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. Свойства на ковариационната матрица (15 мин)
4. Задача за намиране на принципните компоненти (25 мин)
5. Влагане на думи и контексти в нискомерно гъсто векторно пространство и латентен семантичен анализ (20 мин)
6. Семантично пространствени релации (15 мин)

Формалности

- Засега ще провеждаме занятията онлайн всяка сряда от 8:15 до 12:00 часа.
- Моля следете редовно обявите в Moodle за евентуални промени.
- Засега ще използваме платформата Google meet: meet.google.com/hue-frfx-axb
- Днес ще използваме едновременно слайдове и бяла дъска. Моля следете съответния екран.
- 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. Свойства на ковариационната матрица (15 мин)
4. Задача за намиране на принципните компоненти (25 мин)
5. Влагане на думи и контексти в нискомерно гъсто векторно пространство и латентен семантичен анализ (20 мин)
6. Семантично пространствени релации (15 мин)

Влагането на думи във векторното пространство на контекстите

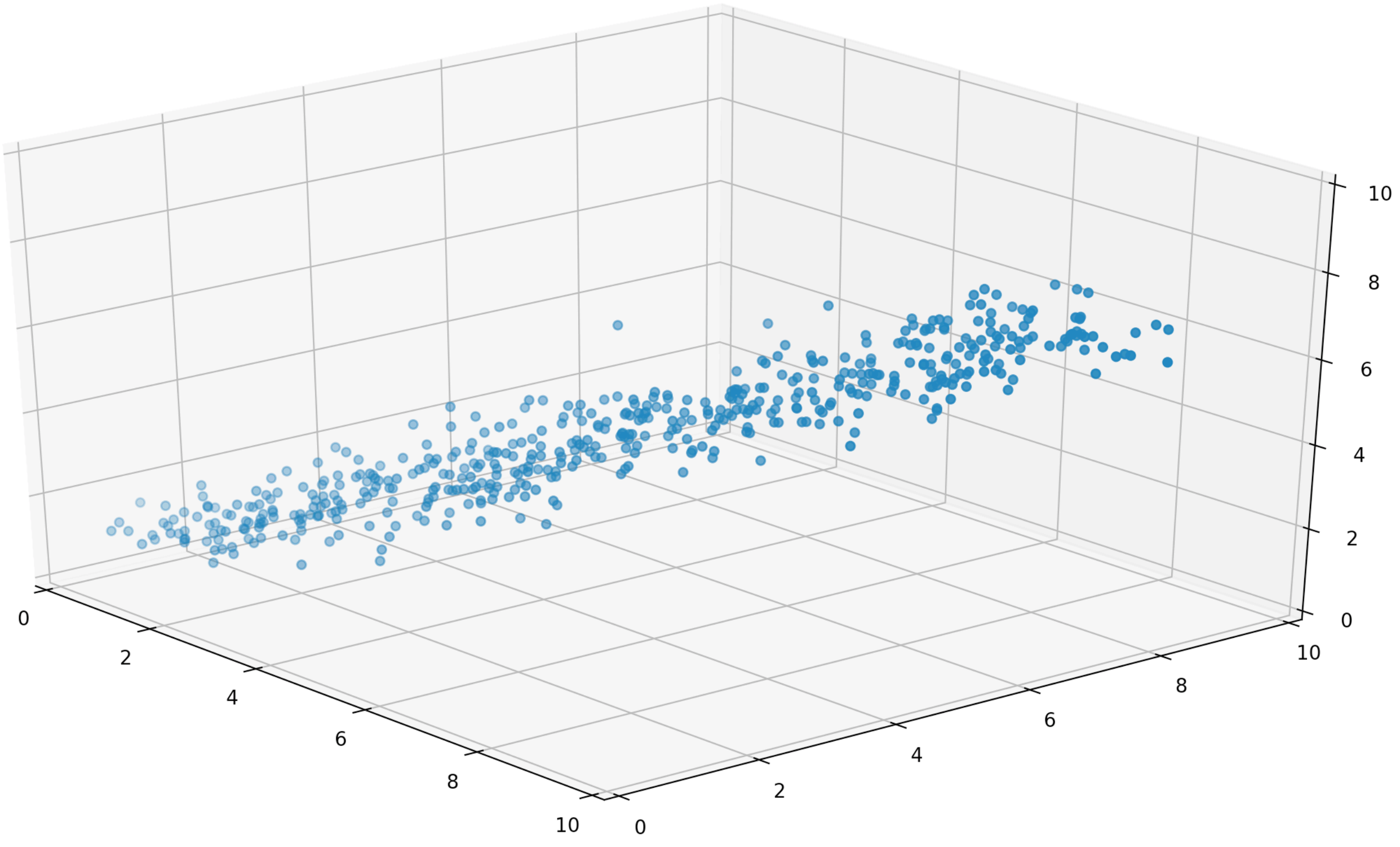
- Контекстът на дадена дума са думите, които са около нея — в рамките на параграф, изречение или фиксиран по размер прозорец.

