Търсене и извличане на информация. Приложение на дълбоко машинно обучение

Стоян Михов

Лекция 6: Принципен компонентен анализ (PCA). Влагане на думи и документи в гъсто нискомерно векторно пространство.

План на лекцията

1. **Формалности за курса (5 мин)**
2. Интуиция за принципния компонентен анализ (10 мин)
3. Смятане с вектори и матрици (15 мин)
4. Свойства на ковариационната матрица (15 мин)
5. Задача за намиране на принципните компоненти (25 мин)
6. Влагане на думи и контексти в нискомерно гъсто векторно пространство и латентен семантичен анализ (10 мин)
7. Семантично пространствени релации (10 мин)

Формалности

- В близките дни ще бъде публикувано първото домашно задание в Мудъл.

- Шестата лекция се базира на глава 18 от първия учебник и секция 10.4 от втория учебник.

План на лекцията

1. Формалности за курса (5 мин)

2. **Интуиция за принципния компонентен анализ (10 мин)**

3. Смятане с вектори и матрици (15 мин)

4. Свойства на ковариационната матрица (15 мин)

5. Задача за намиране на принципните компоненти (25 мин)

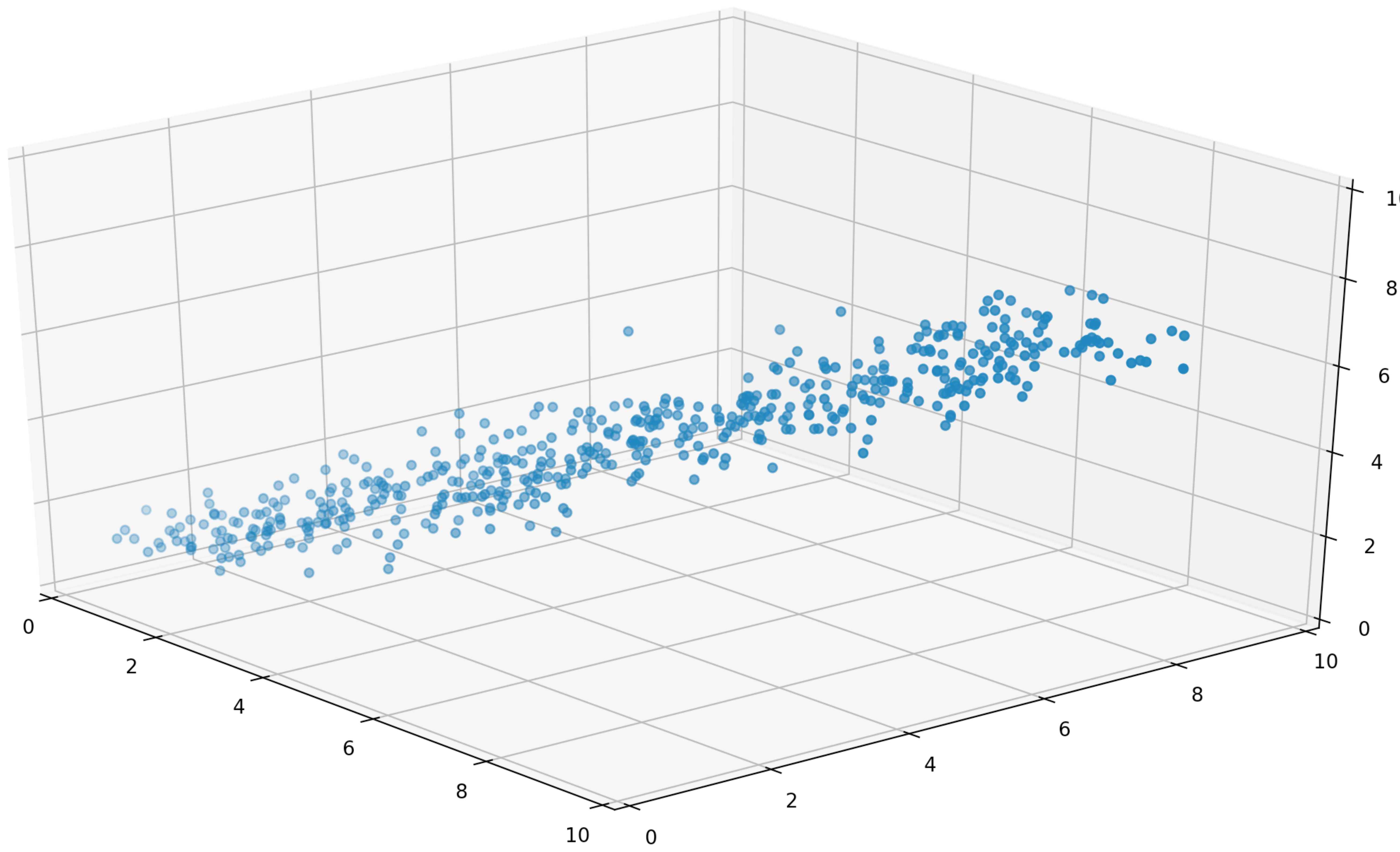
6. Влагане на думи и контексти в нискомерно гъсто векторно пространство и латентен семантичен анализ (10 мин)

7. Семантично пространствени релации (10 мин)

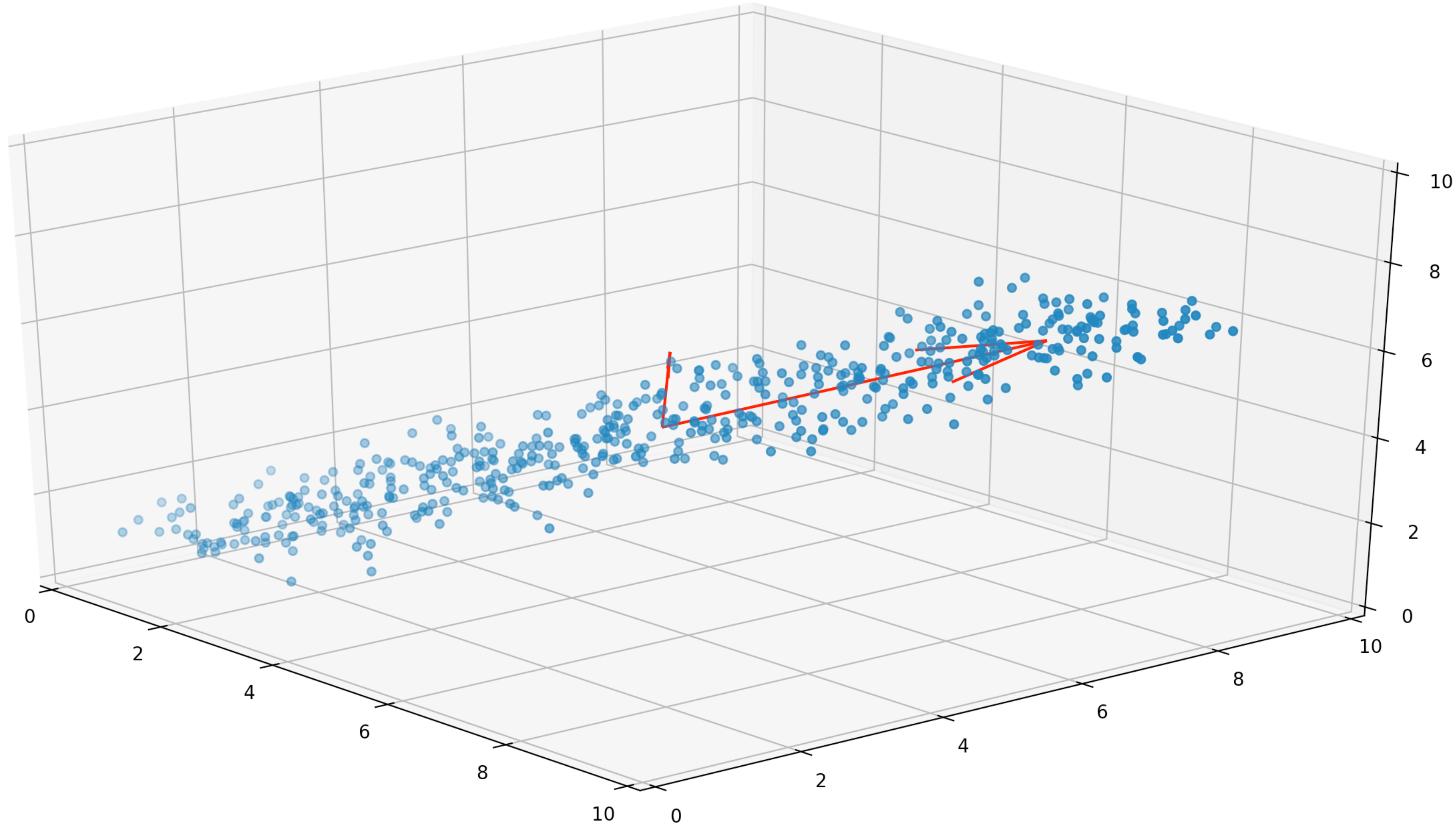
Влагането на думи във векторното пространство на контекстите

