Kapitola 1

Analýza viacrozmerných objektov

1.1 Dôležité súradnice

Úloha, ktorou sa budeme zaoberať je zjednodušenie popisu skupiny viacrozmerných dátových vektorov tým spôsobom, že sa urobí ich lineárna transformácia. Touto transformáciou získame informáciu o tých súradniciach, v ktorých sa vektory najviac líšia. Na ilustráciu načrtnutej myšlienky uvedieme nasledujúci príklad.

Príklad 1.1.1 *Zákazník si chce vybrať spomedzi 3 mobilných telefónov. Parametre, podľa ktorých sa rozhoduje sú cena, výdrž baterky, veľkosť pamäte a váha mobilu. Hodnoty parametrov pre jednotlivé mobily sú v tabuľke 1.2.*

	cena	baterka	pamäť	váha
M1	3700	81	400	40
M2	3700	82	200	20
M3	3700	81	500	50

Tabuľka 1.1: Cena, výdrž baterky, pamäť a váha troch mobilných telefónov

Zákazník sa chce rozhodnúť pre kúpu telefónu len podľa jedného parametra. Ktorý z parametrov si má pre svoje rozhodovanie vybrať?

Riešenie:
Riešenie úlohy vidno už z prvého pohľadu na tabuľku. Na cene nezáleží, pretože sa v nej telefóny nelíšia. Podobne je to s výdržou baterky, tu sú síce rozdiely, ale sú veľmi malé. Podľa prvých dvoch parametrov sa rozhodnúť nedá. Tretí a štvrtý parameter už majú telefóny rozdielne. Stačí však uvažovať iba jeden z nich, lebo druhý je jeho násobkom. Na rozlíšenie medzi telefónmi stačí zákazníkovi uvažovať napríklad váhu mobilu (posledný parameter). □

V príklade 1.1.1 sme riešili úlohu nájsť také premenné, ktoré najlepšie odlíšia viacrozmerné dáta. Našli sme zjednodušenie, ktoré nám umožní zoradiť jednotlivé viacrozmerné dáta podľa jedného parametra.

V nasledujúcom príklade ukážeme ako sa dajú zmeniť viacrozmerné dáta na prehľadnejšie.

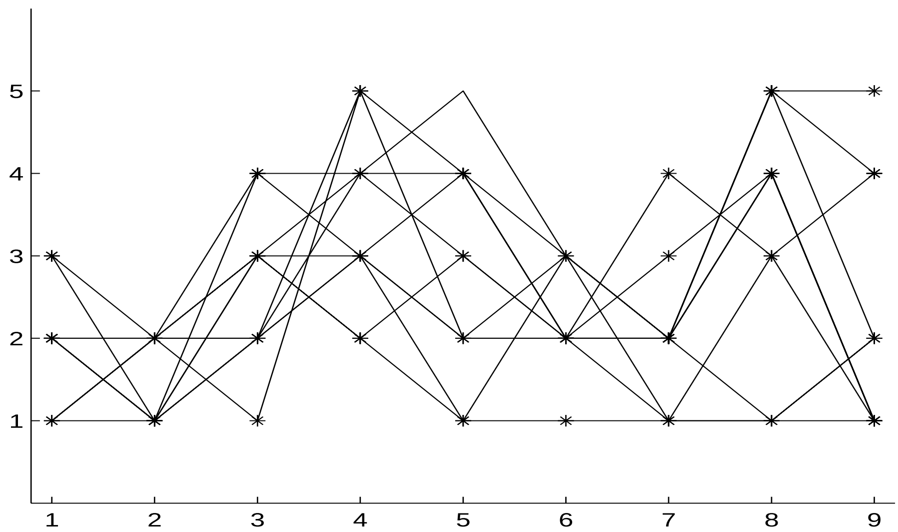
Príklad 1.1.2 *Referenti študijného oddelenia vysokej školy sa rozhodli urobiť rebríček úspešnosti študentov bakalárskeho stupňa. Ako kritérium si zvolili body za známky získané počas štúdia, ktoré sú uvedené v tabuľke* 1.2. *Najlepšie je hodnotenie 5, najhoršie je hodnotenie 1.*

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9
S_1	1	2	3	4	5	3	2	4	1
S_2	3	2	3	2	3	2	3	4	1
S_3	1	1	2	3	2	3	1	1	2
S_4	2	1	4	4	4	3	2	1	2
S_5	3	1	3	2	1	1	1	3	1
S_6	2	1	2	5	4	2	2	5	2
S_7	1	2	4	3	4	2	2	4	1
S_8	2	2	1	5	2	2	2	5	4
S_9	2	2	2	3	2	2	1	1	1
S_{10}	2	1	2	3	1	3	2	5	5
S_{11}	1	2	2	4	3	2	4	3	4

Tabuľka 1.2: Body 11-tich študentov z 9 predmetov

Ako usporiadať študentov podľa študijných výsledkov?

Riešenie:
Riešenie úlohy už nie je také evidentné ako v predchádzajúcom prípade. Načrtnutie grafov výsledkov jednotlivých študentov neurobí úlohu prehľadnejšou, ako to vidno na obrázku 1.1.



Obrázok 1.1: Hodnotenie znázornené graficky

Zo školskej praxe samozrejme každý vie, že školské výsledky možno interpretovať aj jediným číslom, ich aritmetickým priemerom. Hodnotenie reprezentované aritmetickým priemerom, bude prehľadnejšie a študentov môžeme jednoducho usporiadať. Nevýhodou tohoto postupu je, že aritmetický priemer nevypovedá o úspešnosti študenta príliš presne. V našom prípade sú aritmetické priemery jednotlivých študentov často zhodné, ako vidíme v tabuľke 1.3.

S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}
2.27	2.09	1.45	2.09	1.45	2.27	2.09	2.27	1.45	2.27	2.27

Tabuľka 1.3: Priemerné hodnotenie 11-tich študentov

Aritmetickým priemerom sme teda dosiahli rozdelenie študentov na 3 skupiny

$$A_1 = \{S_3, S_5, S_9\}, \qquad A_2 = \{S_2, S_4, S_7\} \qquad A_3 = \{S_1, S_6, S_8, S_{10}, S_{11}\}$$

Študentov z jednotlivých skupín A_1, A_2, A_3 už nedokážeme rozlíšiť. Zo skúsenosti však vieme, že rovnaký priemerný výsledok môžu dosiahnuť aj veľmi rozdielne úspešní študenti, respektíve študenti pri rôznom pracovnom výkone.

Predstavme si, že potrebujeme podstatne jemnejšie usporiadanie študentov podľa ich výsledkov. S riešením nám znovu pomôže bežná prax na vysokých školách, kde sa na ohodnotenie výsledkov štúdia používa takzvaný vážený priemer, ktorým zvýšime význam náročnejších predmetov a potlačíme jednoduchšie predmety, ktoré od študenta vyžadujú menej učenia sa. Predpokladajme, že úlohu riešime na technickej VŠ a z rozloženia úspešnosti v jednotlivých predmetoch vidíme, že predmet P_2 je zrejme matematika a predmet P_4 je telesná výchova. Študijné oddelenie teda zvýši váhu predmetu P_2 tak, že výsledky tohoto predmetu vynásobí dvomi a zníži váhu predmetu P_4 tak, že výsledky tohoto predmetu vydelí dvomi. Po takejto úprave je možné rozlíšiť študentov presnejšie, ako vidíme v tabuľke 1.4.

