BI-VZD přednáška 10

Alexander Kovalenko

FIT ČVUT

25. 1. 2022

Autoři: Karel Klouda, Juan Pablo Maldonado Lopez, Daniel Vašata.
Problémy, návrhy apod. hlaste v GitLabu.
Verze souboru: 25. dubna 2022 10:14.

Co bude v dnešní přednášce

- Redukce dimenzionality
- Ortogonální projekce - metoda hlavních komponent (PCA)
- Manifold learning - lokálně lineární vnoření (LLE)

Problematika vysokých dimenzí

Z předchozích přednášek víme, že:

- U jednotkové krychle v d rozměrném prostoru s rovnoměrným rozdělením bodů se v podkrychli o hraně délky $(1-\varepsilon)$ nalézá $(1-\varepsilon)^d \cdot 100\ \%$ všech bodů.
- Pro velké d je toto číslo blízké nule, což znamená, že se většina bodů nachází v okrajové slupce tlusté $\varepsilon/2$ a tedy velmi blízko hranici krychle.
- Například pro $d = 100$ se v okrajové slupce tloušťky 0.01 nachází 86.7% bodů.
- Dva body ve stejné jednotkové krychli jsou tedy typicky hodně vzdálené.
- Z pohledu strojového učení to znamená, že hustota trénovacích bodů s rostoucí dimenzí klesá. Body jsou od sebe velmi vzdálené.
- Vytváření predikcí je tak méně spolehlivé, protože model musí dělat velké extrapolace.

Problematika vysokých dimenzí

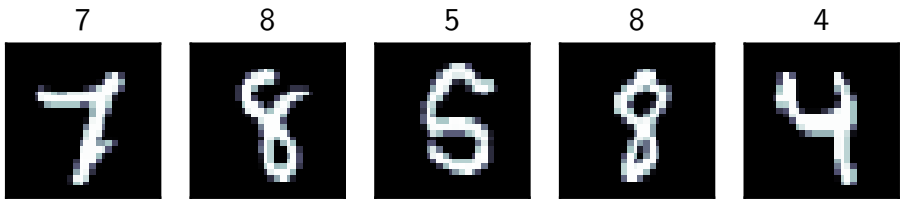
- Co použít větší množství dat? Teoreticky ano. Co je ale dostatečně velký počet?
- Pro data se 100 příznaky potřebujeme 10^{100} bodů k tomu, abychom zajistili, že v jednotkové krychli budou body cca 0.1 daleko od sebe.
- To ale není úplně málo, když vezmeme v úvahu, že odhadovaný počet atomů (protonů) ve vesmíru[1] je 10^{80}.
- Je tedy zřejmé, že pokud by data byla v prostoru vysoké dimenze „hodně“ rozprostřena, může to být velký problém.
- Jak se ale ukazuje, ve většině reálných situací tomu tak není a data se vyskytují pouze v omezených oblastech prostoru menší dimenze.
- Z matematického pohledu tak často předpokládáme, že se vyskytují podél takzvaných variet (angl. manifold), což jsou nelineární obdoby lineárních variet.
- Proto se tento předpoklad nazývá **hypotéza variet** (angl. **manifold hypothesis**).

[1] Známý též jako Eddingtonovo číslo.

Manifold hypothesis (hypotéza variet)

Většina reálných mnohorozměrných dat je ve skutečnosti rozprostřena podél variet mnohem menší dimenze.

- Uvažujme opět známý MNIST dataset s rukou psanými číslicemi zachycenými na černobílých fotkách o rozměrech 28×28 pixelů ve stupních šedi.

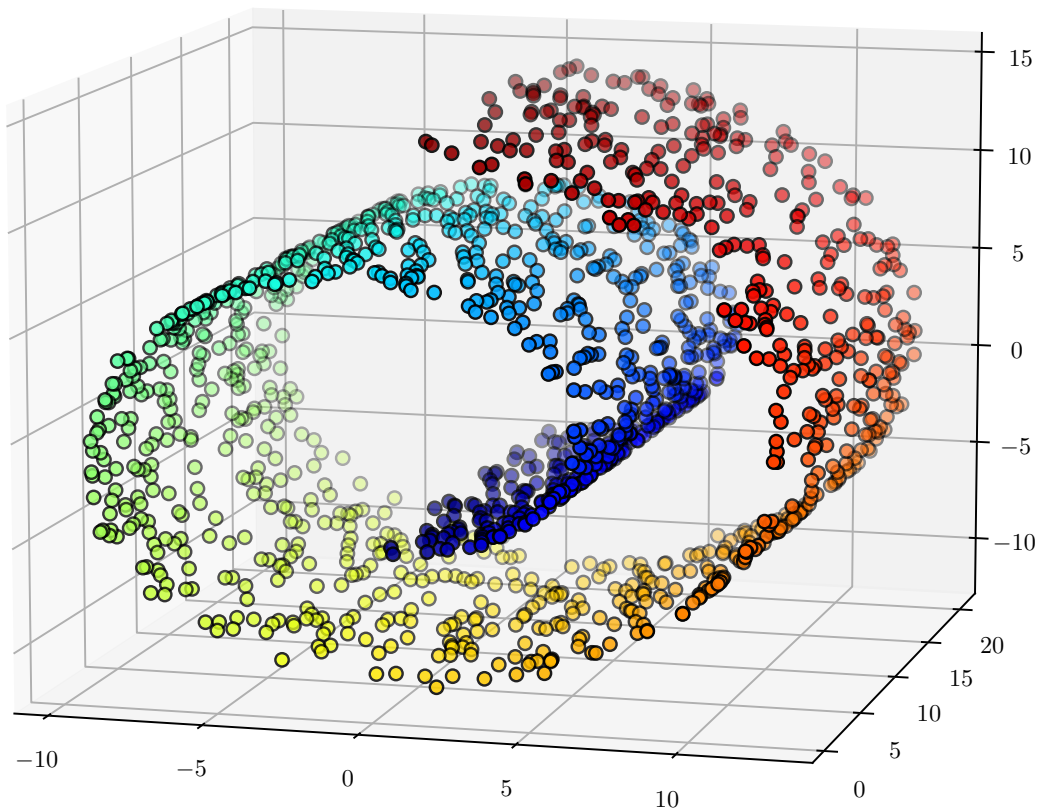


- Datový bod tedy obsahuje $28 \cdot 28 = 784$ příznaků s hodnotami od 0 do 255.

- Pokud takový obrázek nakreslíme náhodně (např. rovnoměrně), je pravděpodobnost toho, že dostaneme něco podobného číslici, extrémně malá. Data jsou evidentně rozprostřena pouze v malinké oblasti prostoru.

Manifold hypothesis - další příklad

Dalším příkladem je dataset nazývaný **švýcarská rolka** (angl. **swiss roll**).



Zde se data vyskytují podél dvourozměrné spirálovitě zatočené oblasti.

Redukce dimenzionality

- Z hypotézy variet plyne, že dává smysl se pokoušet najít vhodné zobrazení dat z původního prostoru do nového prostoru menší dimenze.
- Z pohledu linearity takového zobrazení můžeme používané metody rozdělit na lineární projekce a manifold learning.
- **Lineární projekce** jsou víceméně vždy ortogonální projekce na nějaký lineární podprostor, případně na lineární varietu.
- **Manifold learning** reprezentují nelineární metody, které se snaží zobecnit lineární přístup k zachycení obecné nelineární struktury v datech.
- Výstupem takových metod je nelineární mapování bodů z původního prostoru na prostor menší dimenze.