- Контекстът на дадена дума са думите, които са около нея — в рамките на параграф, изречение или фиксиран по размер прозорец.
- На всяка дума съпоставяме вектора от свързванията на думата с всеки от контекстите.
- Размерността на пространството е огромна, което води до изчислителни трудности.
- **Цел**: Да намерим влагане на векторите в нискомерно гъсто векторно пространство, което възможно най-добре да отразява разстоянията в многомерното контекстно пространство.

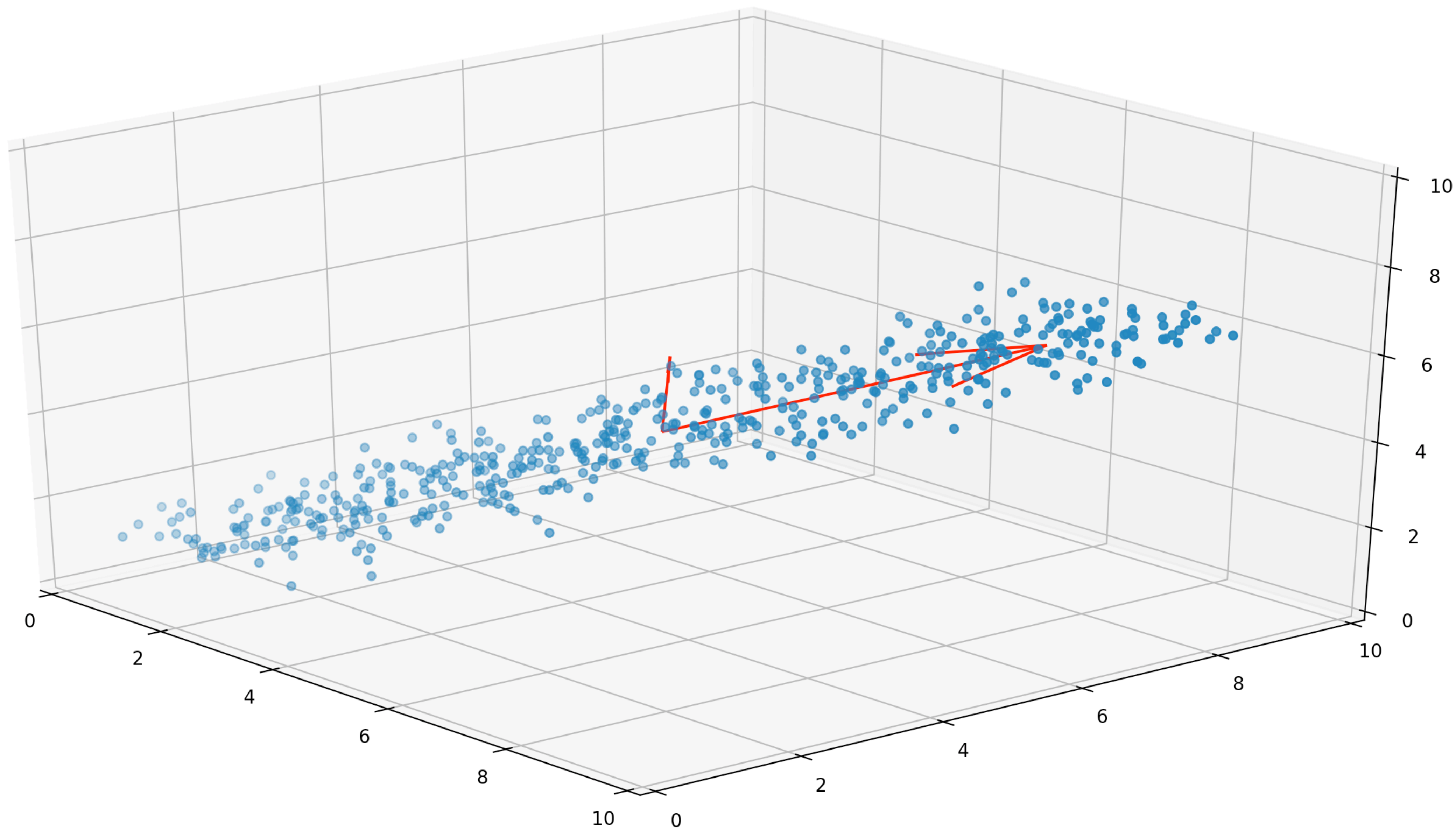
Интуитивна представа за принципен компонентен анализ