S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}
2.78	2.67	1.72	2.44	1.78	2.61	2.61	2.72	1.83	2.72	2.78

Tabuľka 1.4: Vážené priemerné hodnotenie 11-tich študentov

Váženým aritmetickým priemerom sme dosiahli rozloženie študentov podľa prospechu na 8 skupín

$$A_1 = \{S_3\}, \quad A_2 = \{S_5\}, \quad A_3 = \{S_9\}, \quad A_4 = \{S_4\}, \quad A_5 = \{S_6, S_7\},$$

$$A_6 = \{S_2\}, \quad A_7 = \{S_8, S_{10}\}, \quad A_8 = \{S_1, S_{11}\}$$

Študentov z jednotlivých skupín A_5, A_7 a A_8 nedokážeme rozlíšiť, ale študenti sú usporiadaní lepšie ako v predchádzajúcom prípade.

Znovu, ale prehovorí skúsenosť zo škôl, že váha predmetu sa môže zmeniť (skúša prísnejší učiteľ), alebo nezodpovedá náročnosti predmetu. Skúsme preto len zo známych výsledkov, bez toho, aby sme vedeli o aké predmety sa jedná, určiť také váhy jednotlivým predmetom, že rozlíšenie študentov pomocou váženého priemeru bude čo najlepšie. Táto úloha bude témou nasledujúcej podkapitoly. □

1.2 Váhové koeficienty pre maximálny rozptyl

Ukázali sme, že v prípade analýzy viacrozmerných dát, môžu byť niektoré súradnice viac a niektoré menej významné. Vhodnou lineárnou kombináciou môžeme z matice dát získať sadu čísel, ktoré popisujú jednotlivé riadky matice. Tieto čísla sú dostatočne rozptýlené na to, aby sa dali pôvodné dáta od seba navzájom dobre odlíšiť.

Ukážeme ako pre dáta z príkladu 1.1.1 rozdeliť váhy medzi jednotlivé parametre telefónu tak, aby výsledné lineárne kombinácie mali čo najväčší rozptyl. Predpokladáme, že dáta, ktoré ideme spracovávať, sú prvkami matice $\mathbb{M}$. Každý jej riadok sú parametre jedného mobilného telefónu.
Vektor $\mathbf{w} = (w_1, w_2, w_3, w_4)$ je vektor, ktorého zložky sú váhy jednotlivých parametrov.

Chceme, aby vektor $\mathbb{M}\,\mathbf{w}^T$ mal, čo najväčší rozptyl, aby sa vážené priemery jednotlivých telefónov od seba čo najviac líšili.

$$\mathbb{M}\,\mathbf{w}^T = \begin{pmatrix} 3700 & 81 & 400 & 40 \\ 3700 & 82 & 200 & 20 \\ 3700 & 81 & 500 & 50 \end{pmatrix} \mathbf{w}^T = \begin{pmatrix} 3700w_1 + 81w_2 + 400w_3 + 40w_4 \\ 3700w_1 + 82w_2 + 200w_3 + 20w_4 \\ 3700w_1 + 81w_2 + 500w_3 + 50w_4 \end{pmatrix}$$

Úloha 1.2.1 *Maximalizovať rozptyl budeme tak, že budeme maximalizovať súčet vzdialeností všetkých hodnôt od priemernej hodnoty. Pre pohodlnejšie výpočty budeme ako vzdialenosť používať druhú mocninu rozdielu.*

Ukážte, že rovnaký rozptyl ako vektor $\mathbb{M}\,\mathbf{w}^T$ má vektor $\mathbb{M}^\circ\,\mathbf{w}^T$, v ktorom maticu $\mathbb{M}^\circ$ dostaneme z matice $\mathbb{M}$ tak, že od každého stĺpca odčítame jeho aritmetický priemer.

$$\mathbb{M}^\circ = \begin{pmatrix} 0 & -0.3 & 33.3 & 3.3 \\ 0 & 0.7 & -166.7 & -16.7 \\ 0 & -0.3 & 133.3 & 13.3 \end{pmatrix}$$

Príklad 1.2.1 *Ukážte, že aritmetický priemer každého stĺpca matice $\mathbb{M}^\circ$ je 0.*

Riešenie:
Tvrdenie vyplýva priamo z konštrukcie matice $\mathbb{M}^\circ$. Súčet jedného stĺpca matice $\mathbb{M}$ je vlastne N-násobok jeho aritmetického priemeru. Matica $\mathbb{M}^\circ$ vznikne z $\mathbb{M}$ tak, že od každého prvku v stĺpci odpočítame aritmetický priemer celého stĺpca, spolu N-krát. Súčet jedného stĺpca matice $\mathbb{M}^\circ$ preto je:

$$N \times \text{aritmetický priemer} - N \times \text{aritmetický priemer} = 0$$

Preto priemer každého stĺpca bude 0. □

Hľadáme také $\mathbf{w}$, ktoré bude maximalizovať rozptyl vektora $\mathbb{M}\,\mathbf{w}^T$ od jeho priemernej hodnoty, resp. vektora $\mathbb{M}^\circ\,\mathbf{w}^T$ od hodnoty 0.

$$\mathbb{M}^\circ\,\mathbf{w}^T = \begin{pmatrix} 0w_1 & -0.3w_2 & 33.3w_3 & 3.3w_4 \\ 0w_1 & 0.67w_2 & -166.67w_3 & -16.67w_4 \\ 0w_1 & -0.33w_2 & 133.33w_3 & 13.33w_4 \end{pmatrix}$$

Rozptyl ROZ vektora $\mathbb{M}\,\mathbf{w}^T$ okolo hodnoty 0 je

$$ROZ = (-0.3w_2 + 33.3w_3 + 3.3w_4 - 0)^2 + (0.67w_2 - 166.67w_3 - 16.67w_4 - 0)^2 +$$

$$+(-0.33w_2 + 133.33w_3 + 13.33w_4 - 0)^2$$

Maticovo môžeme rozptyl ROZ zapísať ako:

$$\begin{aligned} &ROZ = \mathbf{w}\,\mathbb{M}^{\circ T}\,\mathbb{M}^\circ\,\mathbf{w}^T = (-0.3w_2 + 33.3w_3 + 3.3w_4)^2 \\ &+(0.67w_2 - 166.67w_3 - 16.67w_4)^2 + (-0.33w_2 + 133.33w_3 + 13.33w_4)^2 \end{aligned} \tag{1.2.1}$$

Úloha 1.2.2 *Ako vzdialenosť sme použili*

$$\|\mathbf{v} - \mathbf{u}\|^2 = (v_0 - u_0)^2 + (v_1 - u_1)^2 + \cdots + (v_{N-1} - u_{N-1})^2$$

Ukážte, že táto hodnota bude maximálna vtedy, keď bude maximálna hodnota

$$\|\mathbf{v}\| = \sqrt{(v_0 - u_0)^2 + (v_1 - u_1)^2 + \cdots + (v_{N-1} - u_{N-1})^2}$$

Vidíme, že výraz (1.2.1) dosiahne tým väčšiu hodnotu, čím je väčší vektor $\mathbf{w}$, preto musíme určiť nejakú ďalšiu podmienku, aby sme mohli hľadať maximum. Pretože v našej úlohe vektor $\mathbf{w}$ reprezentoval váhové koeficienty, budeme predpokladať, že veľkosť tohto vektora je 1.
Riešime teda úlohu, nájsť spomedzi všetkých vektorov $\mathbf{w}$ dĺžky 1 taký, ktorý maximalizuje výraz (1.2.1).