- На всяка дума съпоставяме вектора от свързванията на думата с всеки от контекстите.

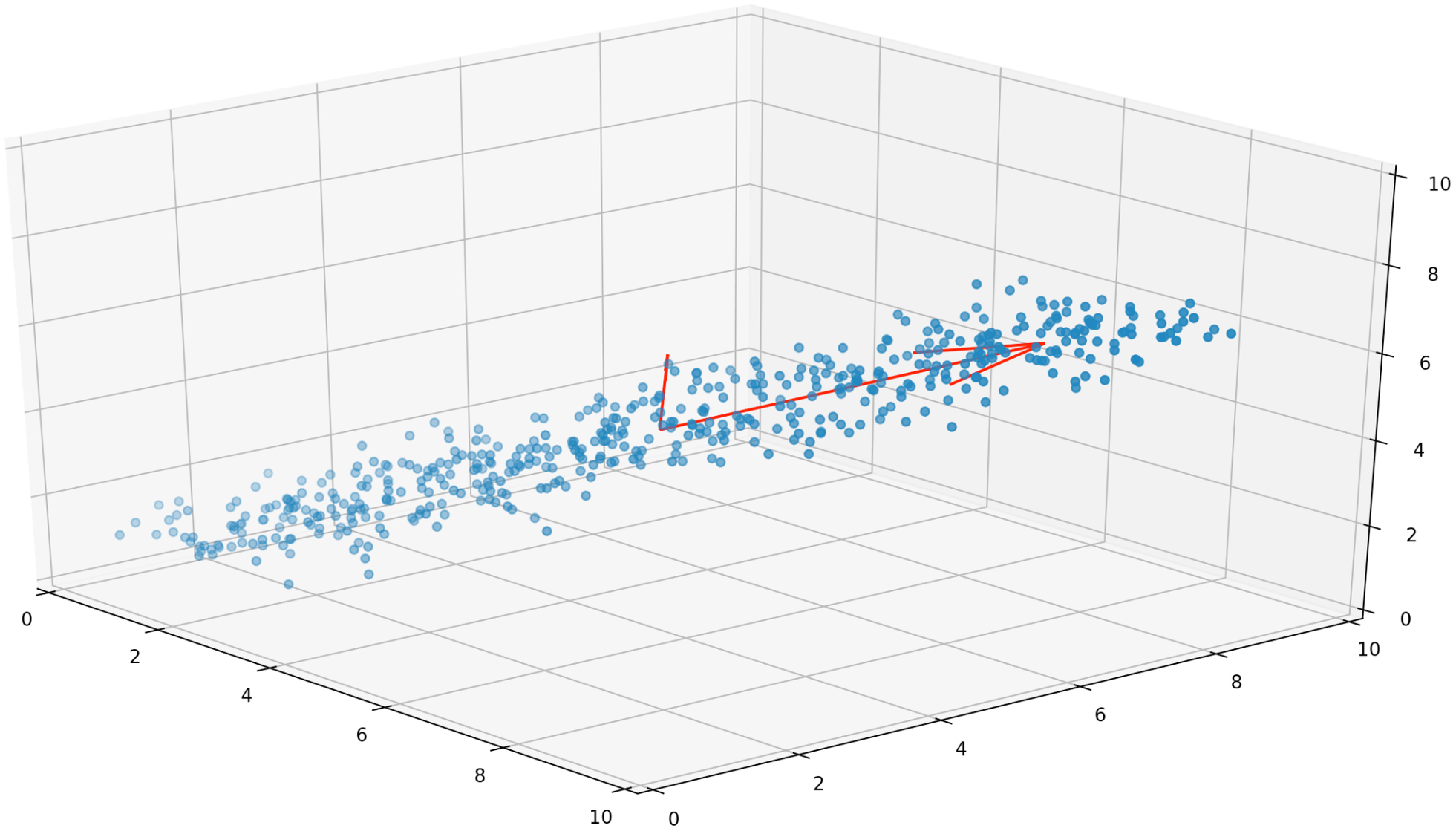
- Размерността на пространството е огромна, което води до изчислителни трудности.