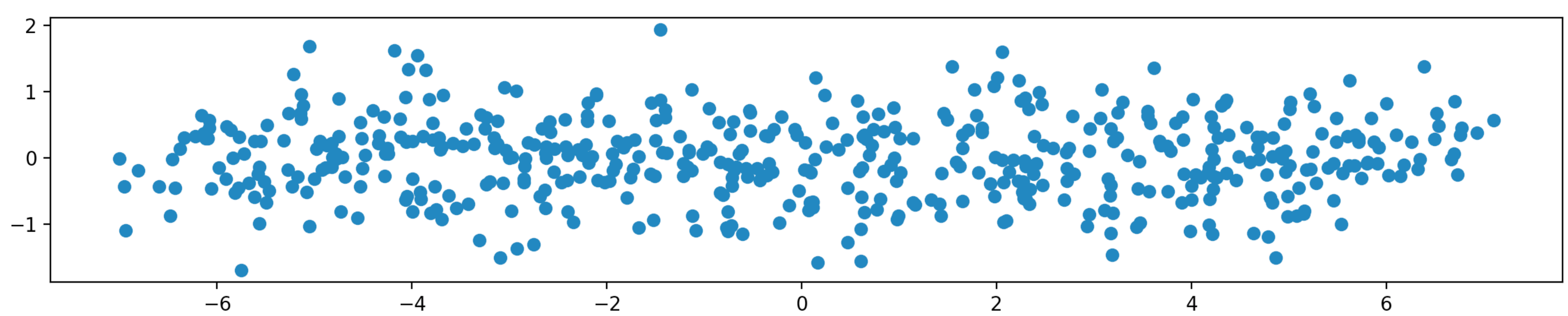


Интуитивна представа за принципен компонентен анализ



Интуитивна представа за принципен компонентен анализ





Основна идея

- Ще използваме техниката на принципния компонентен анализ за да намерим нискомерен базис от ортогонални вектори — размерността обикновено е между 25 и 1000.
- В този базис разстоянията между векторите ще искаме да са максимално близки до съответните разстояния в многомерното пространство.
- Направленията в новия базис ще съответстват на линейни комбинации от оригиналните базисни вектори определени от контекстите.
- Новите вектори ще бъдат “гъсти” — почти няма да има нулеви компоненти.
- Намалянето на размерността може да доведе до намаляне на шума и до постигане на по-висока прецизност.

План на лекцията

1. Формалности за курса (5 мин)
2. Интуиция за принципния компонентен анализ (10 мин)
3. **Смятане с вектори и матрици (15 мин)**
4. Свойства на ковариационната матрица (15 мин)
5. Задача за намиране на принципните компоненти (25 мин)
6. Влагане на думи и контексти в нискомерно гъсто векторно пространство и латентен семантичен анализ (10 мин)
7. Семантично пространствени релации (10 мин)

Означения на вектори и матрици

- Матриците ще бележим с главни удебелени букви — $\mathbf{W}, \mathbf{V}, \mathbf{A}, \mathbf{B}$, единична матрица — $\mathbf{I}$.
  $\mathbf{W}_{i,j}$ — елементът на ред i, стълб j в матрицата $\mathbf{W}$.
  $\mathbf{W}_{i,\bullet}$ — вектор ред i на матрицата $\mathbf{W}$.
  $\mathbf{W}_{\bullet,j}$ — вектор стълб j на матрицата $\mathbf{W}$.

- Векторите ще бележим с малки удебелени букви — $\mathbf{u}, \mathbf{v}, \mathbf{x}, \mathbf{y}$.
  $\mathbf{u}_i$ — i-тия елемент на $\mathbf{u}$.

- Ако не е указано друго, ще подразбираме вектор стълбове. Т.е. $\mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u_2} \\ \vdots \\ \mathbf{u}_n \end{bmatrix}$.

- Векторите може да разглеждаме като матрици — $\mathbf{u} \in \mathbb{R}^{n \times 1}$.

- С $\mathbf{1}_n \in \mathbb{R}^n$ ще бележим вектор, състоящ се само от единици. С $\mathbf{1}_{n \times m} \in \mathbb{R}^{n \times m}$ ще бележим матрица, състояща се само от единици.

Произведения на матрици и вектори

- **Произведение на матрици**: Нека $\mathbf{A} \in \mathbb{R}^{n \times m}, \mathbf{B} \in \mathbb{R}^{m \times k}$, тогава $\mathbf{C} = \mathbf{AB}$, ако $\mathbf{C} \in \mathbb{R}^{n \times k}$ и $\mathbf{C}_{i,j} = \sum_{l=1}^{m} \mathbf{A}_{i,l}\mathbf{B}_{l,j}$.
- **Произведение на матрица с вектор**:
  Нека $\mathbf{A} \in \mathbb{R}^{n \times m}, \mathbf{x} \in \mathbb{R}^{m \times 1}$, тогава $\mathbf{y} = \mathbf{Ax}$, ако $\mathbf{y} \in \mathbb{R}^{n \times 1}$ и $\mathbf{y}_i = \sum_{l=1}^{m} \mathbf{A}_{i,l}\mathbf{x}_l$.
  Нека $\mathbf{u} \in \mathbb{R}^{n \times 1}$, тогава вектора ред $\mathbf{v} = \mathbf{u}^\top \mathbf{A}$, ако $\mathbf{v} \in \mathbb{R}^{1 \times m}$ и $\mathbf{v}_i = \sum_{l=1}^{n} \mathbf{u}_l \mathbf{A}_{l,i}$.
- **Скаларно произведение** на два вектора: Нека $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{n \times 1}$ тогава $\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^\top \mathbf{v} = \sum_{l=1}^{n} \mathbf{u}_l \mathbf{v}_l$.
- **Квадратична форма**: Нека $\mathbf{A} \in \mathbb{R}^{n \times n}, \mathbf{x} \in \mathbb{R}^{n \times 1}$ тогава $\mathbf{x}^\top \mathbf{A} \mathbf{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{x}_i \mathbf{A}_{i,j} \mathbf{x}_j$.
- **Диадно (тензорно, външно) произведение** на два вектора: Нека $\mathbf{x} \in \mathbb{R}^{n \times 1}, \mathbf{y} \in \mathbb{R}^{m \times 1}$, тогава $\mathbf{x} \otimes \mathbf{y} = \mathbf{x}\mathbf{y}^\top = \mathbf{C}$, ако $\mathbf{C} \in \mathbb{R}^{n \times m}$ и $\mathbf{C}_{i,j} = \mathbf{x}_i \mathbf{y}_j$.
- **Поелементно произведение** на вектори или матрици: Ако $\mathbf{x} \in \mathbb{R}^{n \times 1}, \mathbf{y} \in \mathbb{R}^{n \times 1}$, то $\mathbf{x} \odot \mathbf{y} \in \mathbb{R}^{n \times 1}$ и $(\mathbf{x} \odot \mathbf{y})_i = \mathbf{x}_i \mathbf{y}_i$. Ако $A \in \mathbb{R}^{n \times m}, B \in \mathbb{R}^{n \times m}$, то $A \odot B \in \mathbb{R}^{n \times m}$ и $(A \odot B)_{i,j} = A_{i,j} B_{i,j}$.