$$\mathbf{w}\mathbb{M}^{\circ T}\,\mathbb{M}^\circ\,\mathbf{w}^T \to max \qquad \text{za podmienky} \quad \|\mathbf{w}\| = 1$$

Túto úlohu môžeme riešiť metódami matematickej analýzy. Je to úloha nájsť viazaný extrém funkcie. Riešenie pomocou Lagrangeových multiplikátorov vedie na hľadanie extrému funkcie viac premenných tvaru

$$f(\mathbf{w}) = \mathbf{w}\mathbb{M}^{\circ T}\,\mathbb{M}^\circ\,\mathbf{w}^T - \lambda(\mathbf{w}\,\mathbf{w}^T - 1)$$

Hodnoty, pre ktoré táto funkcia dosiahne maximum, musia spĺňať podmienky pre stacionárne body funkcie:

$$\frac{\partial f}{\partial \mathbf{w}} = 2\mathbb{M}^{\circ T}\,\mathbb{M}^\circ\,\mathbf{w}^T - 2\lambda\,\mathbf{w}^T = 0$$

$$\frac{\partial f}{\partial \lambda} = -\mathbf{w}\,\mathbf{w}^T + 1 = 0$$

Úpravami dostaneme rovnicu pre neznámy parameter λ a neznámy vektor $\mathbf{w}$, pre ktorý má byť $\|\mathbf{w}\| = 1$.

$$\mathbb{M}^{\circ T}\,\mathbb{M}^{\circ}\,\mathbf{w}^T = \lambda\,\mathbf{w}^T \tag{1.2.2}$$

Teda hľadáme vektor $\mathbf{w}^T$ taký, že jeho vynásobenie maticou sa dá nahradiť vynásobením jedným číslom. Takýto vektor sa nazýva vlastný vektor matice a λ sa nazýva vlastné číslo tejto matice.

Vzťah 1.2.2 môžeme zapísať v maticovom tvare:

$$\left(\mathbb{M}^{\circ T}\,\mathbb{M}^{\circ} - \lambda\,\mathbb{E}\right)\,\mathbf{w}^T = \mathbf{O} \tag{1.2.3}$$

Riešením rovnice 1.2.3 je napríklad vektor $\mathbf{w} = (0, 0, \dots, 0)$. Takéto riešenie, ale nemôžeme použiť ako váhový vektor. Hľadáme preto nenulové riešenie rovnice 1.2.3. Teda chceme, aby táto rovnica mala aspoň jedno nenulové riešenie. To nastane iba vtedy, ak matica $\mathbb{M}^{\circ T}\,\mathbb{M}^{\circ} - \lambda\,\mathbb{E}$ bude singulárna a teda determinant matice bude 0.

$$\det\left(\mathbb{M}^{\circ T}\,\mathbb{M}^{\circ} - \lambda\,\mathbb{E}\right) = 0 \tag{1.2.4}$$

V našom príklade s mobilnými telefónmi dostaneme úpravami rovnice 1.2.4, rovnicu pre neznámu hodnotu λ:

$$\det\begin{pmatrix} 0-\lambda & 0 & 0 & 0 \\ 0 & 1-\lambda & -167 & -17 \\ 0 & -167 & 46667-\lambda & 4667 \\ 0 & -17 & 4667 & 467-\lambda \end{pmatrix} = 0$$

Vlastné čísla $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ sú riešením rovnice

$$(-\lambda)[(1-\lambda)(46667-\lambda)(467-\lambda) + (-17)(4667)(-17) + (-17)(-17)(4667) -$$

$$-(-\lambda)(4667)(-17) - (-17)(-17)(467-\lambda) - (-17)(46667-\lambda)(-17)] = 0$$

Riešenia sú približne

$$\lambda_1 = 0, \lambda_2 = 0, \lambda_3 = 0, \lambda_4 = 47134$$

V nasledujcej úprave nahradíme násobenie maticou násobením vlastným číslom, dosadíme $\mathbf{w}\,\mathbf{w}^T = \|\mathbf{w}\|^2 = 1$ a upravíme funkciu $f(\mathbf{w})$:

$$f(\mathbf{w}) = \mathbf{w}\mathbb{M}^{\circ T}\,\mathbb{M}^{\circ}\,\mathbf{w}^T - \lambda(\mathbf{w}\,\mathbf{w}^T - 1) = \mathbf{w}(\lambda)^2\,\mathbf{w}^T - \lambda(\mathbf{w}\,\mathbf{w}^T - 1) = (\lambda)^2$$

Túto úvahu urobíme neskôr presnejšie, aj teraz však intuitívne vidíme, že maximálny rozptyl dosiahneme pre najväčšie vlastné číslo $\lambda = 47134$ a jemu zodpovedajúci vlastný vektor, ktorý vypočítame ako riešenie rovnice 1.2.3 pre konkrétnu hodnotu λ.

Vlastný vektor, zodpovedajúci vlastnému číslu $\lambda = 47134$ je vektor
$\mathbf{w} = (0, -0.0036, 0.995, 0.0995)$. Pre najväčší rozptyl lineárnych kombinácií treba použiť v prípade mobilného telefónu váhové koeficienty

$$w_1 = 0, w_2 = -0.0036, w_3 = 0.995, w_4 = 0.0995$$

Keby sme chceli telefóny odlíšiť iba podľa jediného parametra, bol by to tretí parameter (pamäť telefónu). S týmto parametrom dosiahneme najväčšie

	cena	baterka	pamäť	váha	vážený priemer
M1	3700	81	400	40	401.7
M2	3700	82	200	20	200.7
M3	3700	81	500	50	502.2

Tabuľka 1.5: Cena, pamäť, výdrž baterky, váha a vážený priemer troch mobilných telefónov

odlíšenie medzi telefónmi. □

Poznámka:
Z riešenia príkladu o mobilných telefónoch je zrejmé, že najvýraznejší rozptyl dosiahneme, keď dáme vysokú váhu tým premenným, ktoré majú veľký rozptyl a vysoké číselné hodnoty. Ak sú číselné parametre v rôznych jednotkách, ich interpretácia môže byť skreslená.

Lepšie výsledky dá uvedená metóda v prípade, že všetky parametre budú mať rovnaký rozsah hodnôt. Takáto situácia nastane v príklade so známkami (príklad 1.1.2) .

Príklad 1.2.2 *Nájdite váhové koeficienty pre najväčší rozptyl. Použite dáta z príkladu* 1.1.2*, v ktorom sú popísané známky* 11 *študentov z* 9 *predmetov.*

Riešenie:
Podobne ako v predchádzajúcej úlohe vytvoríme z dát maticu nameraných hodnôt.

$$M = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 3 & 2 & 4 & 1 \\ 3 & 2 & 3 & 2 & 3 & 2 & 3 & 4 & 1 \\ 1 & 1 & 2 & 3 & 2 & 3 & 1 & 1 & 2 \\ 2 & 1 & 4 & 4 & 4 & 3 & 2 & 1 & 2 \\ 3 & 1 & 3 & 2 & 1 & 1 & 1 & 3 & 1 \\ 2 & 1 & 2 & 5 & 4 & 2 & 2 & 5 & 2 \\ 1 & 2 & 4 & 3 & 4 & 2 & 2 & 4 & 1 \\ 2 & 2 & 1 & 5 & 2 & 2 & 2 & 5 & 4 \\ 2 & 2 & 2 & 3 & 2 & 2 & 1 & 1 & 1 \\ 2 & 1 & 2 & 3 & 1 & 3 & 2 & 5 & 5 \\ 1 & 2 & 2 & 4 & 3 & 2 & 4 & 3 & 4 \end{pmatrix}$$