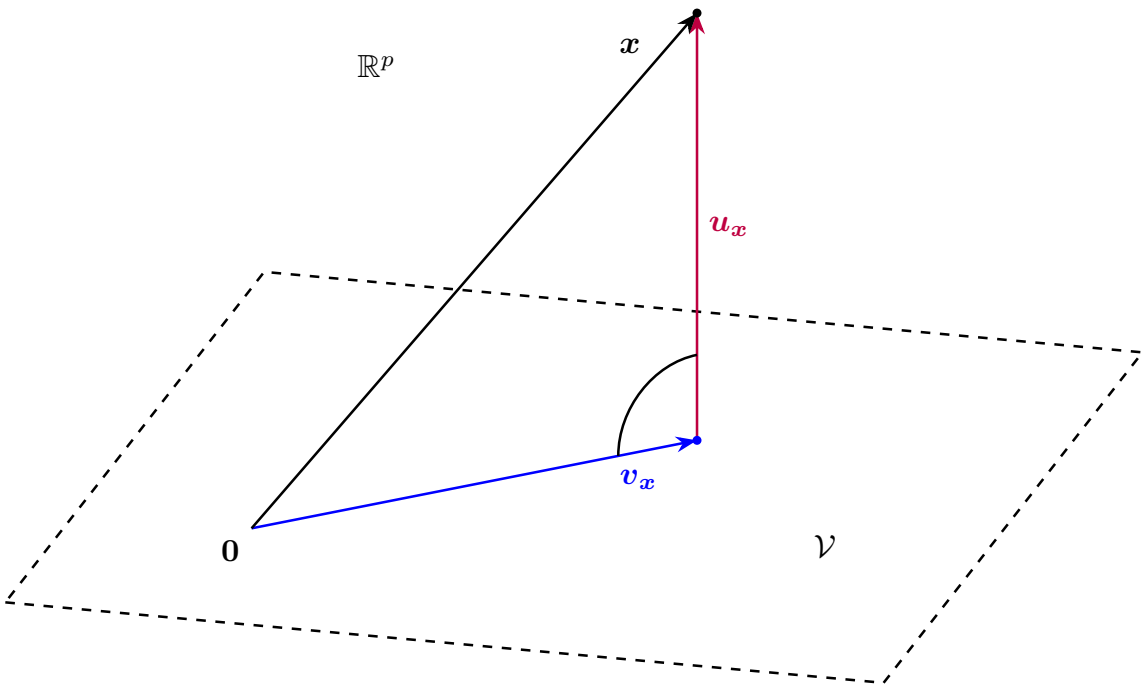
Ortogonální projekce na vektorový podprostor

Zabývejme se ortogonálními projekcemi na nějaký q dimenzionální vektorový podprostor $\mathcal{V}$ vektorového prostoru $\mathbb{R}^p$ vybaveného standardním skalárním součinem.

Z lineární algebry víme, že každý bod $\boldsymbol{x} \in \mathbb{R}^p$ je možné jednoznačně rozložit do součtu

$$\boldsymbol{x} = \boldsymbol{v_x} + \boldsymbol{u_x},$$

kde $\boldsymbol{v_x}$ je bod vektorového prostoru $\mathcal{V}$ a $\boldsymbol{u_x}$ je vektor kolmý na $\mathcal{V}$.



Bod $\boldsymbol{v_x}$ se nazývá **ortogonální projekce** bodu $\boldsymbol{x}$ na podprostor $\mathcal{V}$.

Ortogonální projekce pomocí ortonormální báze

Uvažujme nyní situaci, že máme ortonormální bázi prostoru $\mathbb{R}^p$ tvořenou vektory $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_p$ tak, že prvních q vektorů $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_q$ tvoří bázi podprostoru $\mathcal{V}$ a zbylé vektory $\boldsymbol{b}_{q+1}, \ldots, \boldsymbol{b}_p$ jsou kolmé na $\mathcal{V}$.

Každý bod vektorového prostoru $\boldsymbol{x} \in \mathbb{R}^p$ můžeme jednoznačně vyjádřit jako

$$\boldsymbol{x} = \tau_1 \boldsymbol{b}_1 + \ldots + \tau_q \boldsymbol{b}_q + \tau_{q+1} b_{q+1} + \ldots + \tau_p \boldsymbol{b}_p,$$

kde i-tý koeficient τ_i je určen skalárním součinem $\boldsymbol{x}$ a $\boldsymbol{b}_i$, tj. $\tau_i = \boldsymbol{x}^T \boldsymbol{b}_i$.

Pro ortogonální rozklad $\boldsymbol{x} = \boldsymbol{v_x} + \boldsymbol{u_x}$ tedy dostáváme

$$\boldsymbol{v_x} = \tau_1 \boldsymbol{b}_1 + \ldots + \tau_q \boldsymbol{b}_q \quad \text{a} \quad \boldsymbol{u_x} = \tau_{q+1} b_{q+1} + \ldots + \tau_p \boldsymbol{b}_p.$$

Projekci $\boldsymbol{v_x}$ bodu $\boldsymbol{x}$ na podprostor $\mathcal{V}$ můžeme jednoznačně reprezentovat q rozměrným vektorem koeficientů $\boldsymbol{t_x} = (\tau_1, \ldots, \tau_q)^T \in \mathbb{R}^q$.

Maticově můžeme tento vektor získat vztahem

$$\boldsymbol{t_x} = \mathbf{V}^T \boldsymbol{x}, \quad \text{případně} \quad \boldsymbol{t_x}^T = \boldsymbol{x}^T \mathbf{V},$$

kde $\mathbf{V} \in \mathbb{R}^{p,q}$ je matice, v jejíchž sloupcích jsou zapsané vektory $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_q$ ortonormální báze $\mathcal{V}$.

Poznamenejme, že $\mathbf{V}^T \mathbf{V} = \mathbf{I}_q$, tj. identita v $\mathbb{R}^q$.

Redukce dimenzionality pomocí ortogonální projekce

Reprezentace bodu $\boldsymbol{x} \in \mathbb{R}^p$ pomocí bodu $\boldsymbol{t_x} \in \mathbb{R}^q$ představuje redukci dimenzionality z p na q.

Pro dataset reprezentovaný maticí $\mathbf{X} \in \mathbb{R}^{N,p}$, která obsahuje N bodů o p příznacích zapsané v řádcích získáme transformovaný dataset vztahem

$$\mathbf{T}_q = \mathbf{X}\mathbf{V}.$$

Matice $\mathbf{T}_q \in \mathbb{R}^{N,q}$ tedy obsahuje q příznaků pro N bodů, které představují souřadnice ortogonálních projekcí těchto bodů na podprostor $\mathcal{V}$ v jeho ortonormální bázi $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_q$.