Градиент и Якобиан

- Нека $f : \mathbb{R}^n \to \mathbb{R}$. **Градиент** на f наричаме вектора: $\nabla_{\mathbf{x}} f(\mathbf{x}) = \dfrac{\partial}{\partial \mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial}{\partial x_1} f(\mathbf{x}) \\ \frac{\partial}{\partial x_2} f(\mathbf{x}) \\ \vdots \\ \frac{\partial}{\partial x_n} f(\mathbf{x}) \end{bmatrix} \in \mathbb{R}^n,$

$$\left(\frac{\partial}{\partial \mathbf{x}} f(\mathbf{x})\right)_i = \frac{\partial}{\partial x_i} f(\mathbf{x})$$

- Нека $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$, т.е. $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_1(\mathbf{x}), \ldots, f_1(\mathbf{x})) \in \mathbb{R}^m$,
  Якобиан на $\mathbf{f}$ наричаме матрицата:

$$\nabla_{\mathbf{x}} \mathbf{f}(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}} \mathbf{f}(\mathbf{x}) = \begin{bmatrix} \frac{\partial}{\partial x_1} f_1(\mathbf{x}) & \cdots & \frac{\partial}{\partial x_n} f_1(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1} f_m(\mathbf{x}) & \cdots & \frac{\partial}{\partial x_n} f_m(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \nabla_{\mathbf{x}} f_1(\mathbf{x})^\top \\ \vdots \\ \nabla_{\mathbf{x}} f_m(\mathbf{x})^\top \end{bmatrix} \in \mathbb{R}^{m \times n}.$$

$$\left(\frac{\partial}{\partial \mathbf{x}} \mathbf{f}(\mathbf{x})\right)_{i,j} = \frac{\partial}{\partial x_j} f_i(\mathbf{x})$$

Свойства

- Нека $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{x} \in \mathbb{R}^{n\times 1}$. Тогава: $\dfrac{\partial}{\partial \mathbf{x}}\mathbf{A}\mathbf{x} = \mathbf{A} \in \mathbb{R}^{m\times n}$

- Нека $\mathbf{x} \in \mathbb{R}^{n\times 1}$. Тогава $\dfrac{\partial}{\partial \mathbf{x}}\mathbf{x} = \dfrac{\partial}{\partial \mathbf{x}}\mathbf{I}\mathbf{x} = \mathbf{I}$.

- Нека $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{n\times 1}$. Тогава $\dfrac{\partial}{\partial \mathbf{u}}\mathbf{u}^\top\mathbf{v} = \mathbf{v}$, $\dfrac{\partial}{\partial \mathbf{v}}\mathbf{u}^\top\mathbf{v} = \mathbf{u}$.

- Нека $\mathbf{f}, \mathbf{g} \in \mathbb{R}^n \to \mathbb{R}^m$. Тогава
$$\frac{\partial}{\partial \mathbf{x}}\left(\mathbf{f}(\mathbf{x})^\top\mathbf{g}(\mathbf{x})\right) = \left(\frac{\partial}{\partial \mathbf{x}}\mathbf{f}(\mathbf{x})\right)^\top \mathbf{g}(\mathbf{x}) + \mathbf{f}(\mathbf{x})^\top\left(\frac{\partial}{\partial \mathbf{x}}\mathbf{g}(\mathbf{x})\right)$$

- Нека $\mathbf{A} \in \mathbb{R}^{n\times n}, \mathbf{x} \in \mathbb{R}^{n\times 1}$. Тогава $\dfrac{\partial}{\partial \mathbf{x}}\mathbf{x}^\top\mathbf{A}\mathbf{x} = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$.

План на лекцията

1. Формалности за курса (5 мин)
2. Интуиция за принципния компонентен анализ (10 мин)
3. Смятане с вектори и матрици (15 мин)
4. **Свойства на ковариационната матрица (15 мин)**
5. Задача за намиране на принципните компоненти (25 мин)
6. Влагане на думи и контексти в нискомерно гъсто векторно пространство и латентен семантичен анализ (10 мин)
7. Семантично пространствени релации (10 мин)

Ковариационна матрица

- Ковариационна матрица на вектор от случайни величини $\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}$ е матрица $\mathbb{R}^{N\times N}$, която означаваме с $\mathbf{C}[\mathbf{X}]$ и дефинираме като:
  $\mathbf{C}[\mathbf{X}] = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top]$,
  т.е. $\mathbf{C}[\mathbf{X}]_{i,j} = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])^\top] = \mathrm{Cov}(X_i, X_j)$

Свойство:

- $\mathbf{C}[\mathbf{X}] = \mathbb{E}[\mathbf{X}\mathbf{X}^\top] - \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{X}]^\top$
  - защото: $\mathbf{C}[\mathbf{X}]_{i,j} = \mathrm{Cov}(X_i, X_j) = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]$

Забележка: $\mathbf{X}\mathbf{X}^\top = \mathbf{X} \otimes \mathbf{X}$

- **Свойство**: Нека $u, v \in \mathbb{R}^n$. Тогава:$(u \cdot v)^2 = v^\top (u \otimes u) v = v^\top (uu^\top) v$
  доказателство:
$$(u \cdot v)^2 = (u^\top v)(u^\top v) = (\sum_{i=1}^{n} u_i v_i)(\sum_{j=1}^{n} u_j v_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} u_i v_i u_j v_j$$
$$v^\top (uu^\top) v = v^\top ((uu^\top)v) = \sum_{i=1}^{n} v_i ((uu^\top)v)_i = \sum_{i=1}^{n} v_i \sum_{j=1}^{n} (u_i u_j) v_j = \sum_{i=1}^{n}\sum_{j=1}^{n} u_i v_i u_j v_j$$

- **Дефиниция**: Матрицата (определяща квадратична форма) $A \in \mathbb{R}^{n \times n}$ наричаме положително дефинитна, ако за всеки вектор $v \in \mathbb{R}^n$ е изпълнено: $v^\top A v \geq 0$.

- **Твърдение**: Всяка ковариационна матрица е симетрична и положително дефинитна.

- **Доказателство**:
$$\mathbf{C}(\mathbf{X}) = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])[(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T] = \sum_{\mathbf{x} \in \mathbf{X}(\Omega)} \Pr[\mathbf{X} = \mathbf{x}]\,(\mathbf{x} - \mathbb{E}[\mathbf{X}])(\mathbf{x} - \mathbb{E}[\mathbf{X}])^\top$$
$$v^\top \mathbf{C}(\mathbf{X}) v = \sum_{\mathbf{x} \in \mathbf{X}(\Omega)} \Pr[\mathbf{X} = \mathbf{x}]\, v^\top (\mathbf{x} - \mathbb{E}[\mathbf{X}])(\mathbf{x} - \mathbb{E}[\mathbf{X}])^\top v =$$
$$= \sum_{\mathbf{x} \in \mathbf{X}(\Omega)} \Pr[\mathbf{X} = \mathbf{x}]\, ((\mathbf{x} - \mathbb{E}[\mathbf{X}]) \cdot v)^2 \geq 0$$

- **Теорема**: (от курса по Линейна алгебра) Нека $A \in \mathbb{R}^{n \times n}$ е симетрична матрица ($A^\top = A$). Тогава:

    1. Всички собствени стойности на A — корените на характеристичното уравнение $|A - \lambda \mathbf{I}| = 0$ — са реални числа.

    2. Съществува ортонормиран базис $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n \in \mathbb{R}^n$ от собствени вектори на A, така че:

        - $A\mathbf{e}_i = \lambda_i \mathbf{e}_i$

        - $A = T \Lambda T^{-1}$, където $T = [\mathbf{e}_1 \quad \mathbf{e}_2 \quad \ldots \quad \mathbf{e}_n]$ и $\Lambda = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ & & \ddots & \\ 0 & 0 & \ldots & \lambda_n \end{bmatrix}$

        - Забележка: Матрицата T е ортогонална: $T^{-1} = T^\top$

- **Твърдение**: Ако $A \in \mathbb{R}^{n \times n}$ е симетрична и положително дефинитна матрица то всички собствени стойности на A са реални неотрицателни числа.