Z centrovanej matice M° vytvoríme súčin matíc $\mathbb{M}^{\circ T}\mathbb{M}^\circ$:

$$\begin{pmatrix} 5.64 & -0.91 & 0.27 & -3.09 & -4.36 & -2.45 & -1 & 1.55 & -1.64 \\ -0.91 & 2.73 & -0.82 & 0.27 & 2.09 & -0.64 & 2 & 1.36 & -1.09 \\ 0.27 & -0.82 & 8.55 & -4.18 & 4.27 & 1.09 & 0 & -1.91 & -5.27 \\ -3.09 & 0.27 & -4.18 & 10.73 & 5.91 & 1.64 & 2 & 4.64 & 6.09 \\ -4.36 & 2.09 & 4.27 & 5.91 & 17.64 & 2.55 & 4 & 1.55 & -7.64 \\ -2.45 & -0.64 & 1.09 & 1.64 & 2.55 & 4.18 & 0 & -1.82 & 2.45 \\ -1 & 2 & 0 & 2 & 4 & 0 & 8 & 5 & 5 \\ 1.55 & 1.36 & -1.91 & 4.64 & 1.55 & -1.82 & 5 & 26.18 & 8.45 \\ -1.64 & -1.09 & -5.27 & 6.09 & -7.64 & 2.45 & 5 & 8.45 & 21.64 \end{pmatrix}$$

Nájdeme vlastné čísla matice $\mathbb{M}^{\circ T}\mathbb{M}^\circ$:

$$\lambda_1 = 0, \lambda_2 = 0.1, \lambda_3 = 0.17, \lambda_4 = 0.37, \lambda_5 = 0.66, \lambda_6 = 0.94, \lambda_7 = 1.8,$$

$$\lambda_8 = 2.69, \lambda_9 = 3.81$$

Nájdeme vlastný vektor zodpovedajúci najväčšiemu vlastnému číslu λ:

$$\mathbf{w} = (-0.02, 0.02, -0.21, 0.28, -0.09, 0.01, 0.22, 0.66, 0.62)$$

Váhy jednotlivých vyučovacích predmetov, použiteľné pre vážený priemer s najväčším rozptylom sú:

$$w_1 = -0.02, w_2 = 0.02, w_3 = -0.21, w_4 = 0.28, w_5 = -0.09, w_6 = 0.01,$$

$$w_7 = 0.22, w_8 = 0.66, w_9 = 0.62$$

Vážené priemery jednotlivých riadkov matice M, resp. priemery známok pre jednotlivých študentov sú

3.8, 3.58, 2.4, 2.28, 2.63, 5.6, 3.39, 7.25, 1.76, 6.98, 5.82,

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	priem	vp1	vp2
S_1	1	2	3	4	5	3	2	4	1	2.27	2.78	3.8
S_2	3	2	3	2	3	2	3	4	1	2.09	2.67	3.58
S_3	1	1	2	3	2	3	1	1	2	1.45	1.72	2.4
S_4	2	1	4	4	4	3	2	1	2	2.09	2.44	2.28
S_5	3	1	3	2	1	1	1	3	1	1.45	1.78	2.63
S_6	2	1	2	5	4	2	2	5	2	2.27	2.61	5.6
S_7	1	2	4	3	4	2	2	4	1	2.09	2.61	3.39
S_8	2	2	1	5	2	2	2	5	4	2.27	2.72	7.25
S_9	2	2	2	3	2	2	1	1	1	1.45	1.83	1.76
S_{10}	2	1	2	3	1	3	2	5	5	2.27	2.72	6.98
S_{11}	1	2	2	4	3	2	4	3	4	2.27	2.78	5.82

Tabuľka 1.6: Body 11-tich študentov z 9 predmetov a zodpovedajúce vážené priemery

V tabuľke 1.6 je popísaný prospech študentov pomocou aritmetického priemeru (stĺpec „priem"), váženého priemeru s odhadnutými váhami (stĺpec „vp1") a váženého priemeru s vypočítanými váhami (stĺpec „vp2"). Výpočet váženého priemeru pomocou vlastného vektora rozptylovej matice zabezpečí nielen najväčší rozptyl, ale aj vyššiu váhu tých hodnôt, ktoré majú väčší rozptyl. Konkrétne v tomto príklade to znamená, že ak majú všetci študenti dobrú známku z telesnej výchovy, tak vo váženom priemere vystupuje telesná výchova s nízkym váhovým koeficientom. Naopak ak sú z informatiky dobré aj zlé známky, tak informatika bude mať vysoký váhový koeficient. A nakoniec, ak sú z matematiky iba zlé známky, tak matematika bude mať nízky váhový koeficient.

Pri konkrétnom spracovaní dát treba uvážiť všetky výhody aj nevýhody použitia tejto metódy. V každom prípade však touto metódou môžeme zistiť, ktoré parametre sú pre rozlíšenie dát významné a ktoré menej významné. □

Ukázali sme, ako vybrať váhové koeficienty tak, aby vážený priemer jednotlivých skupín dát mal čo najväčší rozptyl. Zároveň sme však videli, že v prípade rôznych rozsahov jednotlivých parametrov, budú mať väčšiu váhu tie parametre, ktoré sú zadané v jednotkách s väčšími hodnotami. Ak uvedieme váhu v gramoch dosiahneme väčší rozptyl, než keď tú istú váhu uvedieme v kilogramoch.

Z uvedeného dôvodu je zrejmé, že metóda hľadania váhových koeficientov pre najväčší rozptyl bude dávať rozumné výsledky v tom prípade, keď budú mať rôzne parametre rovnaký rozsah hodnôt. To je dôvod, prečo metóda dáva realistickejšie výsledky v príklade o školskom prospechu (1.1.2). V príklade o mobilných telefónoch (1.1.2) bola parametrom „cena" a „výdrž baterky" telefónu priradená váha 0. V prípade ceny to bolo opodstatnené, nakoľko sa telefóny v cene nelíšili. V prípade parametra „výdrž baterky" by sme dostali podstatne väčší váhový koeficient, keby bol tento parameter zadaný v sekundách namiesto v minútach.

Príklad 1.2.3 *Pri kúpe auta si zákazník zostavil tabuľku s hodnotením* 14 *parametrov pre* 5 *rôznych áut. V tabuľke si vyznačil mieru svojej spokojnosti s jednotlivými parametrami auta bodmi od* 0 *po* 10*.*

parameter	1	2	3	4	5	6	7	8	9	10	11	12	13	14
a_1	5	2	1	1	4	10	4	4	6	7	1	2	1	0
a_2	8	9	1	3	1	5	0	9	8	4	0	4	6	7
a_3	1	1	5	4	9	6	1	7	2	0	0	5	6	5
a_4	3	9	8	6	2	4	9	5	8	9	3	3	1	9
a_5	6	10	6	8	5	3	4	8	9	1	4	7	4	2

Tabuľka 1.7: Miera spokojnosti zákazníka s 14 parametrami pre 5 áut

Určte váhy jednotlivých parametrov a vypočítajte vážený priemer pre mieru spokojnosti pre jednotlivé autá. Ktoré z áut má pri takomto výpočte najvyššie hodnotenie?