- **Цел**: Да намерим влагане на векторите в нискомерно гъсто векторно пространство, което възможно най-добре да отразява разстоянията в многомерното контекстно пространство.

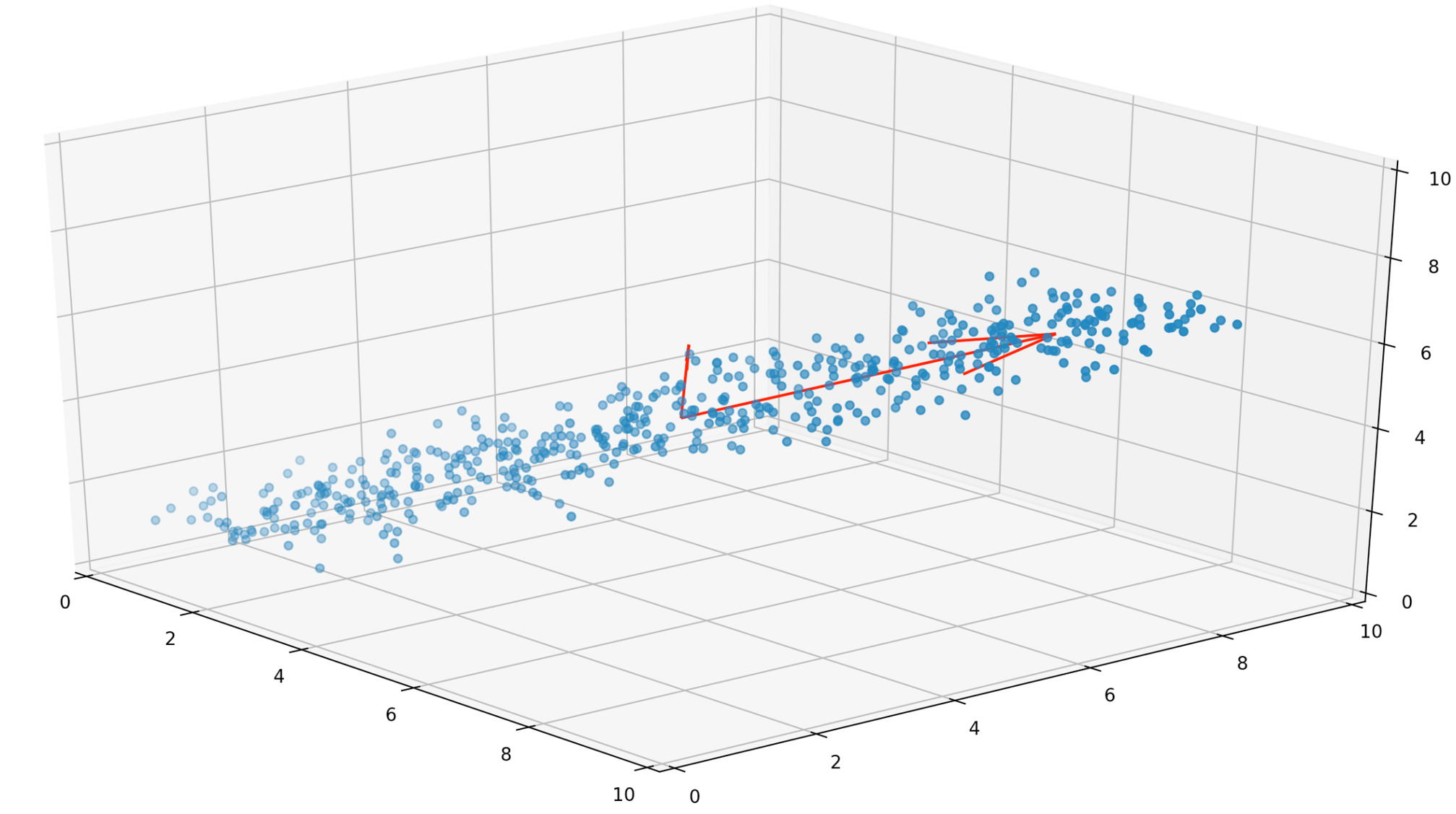
Интуитивна представа за принципен компонентен анализ