- **Доказателство**: $\lambda = \mathbf{e}^\top \lambda \mathbf{e} = \mathbf{e}^\top (A\mathbf{e}) \geq 0$

План на лекцията

1. Формалности за курса (5 мин)
2. Интуиция за принципния компонентен анализ (10 мин)
3. Смятане с вектори и матрици (15 мин)
4. Свойства на ковариационната матрица (15 мин)
5. **Задача за намиране на принципните компоненти (25 мин)**
6. Влагане на думи и контексти в нискомерно гъсто векторно пространство и латентен семантичен анализ (10 мин)
7. Семантично пространствени релации (10 мин)

Постановка на задачата

- Дадени са S вектора $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(S)} \in \mathbb{R}^N$ в N мерно пространство и число $M \in \mathbb{N}^+, M < N$. Можем да разглеждаме векторите $\mathbf{x}^{(i)}$ като S наблюдения на вектор $\mathbf{X}$ от N случайни величини.

- Търсим ортонормиран базис $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_N$ в $\mathbb{R}^N$, и числа $b_{M+1}, b_{M+2}, \ldots, b_N$, така че ако
    - векторите $\mathbf{x}^{(i)}$ се представят в новата координатна система като $\mathbf{x}^{(i)} = \sum_{j=1}^{N} y_j^{(i)} \mathbf{e}_j$, където $y_j^{(i)} = \mathbf{x}^{(i)} \cdot \mathbf{e}_j$, и
    - $\hat{\mathbf{x}}^{(i)} = \sum_{j=1}^{M} y_j^{(i)} \mathbf{e}_j + \sum_{j=M+1}^{N} b_j \mathbf{e}_j$ са проекции на $\mathbf{x}^{(i)}$ върху M мерна хиперравнина

- То $\varepsilon^2 = \dfrac{1}{S} \sum_{i=1}^{S} \|\mathbf{x}^{(i)} - \hat{\mathbf{x}}^{(i)}\|^2$ е минимално.

- $\varepsilon^2 = \frac{1}{S}\sum_{i=1}^{S} \|\mathbf{x}^{(i)} - \hat{\mathbf{x}}^{(i)}\|^2 = \frac{1}{S}\sum_{i=1}^{S}\sum_{j=M+1}^{N} (\mathbf{x}^{(i)} \cdot \mathbf{e}_j - b_j)^2$

- За да намерим b_j търсим къде се нулират производните:

- $\frac{\partial}{\partial b_j}\varepsilon^2 = \frac{1}{S}\sum_{i=1}^{S} - 2(\mathbf{x}^{(i)} \cdot \mathbf{e}_j - b_j) = 0$

- $b_j = \frac{1}{S}\sum_{i=1}^{S} \mathbf{x}^{(i)} \cdot \mathbf{e}_j$

- Можем да разглеждаме компонентите на векторите $\mathbf{x}^{(i)}$ като наблюдения на случайните величини $\mathbf{X}^{(i)}$ еднакво разпределени с $\mathbf{X}$. В такъв случай: $b_j = \mathbb{E}_S[\mathbf{X} \cdot \mathbf{e}_j] = \mathbb{E}_S[Y_j]$, където разглеждаме случайна величина $Y_j = \mathbf{X} \cdot \mathbf{e}_j$ с наблюдения $y_j^{(i)} = \mathbf{x}^{(i)} \cdot \mathbf{e}_j$.

- **Интуитивно**: Заменяме измеренията, които премахваме, със средните стойности по тези измерения.

- В такъв случай, като заместим в ε^2 получаваме:

$$\varepsilon^2 = \frac{1}{S}\sum_{i=1}^{S}\sum_{j=M+1}^{N}(\mathbf{x}^{(i)}\cdot\mathbf{e}_j - b_j)^2 = \sum_{j=M+1}^{N}\frac{1}{S}\sum_{i=1}^{S}(y_j^{(i)} - \mathbb{E}_S[Y_j])^2 =$$

$$= \sum_{j=M+1}^{N}\mathbb{E}_S[(Y_j - \mathbb{E}_S[Y_j])^2]$$

- Заместваме и получаваме:

$$\varepsilon^2 = \sum_{j=M+1}^{N}\mathbb{E}_S[(Y_j - \mathbb{E}_S[Y_j])^2] = \sum_{j=M+1}^{N}\mathbb{E}_S[(\mathbf{X}\cdot\mathbf{e}_j - \mathbb{E}_S[\mathbf{X}\cdot\mathbf{e}_j])^2] =$$

$$= \sum_{j=M+1}^{N}\mathbb{E}_S[((\mathbf{X} - \mathbb{E}_S[\mathbf{X}])\cdot\mathbf{e}_j)^2] = \sum_{j=M+1}^{N}\mathbb{E}_S[\mathbf{e}_j^\top((\mathbf{X} - \mathbb{E}_S[\mathbf{X}])(\mathbf{X} - \mathbb{E}_S[\mathbf{X}])^\top)\mathbf{e}_j] =$$

- $$= \sum_{j=M+1}^{N}\mathbf{e}_j^\top\mathbb{E}_S[((\mathbf{X} - \mathbb{E}_S[\mathbf{X}])(\mathbf{X} - \mathbb{E}_S[\mathbf{X}])^\top)]\mathbf{e}_j = \sum_{j=M+1}^{N}\mathbf{e}_j^\top\mathbf{C}_S(\mathbf{X})\mathbf{e}_j$$

- Търсим ортонормиран базис $\mathbf{e}_j$, който минимизира ε^2. Ще използваме множители на Лагранж, за да си осигурим $\mathbf{e}_j \cdot \mathbf{e}_j = 1$. Дефинираме $N - M$ функции: $g_j(\mathbf{e}_j) = \mathbf{e}_j \cdot \mathbf{e}_j$.
- $$\frac{\partial}{\partial \mathbf{e}_j}\left(\varepsilon^2 - \sum_{k=M+1}^{N} \lambda_k (g_k(\mathbf{e}_k) - 1)\right) = \frac{\partial}{\partial \mathbf{e}_j}\left(\sum_{k=M+1}^{N} \mathbf{e}_k^\top \mathbf{C}_S(\mathbf{X})\mathbf{e}_k - \sum_{k=M+1}^{N} \lambda_k (\mathbf{e}_k \cdot \mathbf{e}_k - 1)\right) =$$
  $$= (\mathbf{C}_S(\mathbf{X}) + \mathbf{C}_S(\mathbf{X})^\top)\mathbf{e}_j - 2\lambda_j \mathbf{e}_j = 2\mathbf{C}_S(\mathbf{X})\mathbf{e}_j - 2\lambda_j \mathbf{e}_j = 0$$
- Така получаваме: $\mathbf{C}_S(\mathbf{X})\mathbf{e}_j = \lambda_j \mathbf{e}_j$
- $$\frac{\partial}{\partial \lambda_j}\left(\varepsilon^2 - \sum_{k=M+1}^{N} \lambda_k (g_k(\mathbf{e}_k) - 1)\right) = \mathbf{e_j} \cdot \mathbf{e_j} - 1 = 0$$
- т.е. $\mathbf{e}_j \cdot \mathbf{e}_j = 1$

Решение

- Ковариационната матрица $\mathbf{C}_S(\mathbf{X})$ е симетрична и положително дефинитна. Следователно на нея съответстват N ортогонални собствени вектори със съответни положителни собствени стойности.