Kdybychom chtěli získat odpovídající projekce (tj. body v původním prostoru), musíme provést ještě jedno vynásobení maticí $\mathbf{V}$:

- Pro jeden bod

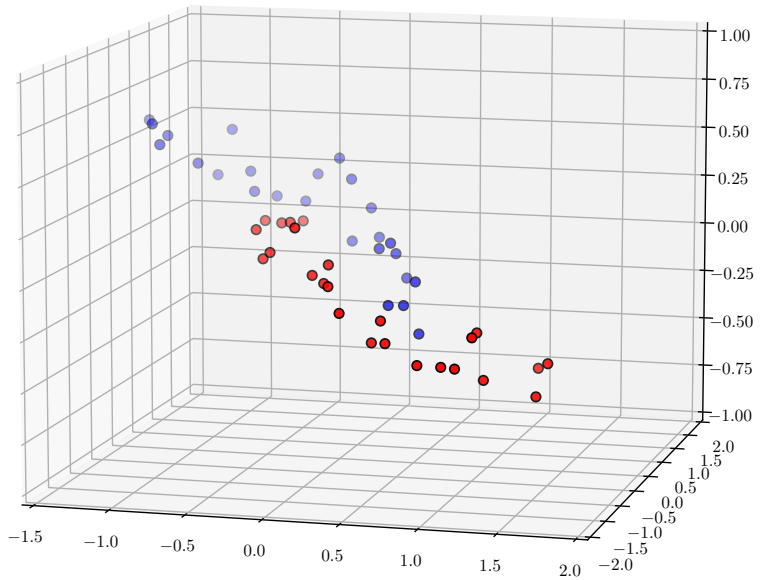
$$\boldsymbol{v_x} = \mathbf{V}\boldsymbol{t_x} = \mathbf{V}\mathbf{V}^T\boldsymbol{x}, \quad \text{případně} \quad \boldsymbol{v_x}^T = \boldsymbol{x}^T\mathbf{V}\mathbf{V}^T.$$

- Pro celý dataset

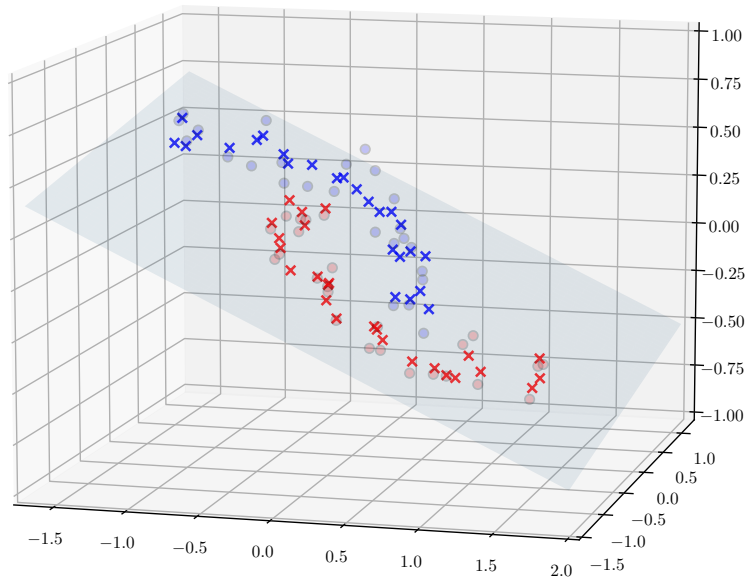
$$\mathbf{X}_\mathcal{V} = \mathbf{X}\mathbf{V}\mathbf{V}^T,$$

kde výsledná matice $\mathbf{X}_\mathcal{V} \in \mathbb{R}^{N,p}$ představuje ortogonální projekce původních bodů na podprostor $\mathcal{V}$, tj. ve sloupcích je p originálních příznaků.

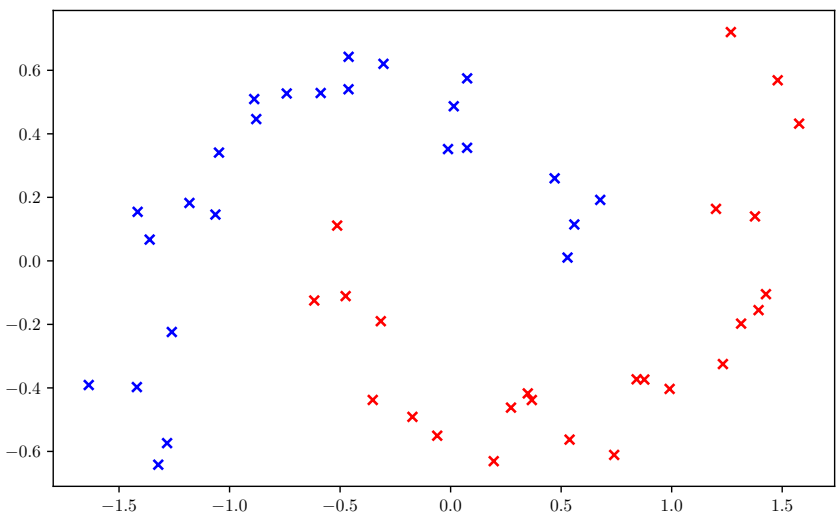
Vizualizace ortogonální projekce



Originální dataset $\mathbf{X}$



Projekce $\mathbf{X}_{\mathcal{V}}$ datasetu na podprostor $\mathcal{V}$



Transformovaný dataset $\mathbf{T}_q$

Ortogonální projekce - pomocná pozorování

Než se podíváme na metodu hlavních komponent, vraťme se ještě k rozkladu $\boldsymbol{x} = \boldsymbol{v_x} + \boldsymbol{u_x}$, kde vektor $\boldsymbol{v_x}$ můžeme získat pomocí matice $\mathbf{V}$ jako $\boldsymbol{v_x} = \mathbf{V}\mathbf{V}^T\boldsymbol{x}$.

Označme dále matici $\mathbf{U} \in \mathbb{R}^{p,p-q}$ v jejíchž sloupcích jsou zapsané vektory $\boldsymbol{b}_{q+1}, \ldots, \boldsymbol{b}_p$.

Analogicky jako u matice $\mathbf{V}$ platí

$$\boldsymbol{u_x} = \mathbf{U}\mathbf{U}^T\boldsymbol{x}$$

a $\mathbf{U}^T\mathbf{U} = \mathbf{I}_{p-q}$ je identita na $\mathbb{R}^{p-q}$.

Rozklad $\boldsymbol{x} = \boldsymbol{v_x} + \boldsymbol{u_x}$ tedy můžeme získat pomocí matic $\mathbf{V}$ a $\mathbf{U}$ jako

$$\boldsymbol{x} = \mathbf{V}\mathbf{V}^T\boldsymbol{x} + \mathbf{U}\mathbf{U}^T\boldsymbol{x}.$$

Pro kvadrát normy navíc z ortogonality rozkladu platí následující vztahy:

$$\begin{aligned}\|\boldsymbol{x}\|^2 &= \boldsymbol{x}^T\boldsymbol{x} = (\boldsymbol{v_x} + \boldsymbol{u_x})^T(\boldsymbol{v_x} + \boldsymbol{u_x}) = \boldsymbol{v_x}^T\boldsymbol{v_x} + \boldsymbol{u_x}^T\boldsymbol{u_x} + \boldsymbol{v_x}^T\boldsymbol{u_x} + \boldsymbol{u_x}^T\boldsymbol{v_x} \\ &= \|\boldsymbol{v_x}\|^2 + \|\boldsymbol{u_x}\|^2 = \boldsymbol{x}^T\mathbf{V}\mathbf{V}^T\mathbf{V}\mathbf{V}^T\boldsymbol{x} + \boldsymbol{x}^T\mathbf{U}\mathbf{U}^T\mathbf{U}\mathbf{U}^T\boldsymbol{x} = \boldsymbol{x}^T\mathbf{V}\mathbf{V}^T\boldsymbol{x} + \boldsymbol{x}^T\mathbf{U}\mathbf{U}^T\boldsymbol{x}.\end{aligned}$$

Metoda hlavních komponent

Otázkou zůstává, jak pro každé q najít podprostor $\mathcal{V}$, na který budeme projekci dělat.

Rozumným požadavkem na hledanou metodu je, aby pro každé q minimalizovala kvadratickou chybu projekce datasetu $\mathbf{X}$ na q rozměrný podprostor $\mathcal{V}$.