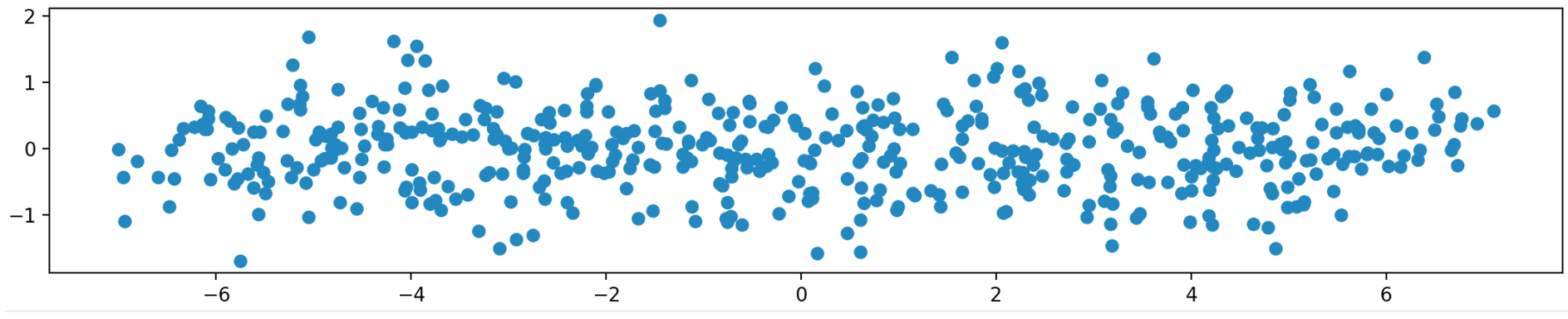


Интуитивна представа за принципен компонентен анализ



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 съответсват на линейни комбинации от оригиналните базисни вектори определени от контекстите.

- Новите вектори ще бъдат “гъсти” — почти няма да има нулеви компоненти.

- Намаляването 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. **Свойства на ковариационната матрица (15 мин)**
4. Задача за намиране на принципните компоненти (25 мин)
5. Влагане на думи и контексти в нискомерно гъсто векторно пространство и латентен семантичен анализ (20 мин)
6. Семантично пространствени релации (15 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:

  $\mathbf{C}[\mathbf{X}] = \mathrm{E}[(\mathbf{X} - \mathrm{E}[\mathbf{X}])(\mathbf{X} - \mathrm{E}[\mathbf{X}])^\mathsf{T}]$,

  т.е. $\mathbf{C}[\mathbf{X}]_{i,j} = \mathrm{Cov}(X_i, X_j)$

Свойство:

- $\mathbf{C}[\mathbf{X}] = \mathrm{E}[\mathbf{X}\mathbf{X}^\mathsf{T}] - \mathrm{E}[\mathbf{X}]\mathrm{E}[\mathbf{X}]^\mathsf{T}$

Забележка: $\mathbf{X}\mathbf{X}^\mathsf{T} = \mathbf{X} \otimes \mathbf{X}$

- Свойство: Нека $A \in \mathbb{R}^{n \times m}, B \in \mathbb{R}^{m \times k}$. Тогава: $(AB)^\top = B^\top A^\top$

- 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 $A \in \mathbb{R}^{n \times n}$ наричаме положително дефинитна, ако за всеки вектор $v \in \mathbb{R}^n$ е изпълнено: $v^\top A v \geq 0$.

- Твърдение: Всяка ковариационна матрица получена чрез емпирична ковариация е положително дефинитна.

