda SVD (Singular Value Decomposition) dokáže obecnou (reálnou) obdélníkovou matici A rozložit na součin tří matic $A = U\Sigma V^{\mathrm{T}}$, kde U a V jsou ortogonální matice a Σ je obdélníková matice mající nenulové σ_k prvky pouze na hlavní diagonále. Aplikujeme-li SVD na matici $X/sqrtn$ dostaneme pro matici C vztah

$$U^{\mathrm{T}} C U = \Sigma\Sigma^{\mathrm{T}}. \tag{3}$$

Matice $\Sigma\Sigma^{\mathrm{T}}$ a U tedy zjevně obsahují vlastní čísla a vlastní vektory C.

Získali jsme tedy, jedním nebo druhým způsobem, vlastní systém autokovariančí matice,

$$C\vec{v}_j = \lambda_j \vec{v}_j, \quad \vec{v}_i \cdot \vec{v}_j = \delta_{ij}.$$

Datové vekrory $\vec{x}_{(k)}$ můžeme vyjádřit v bázi $\vec{v}_j$ jako,

$$\vec{x}_{(k)} = \sum_{j=1}^{d} \lambda_j \xi_j^{(k)} \vec{v}_j, \tag{4}$$

kde d je dimenze prostoru dat a $\xi_j^{(k)}$ jsou po dvou nekorelovaná náhodná čísla s nulovou střední hodnotou a s jednotkovou disperzí,

$$\langle \xi_j^{(k)} \rangle = 0, \quad \langle \xi_i^{(k)} \xi_j^{(k)} \rangle = \delta_{ij}. \tag{5}$$

Zjevně různé vlastní vektory $\vec{v}_j$ přispívají k součtu (6) různě významně. Protože však mají koeficienty $\xi_i^{(k)}$ stejné momenty (do druhého řádu) závisí to, jak moc přispěje daný vlastní vektor k celkové varianci dat zavisí pouze na vlastních číslech λ_j. Často se stává, že několik málo vlastních čísel matice C je výrazně větších, než zbytek. Bez újmy na obecnosti můžeme předpokládat, že λ_j jsou seřazena od největšího k nejmenšímu. Pak můžeme místo původních dat studovat částečný součet

$$\vec{x}_{\mathrm{f}(k)} = \sum_{j=1}^{K} \lambda_j \xi_j^{(k)} \vec{v}_j, \tag{6}$$

nebo prvních K koeficientů $\xi_j^{(k)}$ (případně $\lambda_j \xi_j^{(k)}$), kde $K < d$ je vhodně zvolený index. První přístup odpovídá filtraci dat, druhý je dimenzionální redukce. Místo původních dat tak studujeme několik nejvýznamnějších komponent (6), odtud název metody.

Úkoly:

1) Najděte vlastní vektory a vlastní čísla matice C dat ze souboru `d1k.dat`. Použijte buď SVD, nebo některou z metod pro symetrické čtvercové matice z Numerical Recipes, nebo nějaký jiný vhodný software (třeba Matlab).

2) Prohlédněte si vlatní čísla, vyberte si K největších a pro všechna k spočítejte příslušné koeficienty $\lambda_j \xi_j^{(k)}$. (Jak? Díky ortonormalitě vlastních vektorů C platí, že $\lambda_j \xi_j^{(k)} = \vec{x}_{(k)} \cdot \vec{v}_j$.) Jaké K je vhodné musíte určit experimentálně. Ovšem příklad je umělý, takže má poměrně hezké řešení. V prvních několika komponentách se ukrývá struktura (byť náhodná) a ve zbytku je jen čistý šum.

3) Zkuste určit jaká struktura se v datech ukrývá a případně ji i nějak objektivně popište. (Třeba na ní něco nafitujte pomocí metody nejmenších čtverců.)

4) Soubor `d1k.dat` obsahuje prvních 1000 z 10000 řádek ze souboru `d10k.dat`. Zkuste zopakovat analýzu i na tomto větším souboru dat.