Jak ukážeme, k získání optimálního minima je nejprve nutné provést středování datasetu a získat tak dataset $\mathbf{X}'$, který má v řádcích napsané body $\boldsymbol{x}'_i = \boldsymbol{x}_i - \bar{\boldsymbol{x}}$, kde $\bar{\boldsymbol{x}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i$ je výběrový průměr datasetu $\mathbf{X}$.

Při ortogonálním rozkladu $\boldsymbol{x}'_i = \boldsymbol{v}_{\boldsymbol{x}'_i} + \boldsymbol{u}_{\boldsymbol{x}'_i}$ bodu $\boldsymbol{x}'_i$ tedy chceme minimalizovat

$$\sum_{i=1}^{N} \|\boldsymbol{x}'_i - \boldsymbol{v}_{\boldsymbol{x}'_i}\|^2 = \sum_{i=1}^{N} \|\boldsymbol{u}_{\boldsymbol{x}'_i}\|^2.$$

Řešení spočívá v užití ortonormální báze tvořené vlastními vektory $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_p$ příslušejícími k vlastním číslům matice $\frac{1}{N-1}\mathbf{X}'^T\mathbf{X}'$ seřazeným sestupně podle velikosti.

Podprostor $\mathcal{V}$ a příslušná matice $\mathbf{V}$ jsou pak tvořeny prvními q vektory této báze.

Tento postup se nazývá **metoda hlavních komponent** (angl. **principal component analysis**) (PCA).

Význam středování datasetu

Z přednášky o shlukování víme, že výraz

$$\sum_{i=1}^{N} \|\boldsymbol{x}_i - \boldsymbol{\mu}\|^2$$

nabývá minima, pokud $\mu = \bar{\boldsymbol{x}}$, což je geometrický střed množiny bodů v datasetu.

Zároveň pro libovolný podprostor $\mathcal{V}$ a příslušné matice $\mathbf{V}$ a $\mathbf{U}$ platí

$$\bar{\boldsymbol{u}} = \tfrac{1}{N} \sum_{i=1}^{N} \boldsymbol{u}_{\boldsymbol{x}_i} = \tfrac{1}{N} \sum_{i=1}^{N} \mathbf{U}\mathbf{U}^T \boldsymbol{x}_i = \mathbf{U}\mathbf{U}^T \bar{\boldsymbol{x}}.$$

Protože $\boldsymbol{u}_{\boldsymbol{x}_i} = \mathbf{U}\mathbf{U}^T \boldsymbol{x}_i$ a $\boldsymbol{u}_{\boldsymbol{x}'_i} = \mathbf{U}\mathbf{U}^T(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) = \boldsymbol{u}_{\boldsymbol{x}_i} - \bar{\boldsymbol{u}}$ nabývá

$$\sum_{i=1}^{N} \|\boldsymbol{u}_{\boldsymbol{x}_i} - \bar{\boldsymbol{u}}\|^2 = \sum_{i=1}^{N} \|\boldsymbol{u}_{\boldsymbol{x}'_i}\|^2$$

minima z pohledu možných translací datasetu $\mathbf{X}$.

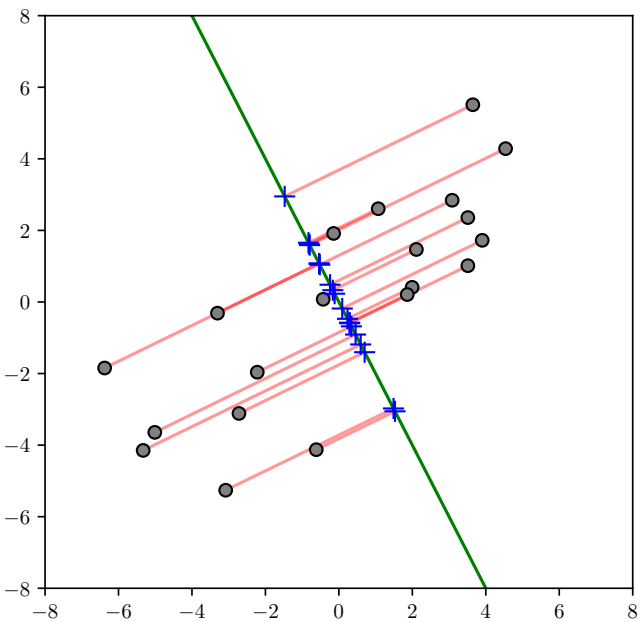
Proto, nezávisle na $\mathcal{V}$, vede středování k nejmenší možné kvadratické chybě projekce posunutého datasetu $\mathbf{X}'$ na $\mathcal{V}$.

Vztah kvadratické chyby a rozptylu

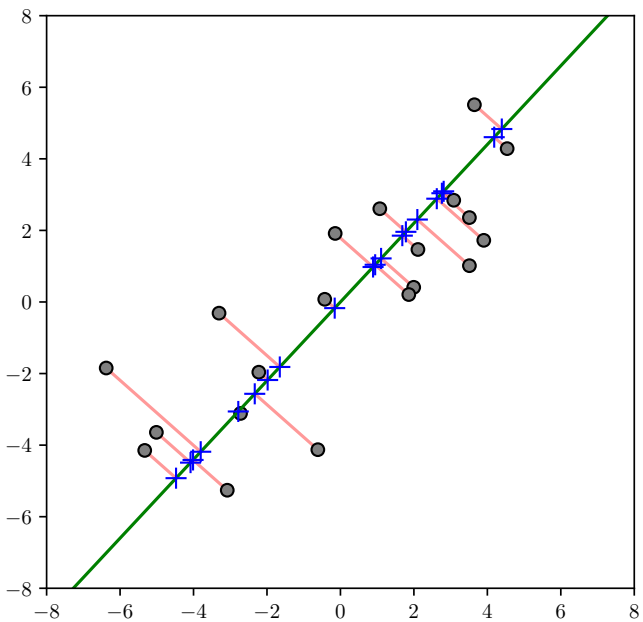
Protože pro normu při ortogonálním rozkladu platí $\|\boldsymbol{x}'_i\|^2 = \|\boldsymbol{v}_{\boldsymbol{x}'_i}\|^2 + \|\boldsymbol{u}_{\boldsymbol{x}'_i}\|^2$ a tedy

$$\frac{1}{N-1}\sum_{i=1}^{N}\|\boldsymbol{x}'_i\|^2 = \frac{1}{N-1}\sum_{i=1}^{N}\|\boldsymbol{v}_{\boldsymbol{x}'_i}\|^2 + \frac{1}{N-1}\sum_{i=1}^{N}\|\boldsymbol{u}_{\boldsymbol{x}'_i}\|^2,$$

odpovídá minimalizace chyb projekce maximalizaci „rozptylu“ projektovaných bodů ve $\mathcal{V}$.



Projekce s větší chybou



Projekce s menší chybou

Varianční matice

- Označme p příznaků v prostoru, ze kterého pochází data, jako $X_1, \ldots, X_p$.
- Dataset $\mathbf{X}$ můžeme chápat jako náhodný výběr z rozdělení těchto příznaků.
- Složky $\bar{x}_1, \ldots, \bar{x}_p$ vektoru $\bar{x}$ jsou tak výběrovými průměry a tedy bodovými odhady středních hodnot příznaků.
- **Výběrová kovariance** příznaků X_i a X_j, která je odhadem $\mathrm{cov}(X_i, X_j)$, je tedy

$$\widehat{\mathrm{cov}}(X_i, X_j) = \frac{1}{N-1} \sum_{k=1}^{N} (x_{k;i} - \bar{x}_i)(x_{k;j} - \bar{x}_j) = \frac{1}{N-1} \sum_{k=1}^{N} x'_{k;i} x'_{k;j} = \frac{1}{N-1} (\mathbf{X'}^T \mathbf{X'})_{ij}.$$

- **Varianční matice** (angl. **covariance matrix**) je matice jejíž (i, j)-tá složka je $\mathrm{cov}(X_i, X_j)$.
- **Výběrová varianční matice** (angl. **sample covariance matrix**) je matice jejíž (i, j)-tá složka je $\widehat{\mathrm{cov}}(X_i, X_j)$ a je vlastně bodovým odhadem varianční matice.
- Z předešlého tedy platí, že výběrová varianční matice je $\frac{1}{N-1} \mathbf{X'}^T \mathbf{X'}$.
- V dalším textu budeme pro jednoduchost používat pojem varianční matice i pro tuto výběrovou varianční matici.
- Varianční matice je symetrická a na diagonále má nezáporné hodnoty (výběrové rozptyly).