Riešenie:
K matici $\mathbb{M}$ predstavujúcej dáta zostrojíme centrovanú maticu $\mathbb{M}^\circ$.
Nájdeme vlastné čísla matice $\mathbb{M}^{\circ T}\,\mathbb{M}^\circ$. Sú to postupne čísla:

$$\lambda_0 = 0, \quad \lambda_1 = 0, \ldots, \quad \lambda_{10} = 0, \quad \lambda_{11} = 58, \quad \lambda_{12} = 103.79,$$
$$\lambda_{13} = 140.18, \quad \lambda_{14} = 200.82$$

K najväčšiemu vlastnému číslu $\lambda = 200.82$ prislúcha vlastný vektor $\mathbf{w} =$

$$= (0.13, 0.56, 0.23, 0.24, -0.31, -0.26, 0.32, 0.03, 0.33, 0.25, 0.18, 0.03, -0.12, 0.27)$$

Keď tento vektor použijeme ako váhový vektor pre hodnotenie parametrov, dostaneme iné poradie, než pri poradí áut podľa aritmetického priemeru. Z hodnôt v tabuľke 1.8 vidíme, že ak usporiadame autá podľa vážených priemerov, najlepšie hodnotené je auto a_4 a najhoršie auto a_3. Pri hodnotení podľa aritmetického priemeru je rovnako najlepšie hodnotené auto a_4 ale najhoršie hodnotené je auto a_1. □

Pri riešení príkladu 1.2.3 sme vypočítali, že matica má iba 4 nenulové vlastné čísla. V nasledujúcom odseku ukážeme, ako vypočítať ďalšie vektory $\mathbf{w}_2, \mathbf{w}_3, \ldots,$

	a_1	a_2	a_3	a_4	a_5
priemer	3.43	4.64	3.71	5.64	5.5
vážený priemer	3.66	10.53	0.38	17.9	13.05

Tabuľka 1.8: Priemerná a vážená priemerná miera spokojnosti zákazníka s 14 parametrami pre 5 áut

ktoré budú tvoriť spolu s vektorom $\mathbf{w} = \mathbf{w}_1$ ortogonálnu bázu. Na veľmi presnú aproximáciu pôvodných dát stačí malý počet vektorov tejto bázy. Na vyjadrenie sa používajú iba tie vektory, ktoré zodpovedajú nenulovým vlastným číslam matice $\mathbb{M}^{\circ T}\mathbb{M}^{\circ}$, pričom $\mathbb{M}$ je matica pôvodných hodnôt.

1.3 Analýza hlavných komponentov

S úlohou riešenou v predchádzajúcej kapitole súvisí metóda, nazývaná analýza hlavných komponentov (PCA). Túto metódu uviedol Angličan Karl Pearson v roku 1901. Rieši úlohu vyjadriť maticu nameraných hodnôt v inej báze. Táto báza bude taká, aby na dobrú aproximáciu stačil menší počet súradníc. V predchádzajúcom odseku sme ukázali, že maximálny rozptyl dosiahne taká lineárna kombinácia nameraných hodnôt, v ktorej sú ako koeficienty použité zložky vlastného vektora, prislúchajúceho najväčšiemu vlastnému číslu matice

$$\mathbb{R} = \mathbb{M}^{\circ T}\mathbb{M}^{\circ}$$

Uvedená lineárna kombinácia obsahuje najväčšie množstvo neurčitosti a teda aj informácie o nameraných hodnotách. Váhový vektor z predchádzajúcej úlohy môžeme chápať aj ako bázový vektor. Násobky tohoto vektora sú najlepšou aproximáciou spomedzi všetkých jednorozmerných aproximácií pôvodných dát. Preto sa táto lineárna kombinácia nazýva prvým hlavným komponentom. Ak prvý hlavný komponent z dát odstránime, znovu môžeme riešiť rovnakú úlohu. Riešením bude druhé najväčšie vlastné číslo a jemu prislúchajúci vlastný vektor. Priemet pôvodných vektorov do podpriestoru generovaného vlastným vektorom s druhým najväčším vlastným číslom sa nazýva druhý hlavný komponent.

Uvedený výsledok nie je samozrejmý, ani triviálny. Na jeho odvodenie sa využije nasledujúce tvrdenie.

Tvrdenie 1.3.1 *Vlastné čísla ľubovolnej matice, ktorá sa dá vyjadriť v tvare $\mathbb{A}^T\mathbb{A}$ sú reálne čísla väčšie alebo rovné 0. Vlastné vektory, ktoré prislúchajú týmto vlastným číslam tvoria ortogonálny systém vektorov (ktorý sa dá upraviť na ortonormálny).*

Celý postup PCA spočíva v niekoľkých krokoch, v ktorých sa spracuje matica nameraných hodnôt

$$\mathbb{M} = \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_m \end{pmatrix}$$

Rozmer matice $\mathbb{M}$ je $m \times n$, matica obsahuje dáta pre m objektov, každý s n parametrami.

- z matice $\mathbb{M}$ vytvoríme centrovanú maticu $\mathbb{M}^\circ$ tak, že od každého stĺpca matice odčítame jeho aritmetický priemer
- vytvoríme maticu $\mathbb{R} = \mathbb{M}^{\circ T}\mathbb{M}^\circ$, ktorej rozmer je $n \times n$
- nájdeme vlastné čísla λ_k a vlastné vektory $\mathbf{w}_k$ matice $\mathbb{R}$
- zoradíme vlastné čísla podľa veľkosti a označíme tak, aby platilo
$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$$
- zodpovedajúce vlastné vektory označíme
$$\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n$$
dá sa overiť, že tieto vektory sú ortogonálne
- riadky matice $\mathbb{M}^\circ$ označíme ako vektory $\mathbf{v}_1^\circ, \mathbf{v}_2^\circ, \ldots, \mathbf{v}_m^\circ$
- najpresnejšia aproximácia vektorov $\mathbf{v}_1^\circ, \mathbf{v}_2^\circ, \ldots, \mathbf{v}_m^\circ$ spomedzi všetkých jednorozmerných podpriestorov bude v podpriestore generovanom vektorom $\mathbf{w}_1$
- najpresnejšia aproximácia vektorov $\mathbf{v}_1^\circ, \mathbf{v}_2^\circ, \ldots, \mathbf{v}_m^\circ$ spomedzi všetkých dvojrozmerných podpriestorov bude v podpriestore generovanom vektormi $\mathbf{w}_1, \mathbf{w}_2$
- najpresnejšia aproximácia vektorov $\mathbf{v}_1^\circ, \mathbf{v}_2^\circ, \ldots, \mathbf{v}_m^\circ$ spomedzi všetkých k-rozmerných podpriestorov bude v podpriestore generovanom vektormi $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k$
- vypočítame koeficienty vektorov $\tilde{\mathbf{v}}_1^\circ, \tilde{\mathbf{v}}_2^\circ, \ldots, \tilde{\mathbf{v}}_m^\circ$ v ortogonálnej báze $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k$ ako koeficienty priemetov vektorov $\mathbf{v}^\circ$ na jednotlivé vektory tejto bázy, teda podľa vzorcov
$$\tilde{\mathbf{v}}^\circ = c_1 \cdot \mathbf{w}_1 + c_2 \cdot \mathbf{w}_2 + \cdots + c_k \cdot \mathbf{w}_k$$
$$c_j = \frac{\langle \mathbf{v}^\circ, \mathbf{w}_j \rangle}{\langle \mathbf{w}_j, \mathbf{w}_j \rangle}$$
- aproximáciu pôvodných vektorov $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$ dostaneme, ak ku vektorom $\mathbf{v}_1^\circ, \mathbf{v}_2^\circ, \ldots, \mathbf{v}_m^\circ$ pripočítame aritmetické priemery jednotlivých stĺpcov matice $\mathbb{M}$

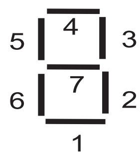
Metódu PCA odvodíme neskôr, zatiaľ ju iba ilustrujeme na úlohe zmenšenia počtu koeficientov potrebných na vyjadrenie sedemrozmerných vektorov.