- От нулирането на производните следва, че необходимо условие за търсеният базис е да се състои от собствени вектори на $\mathbf{C}_S(\mathbf{X})$. В такъв случай:
$$\varepsilon^2 = \sum_{j=M+1}^{N} \mathbf{e}_j^\top \mathbf{C}_S(\mathbf{X})\mathbf{e}_j = \sum_{j=M+1}^{N} \mathbf{e}_j^\top \lambda_j \mathbf{e}_j = \sum_{j=M+1}^{N} \lambda_j.$$

- Тъй като всички собствени стойности са неотрицателни, минималната стойност за ε^2 се получава, като за $\lambda_{M+1}\ldots, \lambda_N$ се изберат най-малките $N - M$ собствени стойности.

- Избираме безисните вектори $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_M$, така че на тях да им съответстват най-големите M собствени стойности.

План на лекцията

1. Формалности за курса (5 мин)
2. Интуиция за принципния компонентен анализ (10 мин)
3. Смятане с вектори и матрици (15 мин)
4. Свойства на ковариационната матрица (15 мин)
5. Задача за намиране на принципните компоненти (25 мин)
6. **Влагане на думи и контексти в нискомерно гъсто векторно пространство и латентен семантичен анализ (10 мин)**
7. Семантично пространствени релации (10 мин)

Влагане — проекция на контекстите

- Нека $\mathbf{X} \in \mathbb{R}^{|V| \times S}$ е терм / контекст матрица. На всеки терм съответства ред от матрицата със свързванията на терма към съответните S контекста. На всеки контекст съответства стълб от матрицата със свързванията на контекста към съответните $|V|$ терма.

- Нека предварително сме центрирали наблюденията за термовете около 0. Т.е. $\mathbb{E}[\mathbf{X}_{j,\bullet}] = 0$ за $j = 1,2,\ldots,|V|$.

- Нека първите M принципни компоненти на $\mathbf{X}\mathbf{X}^\top = \mathbf{C}(\mathbf{X}) \in \mathbb{R}^{|V| \times |V|}$ са ортонормираните вектори $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_M \in \mathbb{R}^{|V|}$.

- Дефинираме матрицата $U_M = [\mathbf{e}_1 \quad \mathbf{e}_2 \quad \cdots \quad \mathbf{e}_M] \in \mathbb{R}^{|V| \times M}$.

- Проекцията на контекстите в M-мерно пространство получаваме: $\tilde{\mathbf{X}}_M = U_M^\top \mathbf{X} \in \mathbb{R}^{M \times S}$. На всеки стълб (контекст) в $\tilde{\mathbf{X}}_M$ съпоставяме M-мерен вектор.

Влагане — проекция на термовете

- Нека $\mathbf{X} \in \mathbb{R}^{|V| \times S}$ е терм / контекст матрица. На всеки терм съответства ред от матрицата със свързванията на терма към съответните S контекста. На всеки контекст съответства стълб от матрицата със свързванията на контекста към съответните $|V|$ терма.

- Нека предварително сме центрирали наблюденията за контекстите около 0. Т.е. $\mathbb{E}[\mathbf{X}_{\bullet,j}] = 0$ за $j = 1,2,\ldots,S$.

- Нека първите M принципни компоненти на $\mathbf{X}^\top \mathbf{X} = \mathbf{C}(\mathbf{X}^\top) \in \mathbb{R}^{S \times S}$ са ортонормираните вектори $\mathbf{e}'_1, \mathbf{e}'_2, \ldots, \mathbf{e}'_M \in \mathbb{R}^S$.

- Дефинираме матрицата $V_M = \begin{bmatrix} \mathbf{e}'_1 & \mathbf{e}'_2 & \ldots & \mathbf{e}'_M \end{bmatrix} \in \mathbb{R}^{S \times M}$.

- Проекцията на термовете в M-мерно пространство получаваме: $\bar{\mathbf{X}}_M = V_M^\top \mathbf{X}^\top \in \mathbb{R}^{M \times |V|}$. На всеки стълб (терм) в $\bar{\mathbf{X}}_M$ съпоставяме M-мерен вектор.

Singular Value Decomposition (SVD)

- Съществува по-директен алгебричен метод за декомпозиция на всяка правоъгълна матрица $\mathbf{X}$.
- Може да се покаже, че ненулевите собствени стойности на $\mathbf{X}^\top\mathbf{X}$ и $\mathbf{X}\mathbf{X}^\top$ съвпадат.
- Нека $\mathbf{X}\mathbf{X}^\top = U\Lambda U^\top$ и $\mathbf{X}^\top\mathbf{X} = V\Lambda V^\top$. Тогава (грубо):
$$\mathbf{X}\mathbf{X}^\top = U\Lambda U^\top = U\sqrt{\Lambda}\sqrt{\Lambda}U^\top =$$
$$= U\sqrt{\Lambda}V^\top V\sqrt{\Lambda}U^\top = (U\sqrt{\Lambda}V^\top)(U\sqrt{\Lambda}V^\top)^\top \Longrightarrow \mathbf{X} = U\sqrt{\Lambda}V^\top$$
- Ако се ограничим до най-големите M собствени стойности получаваме: $\mathbf{X}_M = U_M\sqrt{\Lambda_M}V_M^\top$.
- Матрицата $\mathbf{X}_M$ е най-близката до $\mathbf{X}$ спрямо Фробениус норма $\|A\| = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{S} A_{i,j}^2}$ с ранк $\leq M$
- Доказателствата за SVD може да се намерят в по-задълбочените учебници по линейна алгебра.
- От изчислителна гледна точка SVD е по-ефективен.

Латентен семантичен анализ (LSA) и латентно семантично индексиране (LSI)

- Използва се когато имаме матрица терм / документ

- Нека за терм / документ матрицата $\mathbf{X} \in \mathbb{R}^{|V| \times S}$ сме намерили декомпозиция $\mathbf{X} = U\sqrt{\Lambda}V^\top$ и сме я приближили в M-мерно пространство $\mathbf{X}_M = U_M\sqrt{\Lambda_M}V_M^\top$, където $\mathbf{X}_M \in \mathbb{R}^{|V| \times S}, U_M \in \mathbb{R}^{|V| \times M}, \Lambda_M \in \mathbb{R}^{M \times M}, V_M \in \mathbb{R}^{S \times M}$

- Нека ни е дадена заявка $q \in \mathbb{R}^{|V|}$. Дефинираме $q_M \in \mathbb{R}^M$ като $q_M = U_M^\top q$

- Намираме скаларното произведение (косинусова близост) на q_M с документите от колекцията като умножим $\tilde{\mathbf{X}}_M^\top q_m$, където $\tilde{\mathbf{X}}_M = U_M^\top \mathbf{X} \in \mathbb{R}^{M \times S}$

- По аналогичен начин можем да постъпваме с термовете.