Optimalita PCA

Jelikož je varianční matice symetrická, je také diagonalizovatelná a z příslušných vlastních vektorů můžeme sestavit ortonormální bázi prostoru $\mathbb{R}^p$. Protože je také pozitivně semi-definitní (viz přednáška o lineární regresi), jsou všechna vlastní čísla nezáporná.

Označme si vlastní čísla seřazená sestupně podle velikosti jako $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ a příslušné vlastní vektory jako $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_p$.

Platí tedy $\mathbf{X'}^T \mathbf{X'} \boldsymbol{b}_i = (N-1)\lambda_i \boldsymbol{b}_i$.

Podprostor $\mathcal{V}$ nyní vyrobíme z prvních q vektorů této báze. Příslušná matice $\mathbf{V}$ má ve sloupcích zapsané vektory $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_q$ a matice $\mathbf{U}$ potom zbylé vektory $\boldsymbol{b}_{q+1}, \ldots, \boldsymbol{b}_p$.

Pro výraz, který jsme chtěli minimalizovat platí:

$$\sum_{i=1}^{N} \|\boldsymbol{u}_{\boldsymbol{x}'_i}\|^2 = \sum_{i=1}^{N} \boldsymbol{x}_i'^T \mathbf{U}\mathbf{U}^T \boldsymbol{x}'_i = \sum_{i=1}^{N} \sum_{j=q+1}^{p} \boldsymbol{x}_i'^T \boldsymbol{b}_j \boldsymbol{b}_j^T \boldsymbol{x}'_i = \sum_{j=q+1}^{p} \sum_{i=1}^{N} \boldsymbol{b}_j^T \boldsymbol{x}'_i \boldsymbol{x}_i'^T \boldsymbol{b}_j$$

$$= \sum_{j=q+1}^{p} \boldsymbol{b}_j^T \mathbf{X'}^T \mathbf{X'} \boldsymbol{b}_j = \sum_{j=q+1}^{p} (N-1)\lambda_j \boldsymbol{b}_j^T \boldsymbol{b}_j = (N-1)(\lambda_{q+1} + \ldots + \lambda_p).$$

Lze ukázat, že pro jakýkoliv q rozměrný podprostor je tento součet roven výrazu $(N-1)(\gamma_1\lambda_1 + \ldots \gamma_p\lambda_p)$, kde $0 \leq \gamma_i \leq 1$ a $\sum_i \gamma_i = p - q$, a tedy nikdy nemůže být menší.

Hlavní komponenty

Dataset $\mathbf{X}$ obsahující body $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^p$ můžeme chápat jako náhodný výběr z rozdělení příznaků $\boldsymbol{X} = (X_1, \ldots, X_p)^T$.

Středovaný dataset $\mathbf{X}'$ obsahující body $\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_N \in \mathbb{R}^p$ odpovídá náhodnému výběru z rozdělení posunutých příznaků $\boldsymbol{X}' = (X'_1, \ldots, X'_p)^T$.

Transformovaný dataset $\mathbf{T}_q = \mathbf{X}'\mathbf{V}$, obsahuje body $\boldsymbol{t}_{\boldsymbol{x}'_1}, \ldots \boldsymbol{t}_{\boldsymbol{x}'_N} \in \mathbb{R}^q$ a odpovídá náhodnému výběru z rozdělení příznaků $\boldsymbol{T} = (T_1, \ldots, T_q)^T$, kde

$$T_i = b_{i;1}X'_1 + \ldots b_{i;p}X'_p = \boldsymbol{b}_i^T \boldsymbol{X}' \quad \text{a tedy} \quad \boldsymbol{T} = \mathbf{V}^T \boldsymbol{X}'.$$

Tyto nové příznaky nazýváme hlavní komponenty (angl. **principal components**) a konkrétně T_i nazýváme i-tou **hlavní komponentou** (angl. i**th principal component**).

Hlavní komponenty mají významnou statistickou interpretaci, jak si nyní ukážeme.

Statistická interpretace hlavních komponent

Pro výběrové průměry a výběrové kovariance hlavních komponent platí

$$\bar{T}_i = \frac{1}{N}\sum_{k=1}^{N} t_{\boldsymbol{x}'_k;i} = \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{b}_i^T \boldsymbol{x}'_k = \boldsymbol{b}_i^T \left(\frac{1}{N}\sum_{k=1}^{N} \boldsymbol{x}'_k\right) = 0,$$

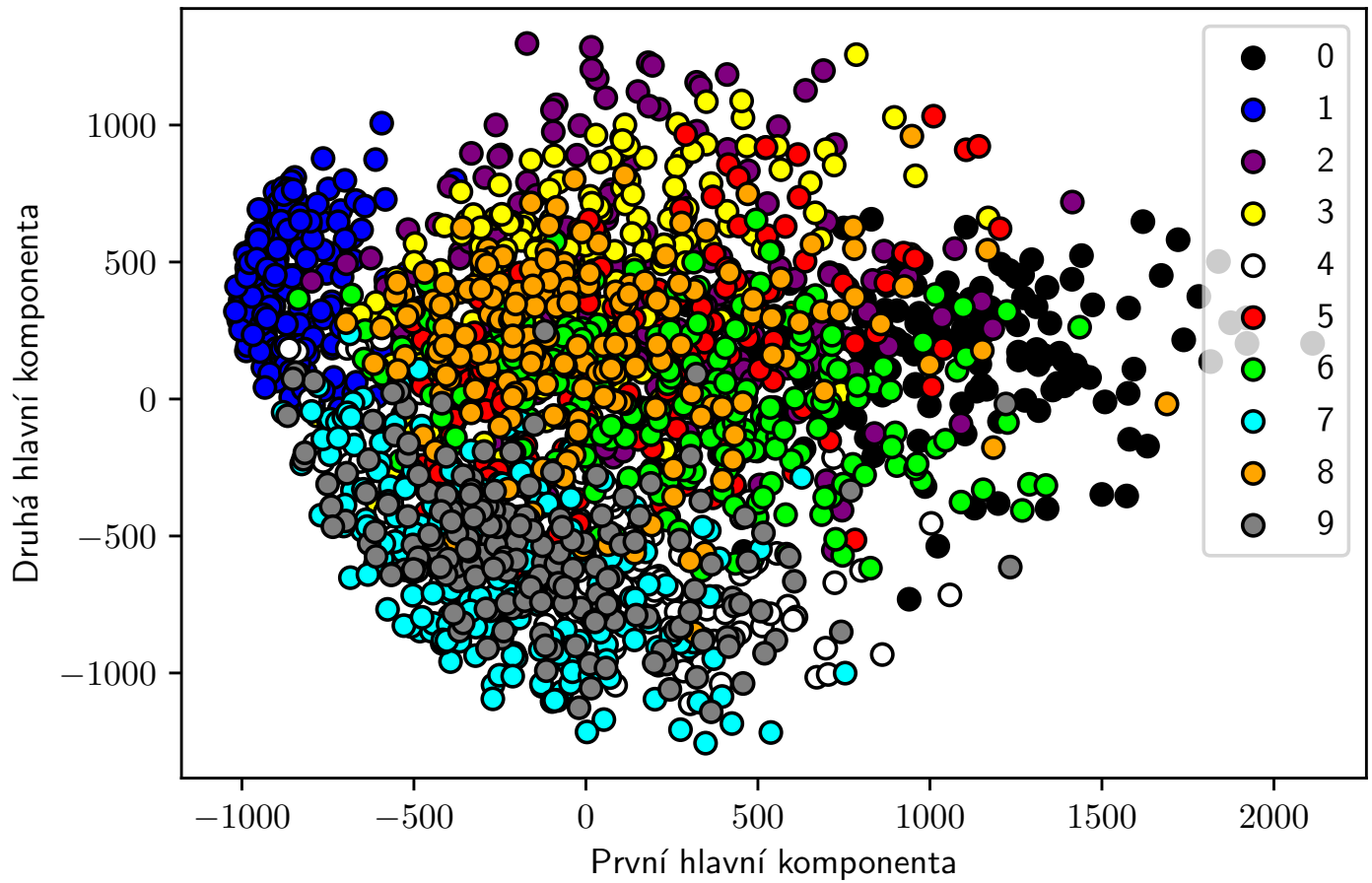
$$\widehat{\mathrm{cov}}(T_i, T_j) = \frac{1}{N-1}\sum_{k=1}^{N} t_{\boldsymbol{x}'_k;i} t_{\boldsymbol{x}'_k;j} = \frac{1}{N-1}(\mathbf{T}_q^T \mathbf{T}_q)_{ij} = \frac{1}{N-1}(\mathbf{V}^T \mathbf{X}'^T \mathbf{X}' \mathbf{V})_{ij}$$