Príklad 1.3.1 *Číslice na digitálnych hodinkách sú tvorené siedmimi čiarkami, ktoré buď svietia, alebo nesvietia. Na určenie o akú číslicu sa jedná použijeme* 7 *hodnôt. Hodnota* 0 *znamená, že príslušná čiarka nesvieti, hodnota* 1 *znamená, že príslušná čiarka svieti. Nájdite pomocou metódy PCA hlavné komponenty tak, aby bolo možné určiť každú z digitálnych číslic pomocou menšieho počtu koeficientov než* 7.

Riešenie:
Vyobrazenia číslic na digitálnych hodinkách, tak ako ich budeme v úlohe používať sú na obrázku 1.2. V každej číslici sú niektoré z čiarok zasvietené, iné nie. Hrubá čiara na obrázku znamená zasvietenú čiarku a bude jej priradená hodnota 1. Tenká čiara znamená, že príslušné miesto je nezasvietené a bude mu priradená hodnota 0.

Obrázok 1.2: Vyobrazenia číslic $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$

Na obrázku 1.3 sú pozície jednotlivých čiarok označené číslami.



Obrázok 1.3: Čísla $1, 2, 3, 4, 5, 6, 7$ označujú pozíciu čiarky v digitálnom čísle

Jednotlivé číslice budú reprezentované vektormi, ktorých súradnice sú hodnoty 0, alebo 1 na zodpovedajúcich miestach.

$$
\begin{aligned}
\mathbf{v}_0 &= (1,1,1,1,1,1,0)\\
\mathbf{v}_1 &= (0,1,1,0,0,0,0)\\
\mathbf{v}_2 &= (1,0,1,1,0,1,1)\\
\mathbf{v}_3 &= (1,1,1,1,0,0,1)\\
\mathbf{v}_4 &= (0,1,1,0,1,0,1)\\
\mathbf{v}_5 &= (1,1,0,1,1,0,1)\\
\mathbf{v}_6 &= (1,1,0,1,1,1,1)\\
\mathbf{v}_7 &= (0,1,1,1,0,0,0)\\
\mathbf{v}_8 &= (1,1,1,1,1,1,1)\\
\mathbf{v}_9 &= (0,1,1,1,1,0,1)
\end{aligned}
$$

Matica $\mathbb{M}$ reprezentujúca namerané dáta, v našom prípade vektory $\mathbf{v}_0, \ldots, \mathbf{v}_9$ má tvar:

$$
\mathbb{M} = \begin{pmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 0\\
0 & 1 & 1 & 0 & 0 & 0 & 0\\
1 & 0 & 1 & 1 & 0 & 1 & 1\\
1 & 1 & 1 & 1 & 0 & 0 & 1\\
0 & 1 & 1 & 0 & 1 & 0 & 1\\
1 & 1 & 0 & 1 & 1 & 0 & 1\\
1 & 1 & 0 & 1 & 1 & 1 & 1\\
0 & 1 & 1 & 1 & 0 & 0 & 0\\
1 & 1 & 1 & 1 & 1 & 1 & 1\\
0 & 1 & 1 & 1 & 1 & 0 & 1
\end{pmatrix}
$$

Vypočítame aritmetický priemer každého stĺpca matice. Dostaneme vektor aritmetických priemerov $\mathbf{m}$.

$$\mathbf{m} = (0.6, 0.9, 0.8, 0.7, 0.6, 0.4, 0.7)$$

Z matice $\mathbb{M}$ vypočítame centrovanú maticu $\mathbb{M}^\circ = \mathbb{M} - \mathbf{m}$.

$$\mathbb{M}^\circ = \begin{pmatrix} 0.4 & 0.1 & 0.2 & 0.3 & 0.4 & 0.6 & -0.7 \\ -0.6 & 0.1 & 0.2 & -0.7 & -0.6 & -0.4 & -0.7 \\ 0.4 & -0.9 & 0.2 & 0.3 & -0.6 & 0.6 & 0.3 \\ 0.4 & 0.1 & 0.2 & 0.3 & -0.6 & -0.4 & 0.3 \\ -0.6 & 0.1 & 0.2 & -0.7 & 0.4 & -0.4 & 0.3 \\ 0.4 & 0.1 & -0.8 & 0.3 & 0.4 & -0.4 & 0.3 \\ 0.4 & 0.1 & -0.8 & -0.7 & 0.4 & 0.6 & 0.3 \\ -0.6 & 0.1 & 0.2 & 0.3 & -0.6 & -0.4 & -0.7 \\ 0.4 & 0.1 & 0.2 & 0.3 & 0.4 & 0.6 & 0.3 \\ -0.6 & 0.1 & 0.2 & 0.3 & 0.4 & -0.4 & 0.3 \end{pmatrix}$$

Vypočítame maticu $\mathbb{R} = \mathbb{M}^{\circ T}\mathbb{M}^\circ$.

$$\mathbb{R} = \begin{pmatrix} 0.27 & -0.04 & -0.09 & 0.09 & 0.04 & 0.18 & 0.09 \\ -0.04 & 0.1 & -0.02 & -0.03 & 0.07 & -0.07 & -0.03 \\ -0.09 & -0.02 & 0.18 & 0.04 & -0.09 & -0.02 & -0.07 \\ 0.09 & -0.03 & 0.04 & 0.23 & -0.02 & 0.02 & 0.01 \\ 0.04 & 0.07 & -0.09 & -0.02 & 0.27 & 0.07 & 0.09 \\ 0.18 & -0.07 & -0.02 & 0.02 & 0.07 & 0.27 & 0.02 \\ 0.09 & -0.03 & -0.07 & 0.01 & 0.09 & 0.02 & 0.23 \end{pmatrix}$$

Nájdeme všetky vlastné čísla matice $\mathbb{R}$ a zoradíme ich podľa veľkosti.

$$\lambda_1 = 0.57,\ \lambda_2 = 0.37,\ \lambda_3 = 0.22,\ \lambda_4 = 0.19,\ \lambda_5 = 0.14,\ \lambda_6 = 0.05,\ \lambda_7 = 0.01$$

K vlastným číslam vypočítame zodpovedajúce vlastné vektory.

$$\begin{array}{lrrrrrrrr}
\mathbf{w}_1 & = (& 0.6, & -0.09, & -0.29, & 0.15, & 0.36, & 0.52, & 0.35) \\
\mathbf{w}_2 & = (& -0.27, & 0.32, & -0.36, & -0.46, & 0.59, & -0.28, & 0.25) \\
\mathbf{w}_3 & = (& 0.05, & 0.02, & -0.04, & 0.6, & -0.05, & -0.59, & 0.53) \\
\mathbf{w}_4 & = (& 0.11, & -0.35, & -0.27, & -0.53, & -0.57, & -0.09, & 0.43) \\
\mathbf{w}_5 & = (& -0.36, & -0.3, & 0.64, & -0.08, & 0.23, & 0.26, & 0.49) \\
\mathbf{w}_6 & = (& 0.49, & 0.62, & 0.51, & -0.27, & -0.13, & -0.1, & 0.15) \\
\mathbf{w}_7 & = (& 0.43, & -0.55, & 0.23, & -0.2, & 0.36, & -0.48, & -0.27)
\end{array}$$