- Доказателство:

$$\mathbf{C}(\mathbf{X}) = \mathrm{E}[(\mathbf{X} - \mathrm{E}[\mathbf{X}])^2] = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \mathrm{E}[\mathbf{X}])(\mathbf{x}_i - \mathrm{E}[\mathbf{X}])^\top$$

$$v^\top \mathbf{C}(\mathbf{X}) v = \frac{1}{n} \sum_{i=1}^{n} v^\top (\mathbf{x}_i - \mathrm{E}[\mathbf{X}])(\mathbf{x}_i - \mathrm{E}[\mathbf{X}])^\top v =$$

$$= \frac{1}{n} \sum_{i=1}^{n} ((\mathbf{x}_i - \mathrm{E}[\mathbf{X}]) v)^2 \geq 0$$

- **Теорема**: 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 $T = [\mathbf{e}_1 \quad \mathbf{e}_2 \quad \cdots \quad \mathbf{e}_n]$ 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. Свойства на ковариационната матрица (15 мин)
4. **Задача за намиране на принципните компоненти (25 мин)**
5. Влагане на думи и контексти в нискомерно гъсто векторно пространство и латентен семантичен анализ (20 мин)
6. Семантично пространствени релации (15 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 $\varepsilon^2 = \frac{1}{S} \sum_{i=1}^{S} \|\mathbf{x}^{(i)} - \hat{\mathbf{x}}^{(i)}\|^2$ 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 случайни величини. В такъв случай:
  $b_j = \mathrm{E}[\mathbf{x}^{(i)} \cdot \mathbf{e}_j] = \mathrm{E}[Y_j]$, където разглеждаме случайна величина Y_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:

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

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

- Разглеждаме вектор от N случайни величини $\mathbf{X}$ с наблюдения $\mathbf{x}^{(i)}$. В такъв случай: $Y_j = \mathbf{X}\cdot\mathbf{e}_j$.

- Заместваме и получаваме:

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

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

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

- 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: $g_j(\mathbf{e}_j) = 1 - \mathbf{e}_j \cdot \mathbf{e}_j$.
- $$\frac{\partial}{\partial \mathbf{e}_j}\left(\varepsilon^2 + \sum_{k=M+1}^{N} \lambda_k g_k(\mathbf{e}_k)\right) = \frac{\partial}{\partial \mathbf{e}_j}\left(\sum_{k=M+1}^{N} \mathbf{e}_k^\top \mathbf{C}(\mathbf{X})\mathbf{e}_k + \sum_{k=M+1}^{N} \lambda_k(1 - \mathbf{e}_k \cdot \mathbf{e}_k)\right) =$$
- $$= (\mathbf{C}(\mathbf{X}) + \mathbf{C}(\mathbf{X})^\top)\mathbf{e}_j - 2\lambda_j \mathbf{e}_j = 2\mathbf{C}(\mathbf{X})\mathbf{e}_j - 2\lambda_j \mathbf{e}_j = 0$$
- Така получаваме: $\mathbf{C}(\mathbf{X})\mathbf{e}_j = \lambda_j \mathbf{e}_j$
- $$\frac{\partial}{\partial \lambda_j}\left(\varepsilon^2 + \sum_{k=M+1}^{N} \lambda_k g_k(\mathbf{e}_k)\right) = 1 - \mathbf{e_j} \cdot \mathbf{e_j} = 0$$
- т.е. $\mathbf{e}_j \cdot \mathbf{e}_j = 1$

Решение

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

- От нулирането на производните следва, че търсеният базис се състои от собствени вектори. В такъв случай:

$$\varepsilon^2 = \sum_{j=M+1}^{N} \mathbf{e}_j^\top \mathbf{C}(\mathbf{X}) \mathbf{e}_j = \sum_{j=M+1}^{N} \mathbf{e}_j^\top \lambda_j \mathbf{e}_j = \sum_{j=M+1}^{N} \lambda_j.$$

- Тъй като всички собствени стойности са положителни 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. Свойства на ковариационната матрица (15 мин)
4. Задача за намиране на принципните компоненти (25 мин)
5. **Влагане на думи и контексти в нискомерно гъсто векторно пространство и латентен семантичен анализ (20 мин)**
6. Семантично пространствени релации (15 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. $\mathrm{E}[\mathbf{X}_{j,\bullet}] = \mathbf{0}$ 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 $U_M = [\mathbf{e}_1 \quad \mathbf{e}_2 \quad \ldots \quad \mathbf{e}_M] \in \mathbb{R}^{|V| \times M}$.