$$= \frac{1}{N-1}\boldsymbol{b}_i^T \mathbf{X}'^T \mathbf{X}' \boldsymbol{b}_j = \lambda_j \boldsymbol{b}_i^T \boldsymbol{b}_j = \begin{cases} 0 & \text{pro } i \neq j, \\ \lambda_j & \text{pro } i = j. \end{cases}$$

Platí tedy:

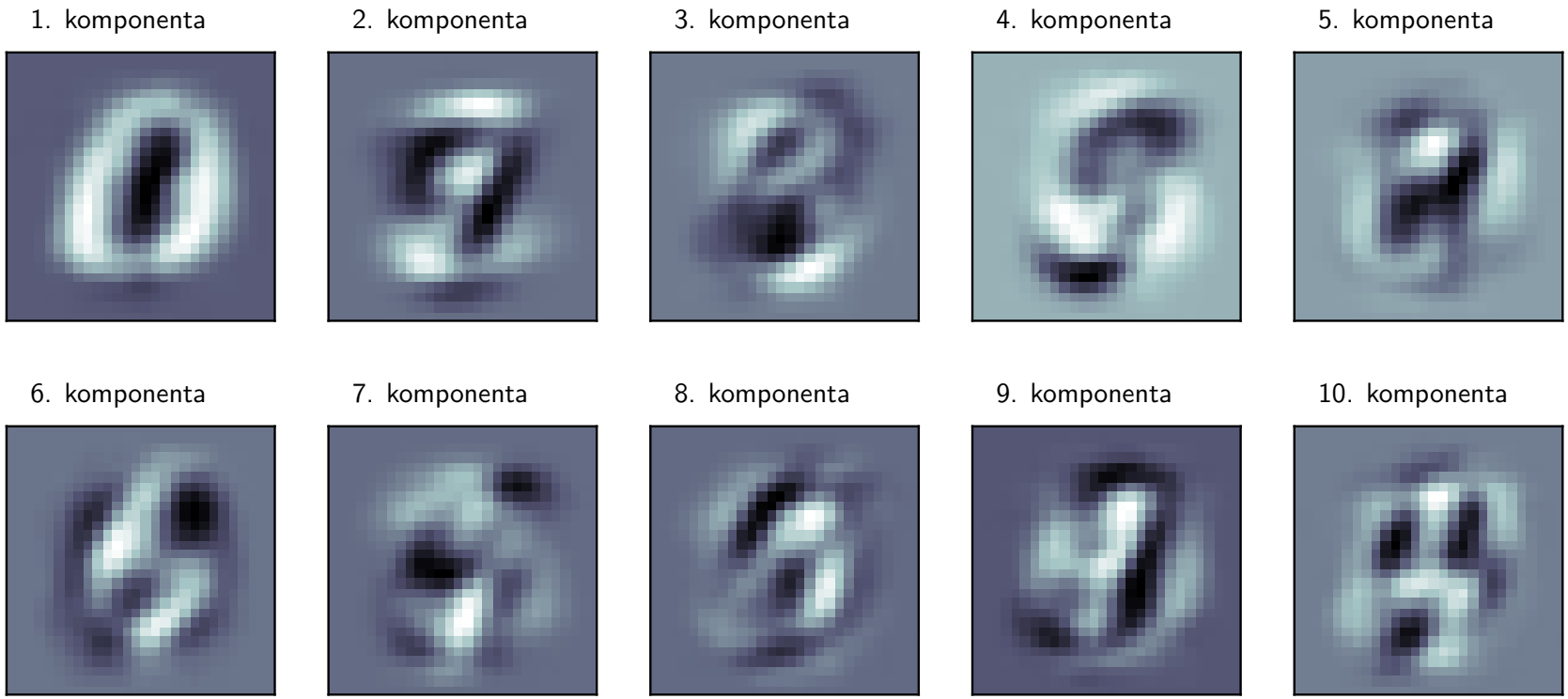
- Rozptyl i-té hlavní komponenty je tedy roven i-tému vlastnímu číslu λ_i.
- Výběrem q hlavních komponent tedy vybereme směry, ve kterých mají data největší rozptyl!
- Různé komponenty (příznaky T_i) jsou nekorelované!
- Podíl rozptylu „vysvětlený“ i-tou komponentou je $\frac{\lambda_i}{\lambda_1+\ldots+\lambda_p}$.