(Presvedčíme sa, že vektory $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_7$ sú ortogonálne.)
Vypočítame koeficienty vektorov $\tilde{\mathbf{v}}_0^\circ, \tilde{\mathbf{v}}_1^\circ, \tilde{\mathbf{v}}_2^\circ, \tilde{\mathbf{v}}_3^\circ, \tilde{\mathbf{v}}_4^\circ, \tilde{\mathbf{v}}_5^\circ, \tilde{\mathbf{v}}_6^\circ, \tilde{\mathbf{v}}_7^\circ, \tilde{\mathbf{v}}_8^\circ, \tilde{\mathbf{v}}_9^\circ$, v ortogonálnej báze $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k$ ako koeficienty priemetov vektorov $\mathbf{v}^\circ$ do tejto bázy, teda podľa vzorca

$$\tilde{\mathbf{v}}^\circ = c_1 \cdot \mathbf{w}_1 + c_2 \cdot \mathbf{w}_2 + \cdots + c_k \cdot \mathbf{w}_k$$

$$c_j = \frac{\langle \mathbf{v}^\circ, \mathbf{w}_j \rangle}{\langle \mathbf{w}_j, \mathbf{w}_j \rangle}$$

V prípade, že do bázy zoberieme iba prvý vlastný vektor, vyjadríme vektory ako násobky vektora $\mathbf{w}_1 = (0.6, -0.09, -0.29, 0.15, 0.36, 0.52, 0.35)$ dostaneme najpresnejší jednorozmerný odhad centrovaných vektorov $\mathbf{v}_0^\circ, \ldots, \mathbf{v}_9^\circ$.

$$\tilde{\mathbf{v}}_0^\circ = c_{01} \cdot \mathbf{w}_1, \text{ kde } \quad c_{01} = \frac{\langle \mathbf{v}_0^\circ, \mathbf{w}_1 \rangle}{\langle \mathbf{w}_1, \mathbf{w}_1 \rangle} = \frac{\langle \mathbf{v}_0^\circ, \mathbf{w}_1 \rangle}{1} = 0.43$$

$$\tilde{\mathbf{v}}_0^\circ = (0.26, -0.04, -0.12, 0.06, 0.15, 0.22, 0.15)$$

Po pripočítaní vektora stredných hodnôt dostaneme odhad pôvodného vektora $\mathbf{v}_0$, reprezentujúceho digitálnu číslicu „0".

$$\tilde{\mathbf{v}}_0 = \tilde{\mathbf{v}}_0^\circ + \mathbf{m} = (0.86, 0.86, 0.68, 0.76, 0.75, 0.62, 0.85)$$

Pre ostatné vektory rovnakým spôsobom dostaneme hodnoty:

$$\begin{aligned}
\tilde{\mathbf{v}}_1 &= (-0.12, 1.01, 1.15, 0.52, 0.17, -0.22, 0.27)\\
\tilde{\mathbf{v}}_2 &= (0.91, 0.85, 0.65, 0.78, 0.78, 0.67, 0.88)\\
\tilde{\mathbf{v}}_3 &= (0.54, 0.9, 0.83, 0.69, 0.57, 0.35, 0.67)\\
\tilde{\mathbf{v}}_4 &= (0.3, 0.95, 0.94, 0.63, 0.42, 0.15, 0.53)\\
\tilde{\mathbf{v}}_5 &= (0.93, 0.85, 0.64, 0.78, 0.8, 0.69, 0.9)\\
\tilde{\mathbf{v}}_6 &= (1.15, 0.81, 0.53, 0.84, 0.93, 0.87, 1.02)\\
\tilde{\mathbf{v}}_7 &= (-0.04, 1, 1.1, 0.54, 0.22, -0.14, 0.33)\\
\tilde{\mathbf{v}}_8 &= (1.07, 0.83, 0.57, 0.82, 0.88, 0.8, 0.98)\\
\tilde{\mathbf{v}}_9 &= (0.4, 0.93, 0.9, 0.65, 0.48, 0.22, 0.58)
\end{aligned}$$

Tým je úloha priemetu do podpriestoru určeného prvým vlastným vektorom ukončená. Každá číslica je v tomto podpriestore určená jediným koeficientom. Keby sme chceli odhadnuté digitálne číslice zakresliť podobne ako na obrázku 1.2, museli by sme použiť odtiene šedej farby.

Teraz urobíme zjednodušenie a hodnoty zaokrúhlime a až potom zakreslíme. Označme $\tilde{\mathbb{M}}_{\mathbf{w}_1}$ maticu zaokrúhlených hodnôt, predstavujúcich priemet na prvý vlastný vektor.

$$\tilde{\mathbb{M}}_{\mathbf{w}_1} = \begin{pmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1\\
0 & 1 & 1 & 1 & 0 & 0 & 0\\
1 & 1 & 1 & 1 & 1 & 1 & 1\\
1 & 1 & 1 & 1 & 1 & 0 & 1\\
0 & 1 & 1 & 1 & 0 & 0 & 1\\
1 & 1 & 1 & 1 & 1 & 1 & 1\\
1 & 1 & 1 & 1 & 1 & 1 & 1\\
0 & 1 & 1 & 1 & 0 & 0 & 0\\
1 & 1 & 1 & 1 & 1 & 1 & 1\\
0 & 1 & 1 & 1 & 0 & 0 & 1
\end{pmatrix}$$

Vyobrazenia číslic, predstavujúce prvý hlavný komponent sú na obrázku 1.4. Táto aproximácia je dosť nepresná, správne zakreslené sú len 2 čísla z 10, úspešnosť je 20%.

V prípade, že do bázy zoberieme prvé dva vlastné vektory, vyjadríme vektory ako lineárnu kombináciu vektorov $\mathbf{w}_1$ a $\mathbf{w}_2$. Dostaneme najpresnejší dvojrozmerný odhad centrovaných vektorov $\mathbf{v}_0^\circ, \ldots, \mathbf{v}_9^\circ$.

$$\tilde{\mathbf{v}}_0^\circ = c_{01} \cdot \mathbf{w}_1 + c_{02} \cdot \mathbf{w}_2, \text{ kde } \quad c_{01} = \frac{\langle \mathbf{v}_0^\circ, \mathbf{w}_1 \rangle}{\langle \mathbf{w}_1, \mathbf{w}_1 \rangle} = 0.43 \quad \text{a} \quad c_{02} = \frac{\langle \mathbf{v}_0^\circ, \mathbf{w}_2 \rangle}{\langle \mathbf{w}_2, \mathbf{w}_1 \rangle} = -1.2$$

Obrázok 1.4: Vyobrazenia číslic $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$ ako prvý hlavný komponent

$$\tilde{\mathbf{v}}_0^\circ = (0.36,\ -0.16,\ 0.02,\ 0.25,\ -0.07,\ 0.33,\ 0.05)$$

Po pripočítaní vektora stredných hodnôt dostaneme odhad pôvodného vektora $\mathbf{v}_0$, reprezentujúceho digitálnu číslicu „0".