- 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. $\mathrm{E}[\mathbf{X}_{\bullet,j}] = 0$ 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 нормата $\|A\| = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{S}A_{i,j}^2}$ с ранк $< M$.
- Доказателствата за SVD може да се намерят в по-задълбочените учебници по линейна алгебра.
- От изчислителна гледна точка SVD е много 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. Свойства на ковариационната матрица (15 мин)
4. Задача за намиране на принципните компоненти (25 мин)
5. Влагане на думи и контексти в нискомерно гъсто векторно пространство и латентен семантичен анализ (20 мин)
6. **Семантично пространствени релации (15 мин)**

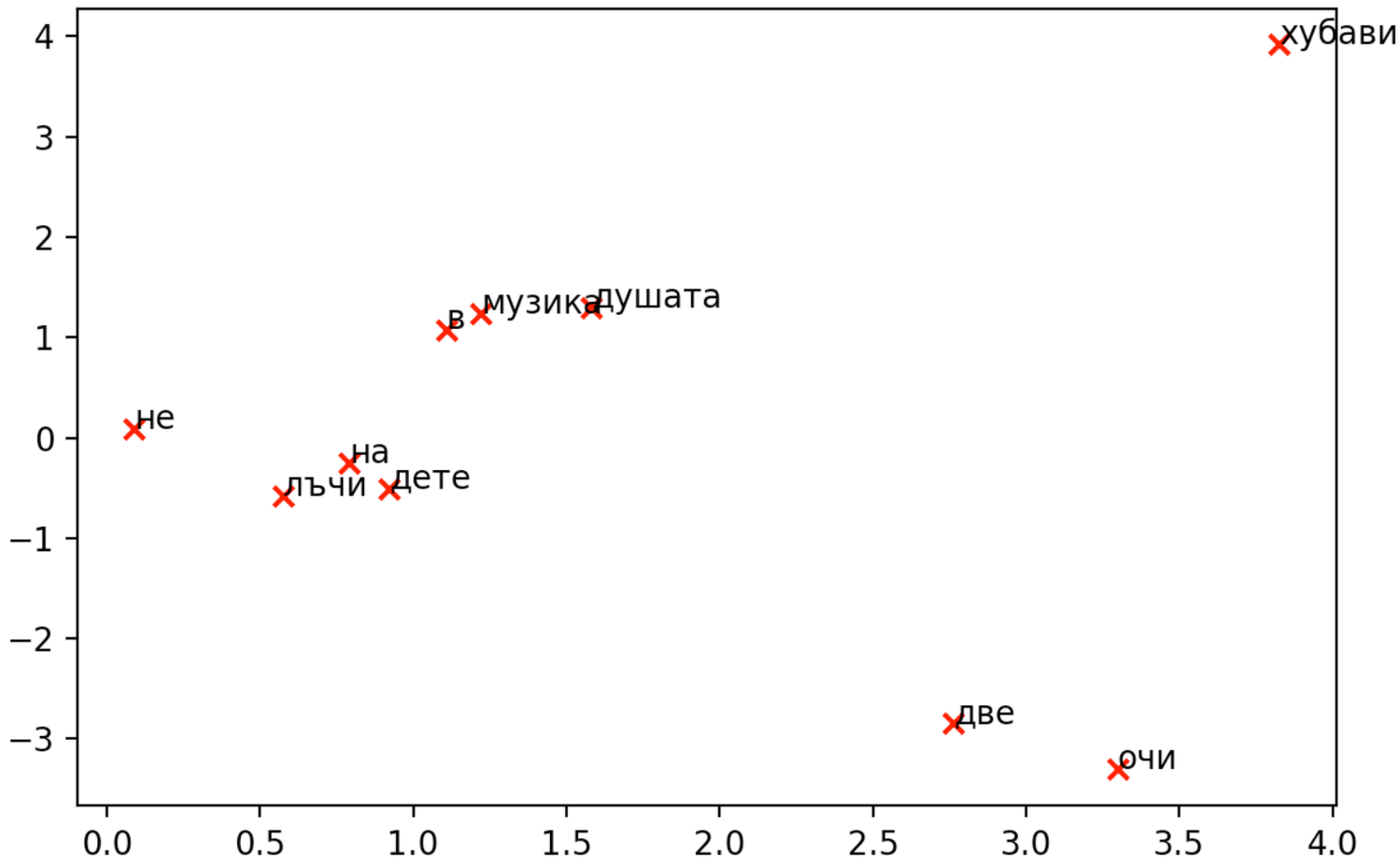
Пример

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

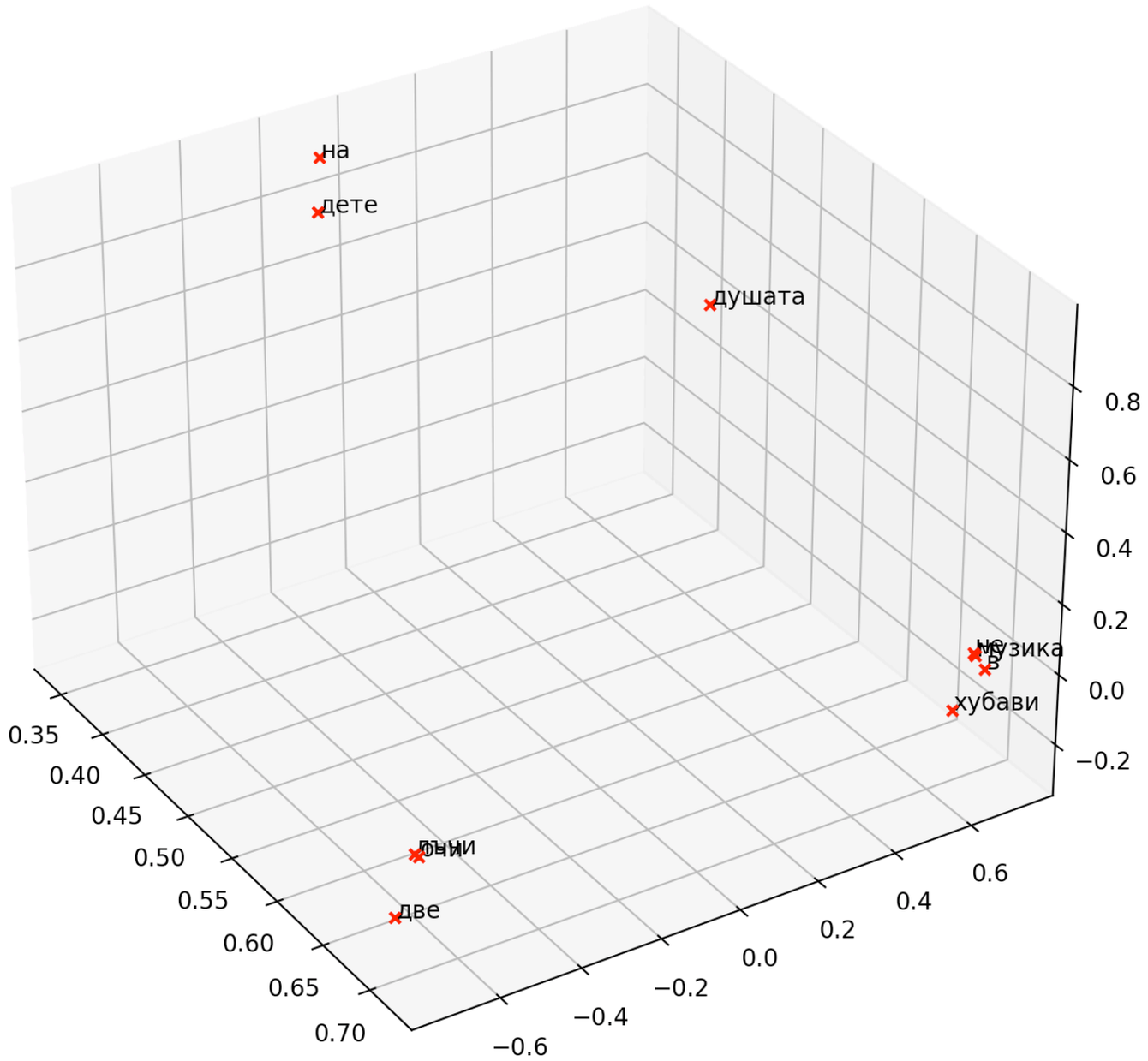
Пример — проектираме в двумерно пространство