Příklad: MNIST



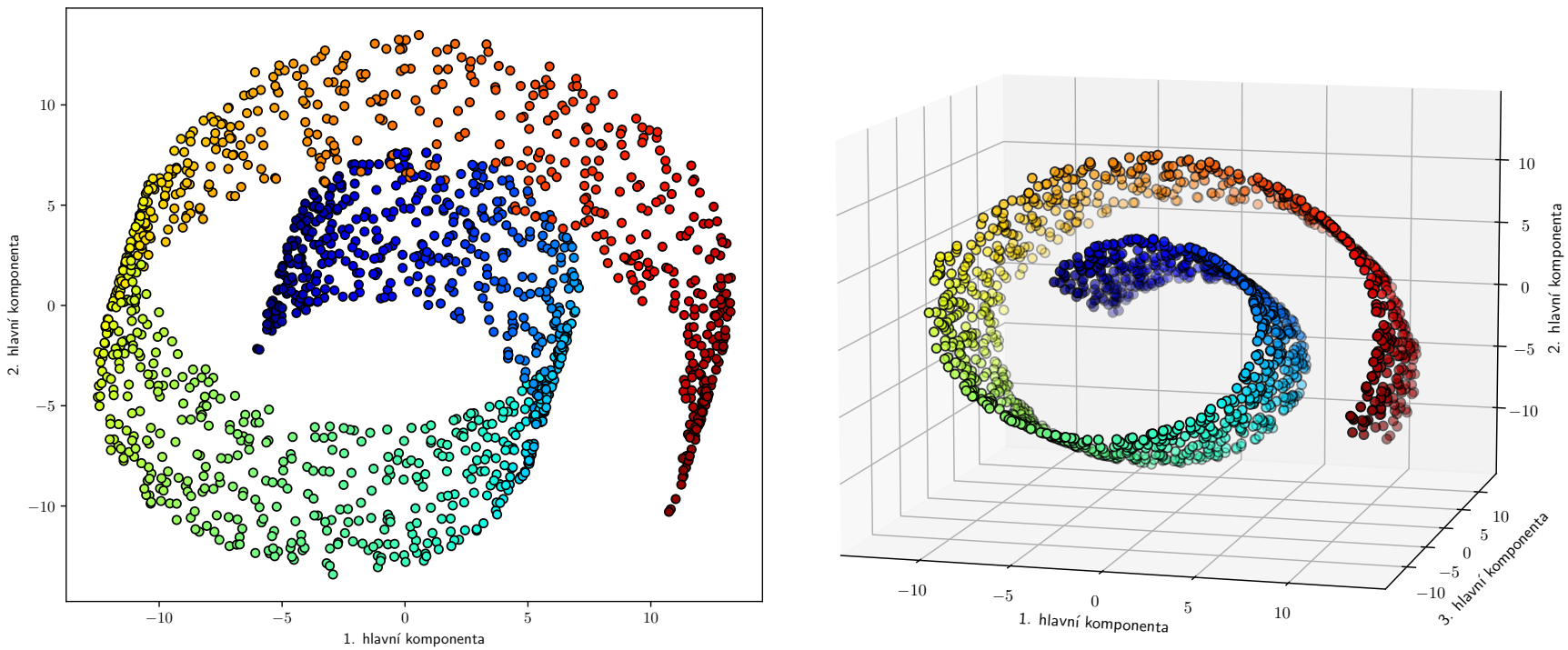
Vykreslení v dvou hlavních komponentách s vyznačením třídy nám dává nějakou informaci o existujících shlucích.

Příklad: MNIST



Pro hlavní komponenty je zde velmi obtížené interpretovat jejich vztah ke třídám!

Příklad: Švýcarská rolka



PCA transformace švýcarské rolky do dvou dimenzí a pak do tří.

Závěrečný komentář

- K numerickému výpočtu PCA je možné jednak využít přímý spektrální rozklad matice $\frac{1}{N-1}\mathbf{X}'^T\mathbf{X}'$ nebo rozklad matice $\mathbf{X}'$ na singulární hodnoty (singular value decomposition - SVD), který je v zásadě numericky stabilnější.

- Počet komponent q je možné volit na základě „elbow“ grafů, ve kterých nakreslíme počet komponent proti celkovému rozptylu vysvětlenému danými komponentami a hledáme bod zlomu.

- Pokud zvolíme $q = p$, tak nezahodíme žádnou informaci, ale pouze přejdeme do reprezentace dat v ortonormální bázi.

- V takovém případě budou sloupce matice $\mathbf{T}_q$ na sebe kolmé, protože

$$\mathbf{T}_q^T\mathbf{T}_q = \mathbf{V}^T\mathbf{X}'^T\mathbf{X}'\mathbf{V} = (N-1)\mathbf{\Lambda},$$

  kde $\mathbf{\Lambda}$ je matice, na jejíž diagonále jsou vlastní čísla $\lambda_1, \ldots, \lambda_p$ matice $\frac{1}{N-1}\mathbf{X}'^T\mathbf{X}'$ a mimo diagonálu 0.

Manifold learning

- Problémem ortogonální projekce (a tedy i PCA) je situace, kdy jsou v datech nelineární závislosti způsobené například komplikovanými interakcemi mezi příznaky.
- V takovém případě je dobré se místo na globální vlastnosti rozdělení dat soustředit na ty lokální.
- O to se obecně snaží metody z manifold learning. Ukažme si jednu z nich.
- **Lokálně lineární vnoření** (angl. **locally linear embedding**), zkracujeme LLE, sleduje, jak jednotlivé trénovací body závisí lineárně na svém okolí a potom hledá méně dimenzionální reprezentaci, která tyto lokální vztahy zachová.

Metoda LLE (1/3)

- Uvažujme i-tý trénovací bod $\boldsymbol{x}_i$ (i-tý řádek matice $\mathbf{X}$).
- Pro dané k spočteme jeho k nejbližších sousedů ($\boldsymbol{x}_i$ mezi ně nepočítáme).
- Dále najdeme váhy $w_{i,j}$ takové aby:
  - $w_{i,j} = 0$ pro j takové, že $\boldsymbol{x}_j$ není mezi k nejbližšími sousedy $\boldsymbol{x}_i$,
  - $\sum_j w_{i,j} = 1$,
  - vzdálenost mezi $\boldsymbol{x}_i$ a $\sum_j w_{i,j}\boldsymbol{x}_j$ byla nejmenší možná.
- Toto provedeme pro všechny body najednou. To ekvivalentně znamená, že hledáme

$$\arg\min_{w_{i,j}} \sum_{i=1}^{N} \left\| \boldsymbol{x}_i - \sum_j w_{i,j}\boldsymbol{x}_j \right\|^2,$$

  za výše uvedené podmínky $w_{i,j} = 0$ pro j takové, že $\boldsymbol{x}_j$ není mezi k nejbližšími sousedy $\boldsymbol{x}_i$, a současně $\sum_j w_{i,j} = 1$.

Metoda LLE (2/3)

- Jako $\mathbf{W}^*$ označme matici která obsahuje optimální hodnoty vah $w^*_{i,j}$.
- V dalším kroku se pro každé $\boldsymbol{x}_i \in \mathbb{R}^p$ snažíme najít q rozměrnou reprezentaci $\boldsymbol{z}_i \in \mathbb{R}^q$ takovou, že je vzdálenost mezi $\boldsymbol{z}_i$ a $\sum_j w^*_{i,j}\boldsymbol{z}_j$ nejmenší možná.
- Tímto způsobem získáme matici $\mathbf{Z}^*$, která má v řádcích zapsané výsledné optimální body $\boldsymbol{z}^*_i$, což představuje finální vnoření.
- Postup je tedy dvoufázový:
  - Nejprve se na základě originálních dat spočítají váhy, které nějakým způsobem zachycují lokální strukturu těch dat.
  - Výsledná reprezentace se potom nalezne s využitím těchto vah, jako d dimenzionální reprezentace, která tu lokální strukturu zachovává nejlépe.