План на лекцията

1. Формалности за курса (5 мин)
2. Интуиция за принципния компонентен анализ (10 мин)
3. Смятане с вектори и матрици (15 мин)
4. Свойства на ковариационната матрица (15 мин)
5. Задача за намиране на принципните компоненти (25 мин)
6. Влагане на думи и контексти в нискомерно гъсто векторно пространство и латентен семантичен анализ (10 мин)
7. **Семантично пространствени релации (10 мин)**

Пример

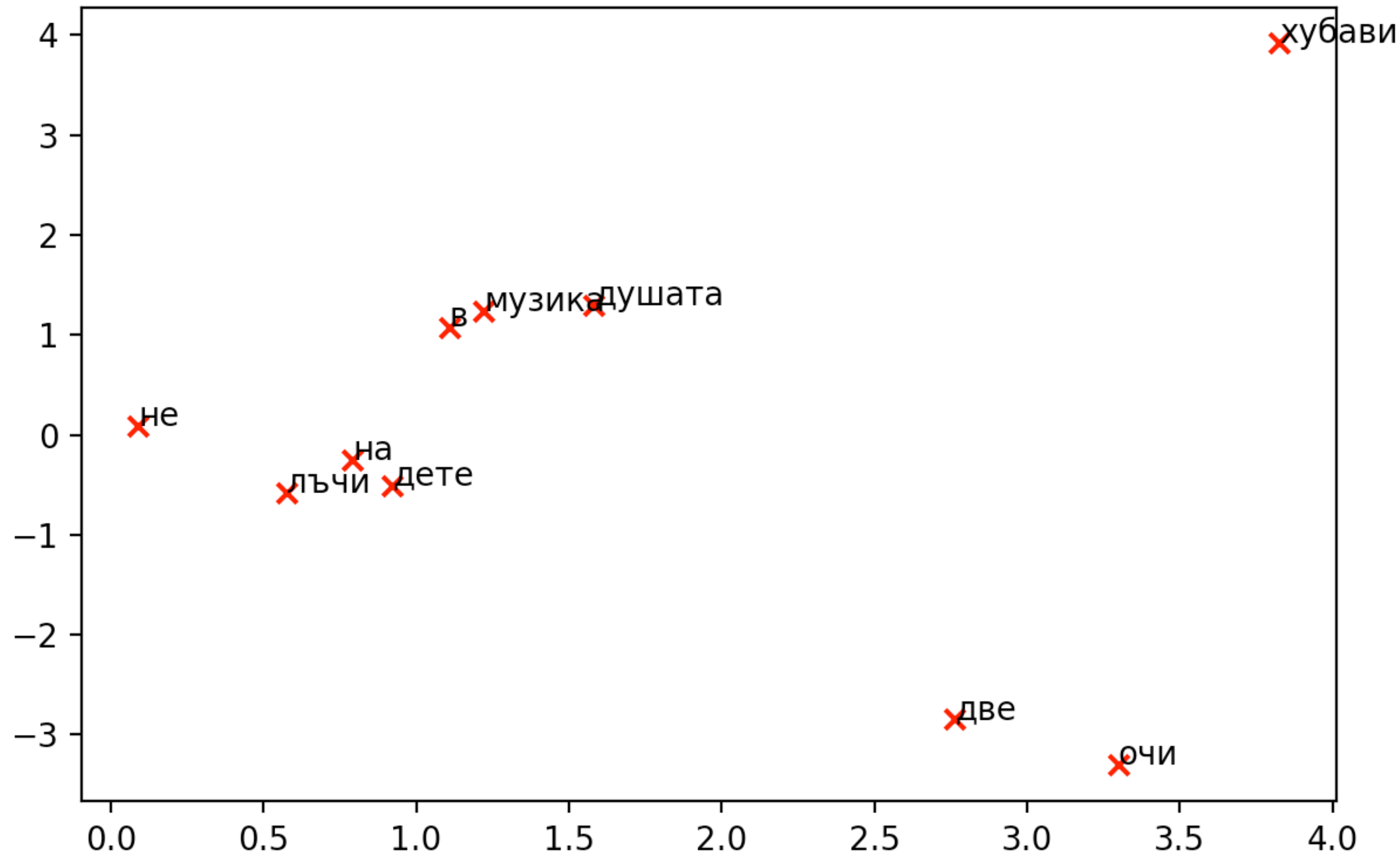
две хубави очи душата на дете
в две хубави очи музика лъчи
не искат и не обещават те
душата ми се моли
дете
душата ми се моли
страсти и неволи
ще хвърлят утре върху тях
булото на срам и грях

булото на срам и грях
не ще го хвърлят върху тях
страсти и неволи
душата ми се моли
дете
душата ми се моли
не искат и не обещават те
две хубави очи музика лъчи
в две хубави очи душата на дете

	дете	**две**	**хубави**	**очи**	**душата**	**на**	**в**	**музика**	**лъчи**	**не**
дете	0	0	0	0	2	2	1	0	0	0
две	0	0	4	0	0	0	2	0	0	0
хубави	0	4	0	4	0	0	0	0	0	0
очи	0	0	4	0	2	0	0	2	0	0
душата	2	0	0	2	0	2	0	0	0	0
на	2	0	0	0	2	0	0	0	0	0
в	1	2	0	0	0	0	0	0	1	0
музика	0	0	0	2	0	0	0	0	2	0
лъчи	0	0	0	0	0	0	1	2	0	1
не	0	0	0	0	0	0	0	0	1	0

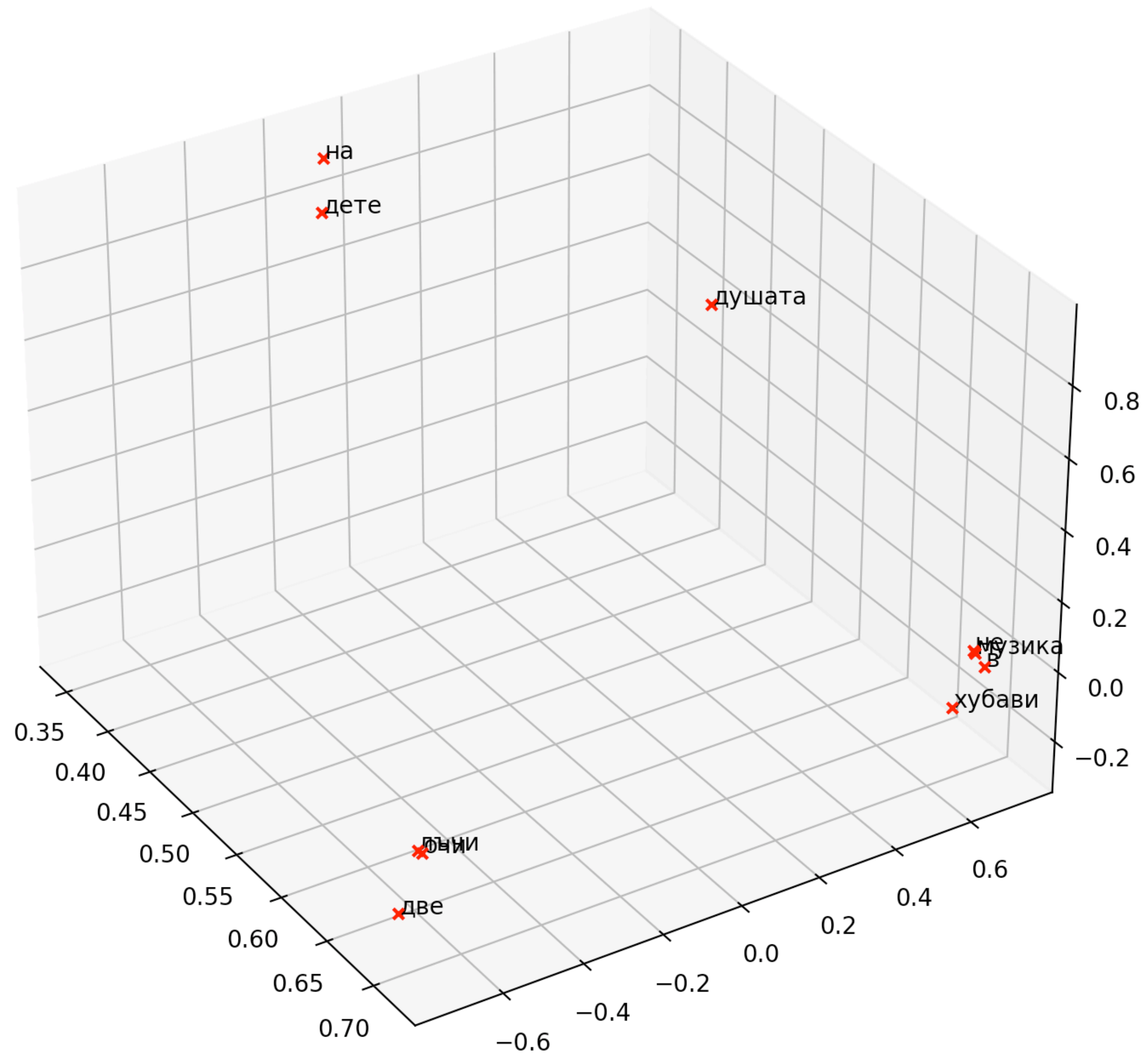
Пример — проектираме в двумерно пространство