	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

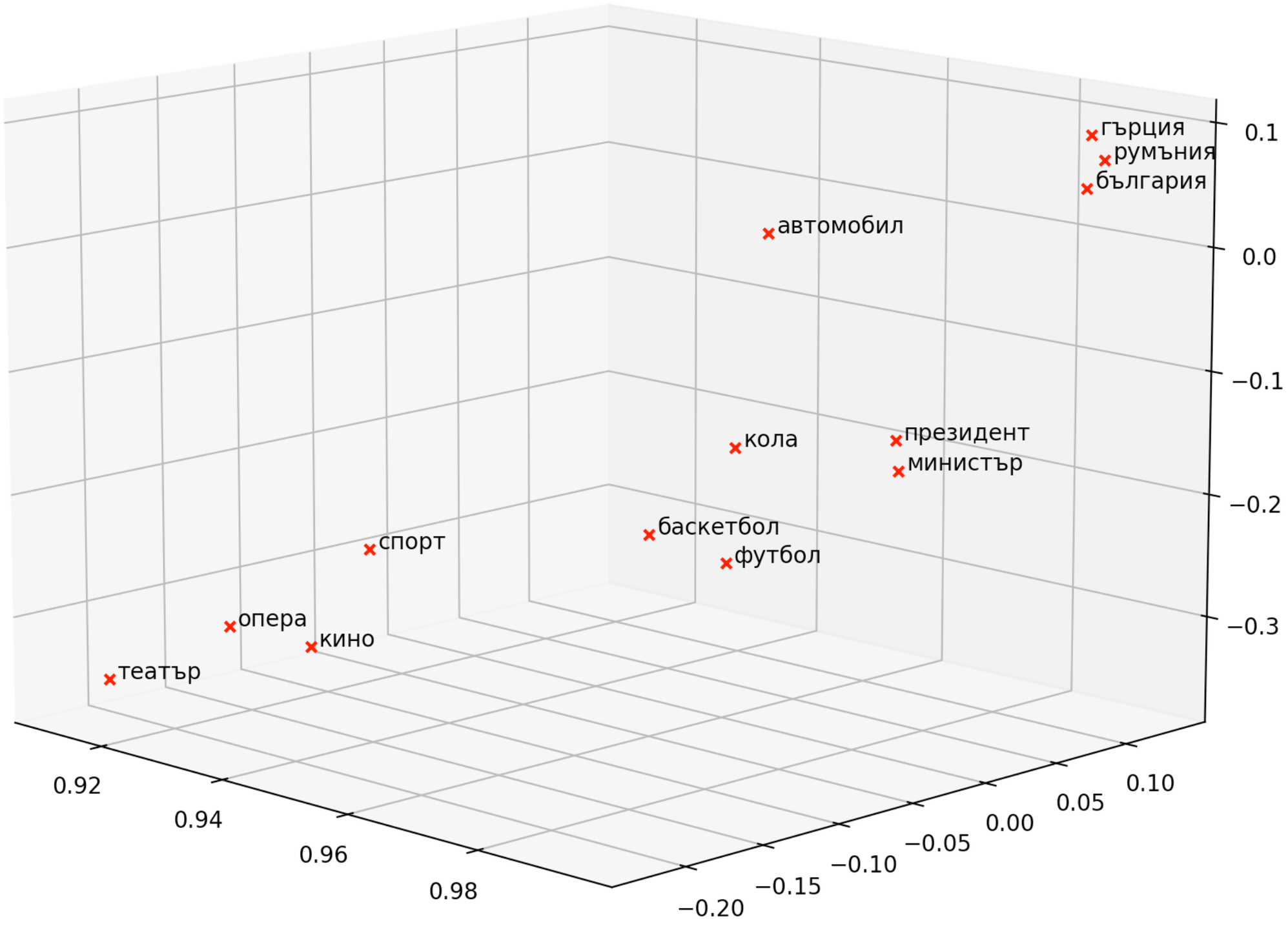
- Косинусовата близост следва да отговаря на семантична близост следствие на сходната дистрибуцията на термовете в контекстите.

- 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.