Metoda LLE (3/3)

- Nalezení vah: Hledáme $\mathbf{W}^* \in \mathbb{R}^{N,N}$ takové, že

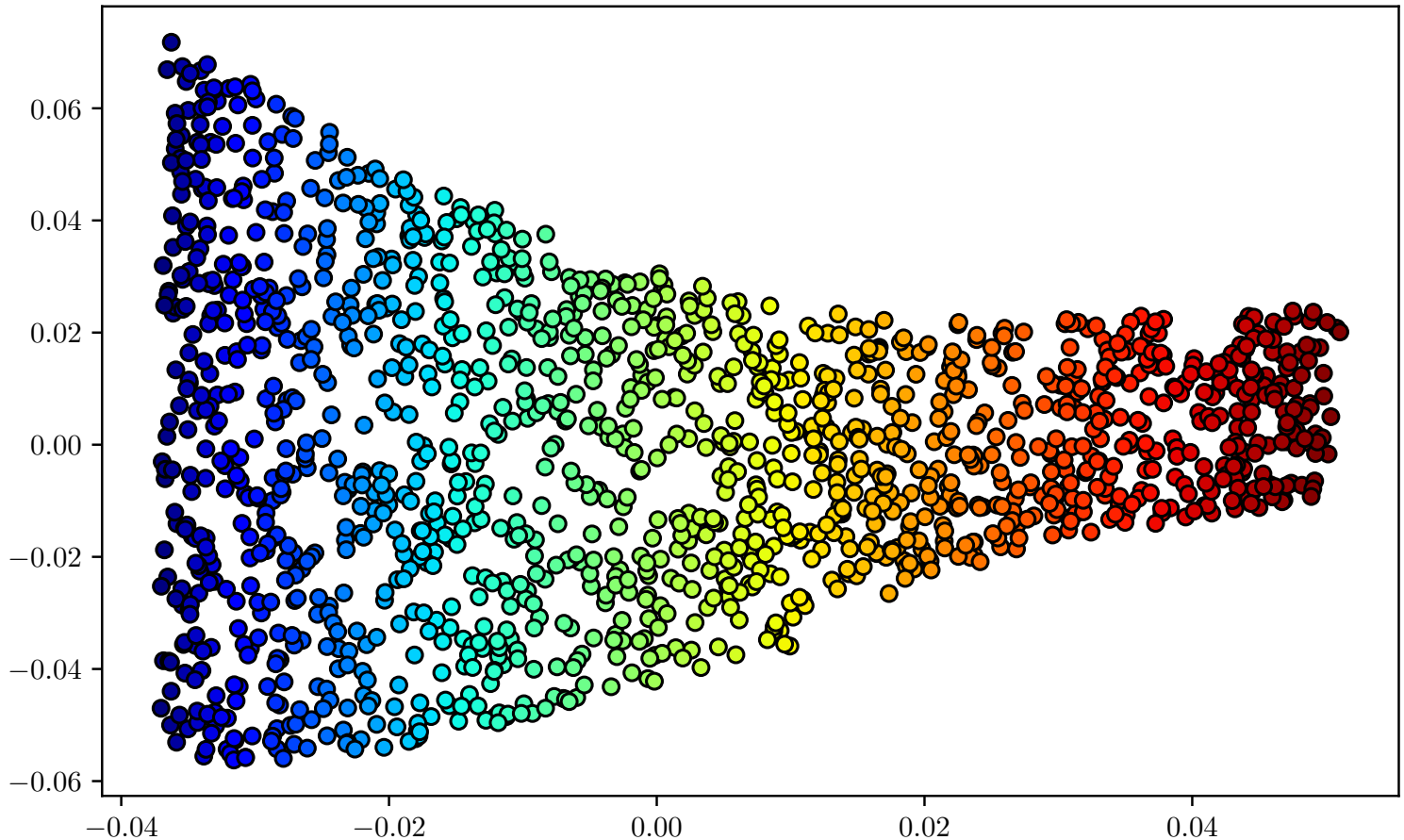
$$\mathbf{W}^* = \arg\min_{\mathbf{W}} \sum_{i=1}^{N} \left\| \boldsymbol{x}_i - \sum_j w_{i,j} \boldsymbol{x}_j \right\|^2$$

  za podmínky $w_{i,j} = 0$ pro j takové, že $\boldsymbol{x}_j$ není mezi k nejbližšími sousedy $\boldsymbol{x}_i$, a současně $\sum_j w_{i,j} = 1$.

- Nalezení vnoření: Najdeme matici $\mathbf{Z}^* \in \mathbb{R}^{N,q}$ takovou, že

$$\mathbf{Z}^* = \arg\min_{\mathbf{Z}} \sum_{i=1}^{N} \left\| \boldsymbol{z}_i - \sum_j w^*_{i,j} \boldsymbol{z}_j \right\|^2.$$

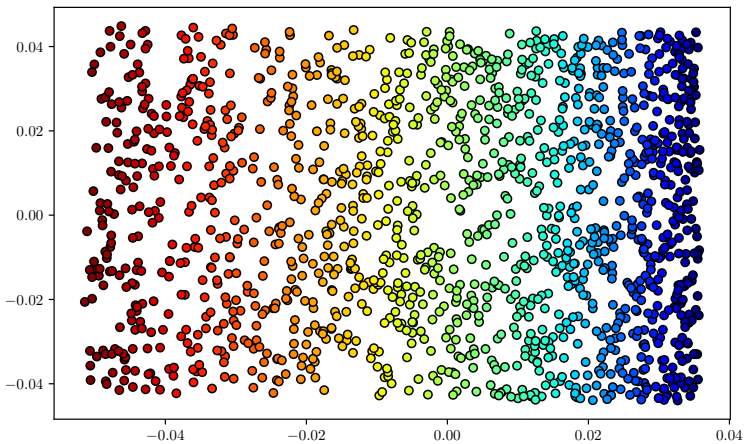
Příklad: Švýcarská rolka



Lokálně lineární vnoření švýcarské rolky do dvou dimenzí pro $k = 15$ sousedů.

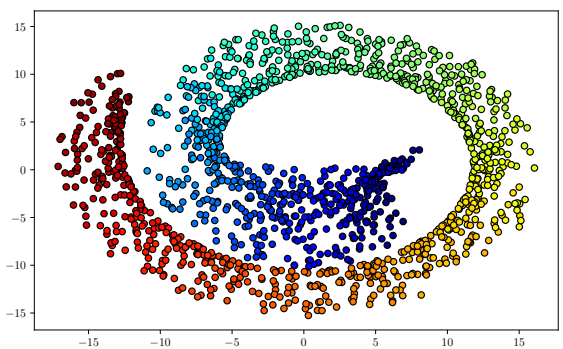
Závěrečné poznámky

V případě, že je počet použitých nejbližších sousedů k větší než dimenze p původního prostoru příznaků, jsou body z okolí lineárně závislé. Pro tento případ je výhodné použít modifikované LLE.

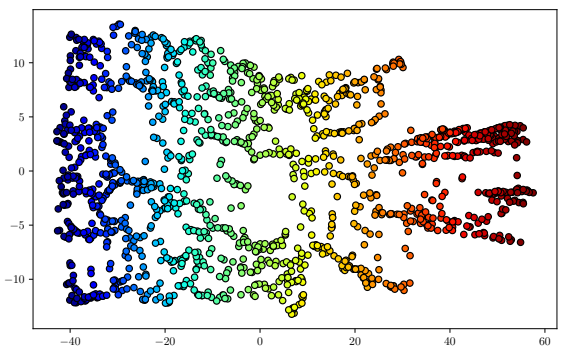


Modifikované LLE

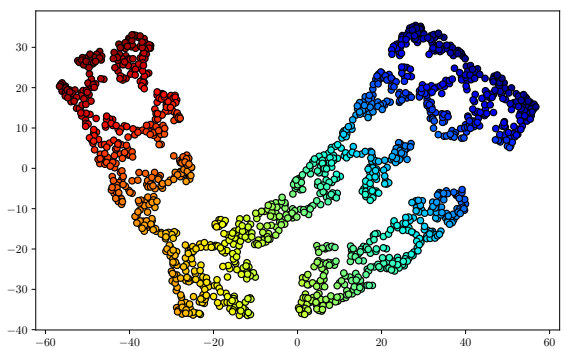
Existují také další metody jako např. Multi-dimensional Scaling (MDS), Isomap, t-distributed Stochastic Neighbor Embedding (t-SNE).



MDS



Isomap



t-SNE