	p1	p2
дете	0.92	-0.50
две	2.76	-2.84
хубави	3.83	3.92
очи	3.30	-3.30
душата	1.58	1.29
на	0.79	-0.25
в	1.11	1.08
музика	1.22	1.24
лъчи	0.57	-0.58
не	0.09	0.09



Пример — нормализираме до единични вектори

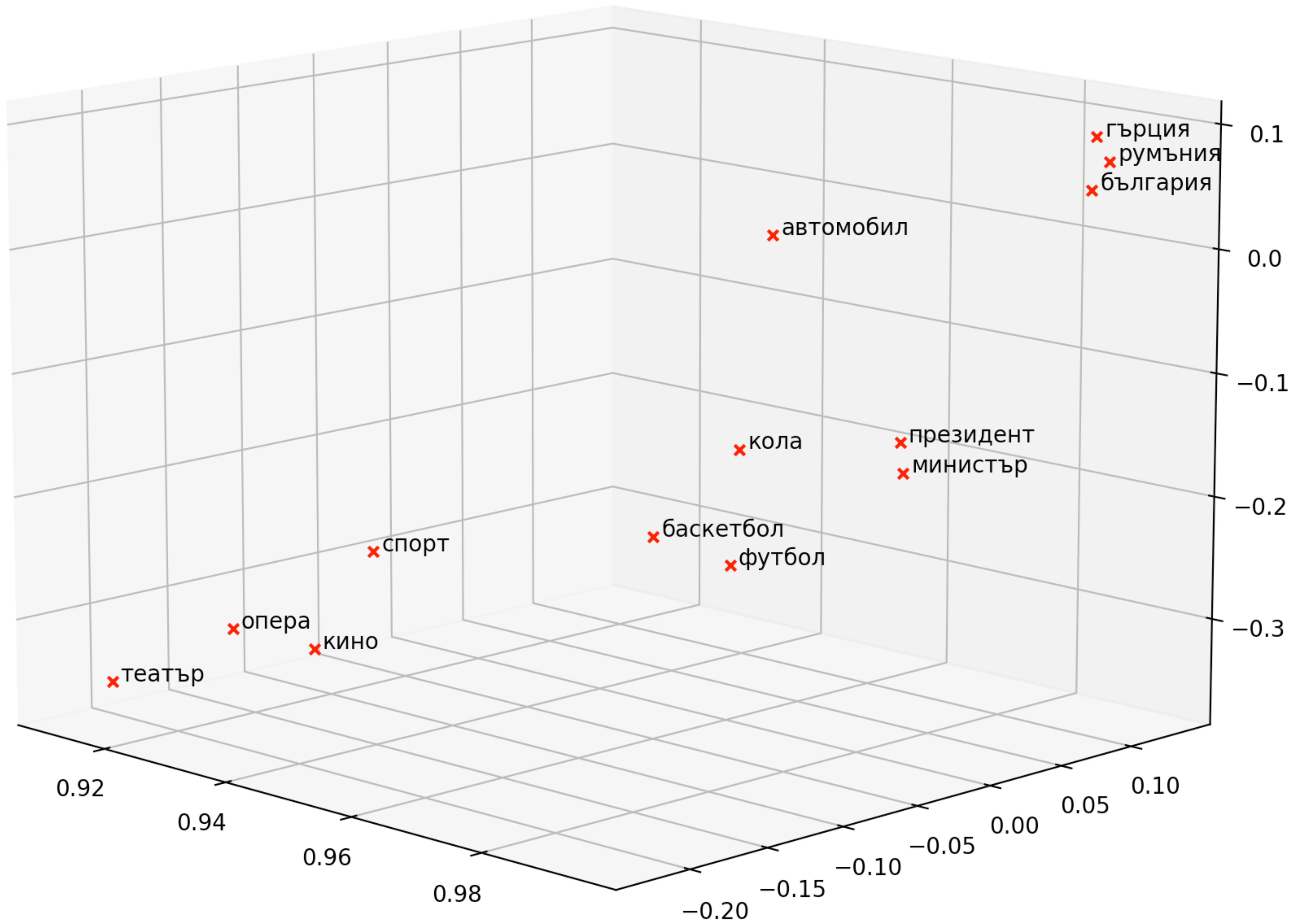
	p1	p2	p3
дете	0.39	-0.21	0.90
две	0.67	-0.69	-0.26
хубави	0.69	0.70	-0.19
очи	0.71	-0.71	0.00
душата	0.53	0.44	0.72
на	0.34	-0.11	0.93
в	0.72	0.70	0.00
музика	0.70	0.71	0.00
лъчи	0.70	-0.71	0.00
не	0.70	0.71	0.00



Семантично пространствени релации

- Косинусовата близост следва да отговаря на семантична близост следствие на сходната дистрибуцията на термовете в контекстите.
- Пример:
  - Най-близките до ***футбол*:**
    баскетбол, 0.9803
    хандбал, 0.9626
    топка, 0.9536
    волейбол, 0.9527
    телевизията, 0.9504
  - Най-близките до ***гърция*:**
    румъния, 0.9921
    българия, 0.9914
    албания, 0.9897
    хърватия, 0.9887
    македония, 0.9860

Семантично пространствени релации



Заключение

- Чрез влагането на термовете в нискомерно гъсто семантично пространство се постига:
  - изчислителна ефективност,
  - подобряване на обхвата,
  - евентуално и подобряване на прецизността.
- Проблеми с методът на принципните компоненти:
  - сложно и изчислително скъпо намиране на принципните компоненти,
  - налага се актуализиране за да се отразят нови езикови феномени.