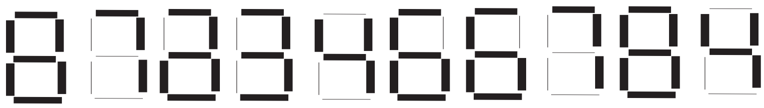
$$\tilde{\mathbf{v}}_0 = \tilde{\mathbf{v}}_0^\circ + \mathbf{m} = (0.96,\ 0.74,\ 0.82,\ 0.95,\ 0.52,\ 0.73,\ 0.75)$$

Pre ostatné vektory rovnakým spôsobom dostaneme hodnoty $\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_9$. Vypočítané hodnoty zaokrúhlime a potom zakreslíme.

Označme $\tilde{\mathbb{M}}_{\mathbf{w}_1,\mathbf{w}_2}$ maticu zaokrúhlených hodnôt, predstavujúcich priemet do podpriestoru určeného prvými dvomi vlastnými vektormi.

$$\tilde{\mathbb{M}}_{\mathbf{w}_1,\mathbf{w}_2} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$$

Vyobrazenia číslic, vyjadrených ako lineárna kombinácia prvých dvoch vlastných vektorov sú na obrázku 1.5. Táto aproximácia je trochu presnejšia, správne zakreslených je 5 čísel z 10, úspešnosť je 50%.



Obrázok 1.5: Vyobrazenia číslic $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$ ako lineárna kombinácia prvých dvoch vlastných vektorov

Matica zaokrúhlených hodnôt, predstavujúcich priemet do podpriestoru určeného vektormi $\mathbf{w}_1$, $\mathbf{w}_2$ a $\mathbf{w}_3$ je matica $\tilde{\mathbb{M}}_{\mathbf{w}_1,\mathbf{w}_2,\mathbf{w}_3}$. Táto matica sa od matice pôvodných dát líši len na jedinom mieste.

$$\tilde{\mathbb{M}}_{\mathbf{w}_1,,\mathbf{w}_2,\mathbf{w}_3} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 0 & 1 \end{pmatrix}$$

Vyobrazenia číslic, vyjadrených ako lineárna kombinácia vektorov $\mathbf{w}_1$, $\mathbf{w}_2$ a $\mathbf{w}_3$ sú na obrázku 1.6. Táto aproximácia je znovu presnejšia, správne zakreslených je 9 čísel z 10, úspešnosť je 90%.

Obrázok 1.6: Vyobrazenia číslic $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$ ako lineárna kombinácia $\mathbf{w}_1$, $\mathbf{w}_2$ a $\mathbf{w}_3$

Matica zaokrúhlených hodnôt, predstavujúcich priemet do podpriestoru určeného vektormi vektorov $\mathbf{w}_1$, $\mathbf{w}_2$, $\mathbf{w}_3$ a $\mathbf{w}_4$ je zhodná s maticou pôvodných hodnôt $\mathbb{M}$.

Vyobrazenia číslic, vyjadrených ako lineárna kombinácia štyroch vlastných vektorov sú na obrázku 1.7. Táto aproximácia je zhodná s pôvodným obrázkom 1.2.

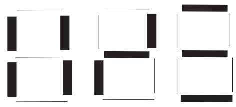
Obrázok 1.7: Vyobrazenia číslic $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$ ako lineárna kombinácia vektorov $\mathbf{w}_1$, $\mathbf{w}_2$, $\mathbf{w}_3$ a $\mathbf{w}_4$

To znamená, že je postačujúce vyjadriť vektory $\mathbf{v}_0, \ldots, \mathbf{v}_9$ pomocou 4 koeficientov v báze tvorenej z prvých štyroch vlastných vektorov $\mathbf{w}_1$, $\mathbf{w}_2$, $\mathbf{w}_3$ a $\mathbf{w}_4$. □

Príklad 1.3.2 *Vyjadrite v báze vytvorenej z prvých štyroch vlastných vektorov z príkladu* 1.3.1 *vektory*

$$\mathbf{f}_1 = (0, 1, 1, 0, 1, 1, 0)$$
$$\mathbf{f}_2 = (0, 0, 1, 0, 0, 1, 1)$$
$$\mathbf{f}_3 = (1, 0, 0, 1, 0, 0, 1)$$

Je báza z príkladu 1.3.1 *vhodná na ich reprezentáciu?*



Obrázok 1.8: Vyobrazenia vektorov $\mathbf{f}_1$, $\mathbf{f}_2$, $\mathbf{f}_3$

Riešenie:
Vektorom $\mathbf{f}_1$, $\mathbf{f}_2$, $\mathbf{f}_3$ zodpovedajú vyobrazenia, ktoré sú na obrázku 1.8.

Použijeme výpočty z príkladu 1.3.1. Vektor stredných hodnôt bude znovu vektor

$$\mathbf{m} = (0.6,\ 0.9,\ 0.8,\ 0.7,\ 0.6,\ 0.4,\ 0.7)$$

Vypočítame koeficienty priemetov centrovaných vektorov $\mathbf{f}_1^\circ$, $\mathbf{f}_2^\circ$, $\mathbf{f}_3^\circ$ do podpriestoru generovaného vektormi $\mathbf{w}_1$, $\mathbf{w}_2$, $\mathbf{w}_3$ a $\mathbf{w}_4$.

$$\tilde{\mathbf{f}}_1^\circ = c_{11}\cdot\mathbf{w}_1 + c_{12}\cdot\mathbf{w}_2 + c_{13}\cdot\mathbf{w}_3 + c_{14}\cdot\mathbf{w}_4 = (-0.39,\ 0.24,\ 0.13,\ -0.74,\ 0.35,\ 0.47,\ -0.83)$$

$$\tilde{\mathbf{f}}_2^\circ = c_{21}\cdot\mathbf{w}_1 + c_{22}\cdot\mathbf{w}_2 + c_{23}\cdot\mathbf{w}_3 + c_{24}\cdot\mathbf{w}_4 = (0.02,\ -0.43,\ -0.05,\ -0.8,\ -0.8,\ 0.25,\ -0.08)$$

$$\tilde{\mathbf{f}}_3^\circ = c_{31}\cdot\mathbf{w}_1 + c_{32}\cdot\mathbf{w}_2 + c_{33}\cdot\mathbf{w}_3 + c_{34}\cdot\mathbf{w}_4 = (0.41,\ -0.46,\ -0.21,\ 0.14,\ -0.69,\ -0.19,\ 0.73)$$

Po pripočítaní vektora $\mathbf{m}$ a zaokrúhlení hodnôt na celé čísla dostávame odhady vektorov $\mathbf{f}_1$, $\mathbf{f}_2$, $\mathbf{f}_3$.

$$\begin{aligned}
\tilde{\mathbf{f}}_1 &= (0, 1, 1, 0, 1, 1, 0)\\
\tilde{\mathbf{f}}_2 &= (1, 0, 1, 0, 0, 1, 1)\\
\tilde{\mathbf{f}}_3 &= (1, 0, 1, 1, 0, 0, 1)
\end{aligned}$$

Na rozdiel od predchádzajúceho príkladu, v tomto prípade aproximácia v štvorrozmernom podpriestore nie je taká presná, ako vidíme na obrázku 1.9. Dôvodom je, že vlastné vektory boli vytvorené k matici $\mathbb{R}$ zostavenej pre číslice, teda pre vektory rôzne od $\mathbf{f}_1$, $\mathbf{f}_2$, $\mathbf{f}_3$. Chyba však nebola príliš veľká. V jednom prípade chyba nenastala, v dvoch prípadoch bola chyba na jednom mieste. Navyše chyby spôsobili, že sa aproximované obrazce viac podobajú na číslice, než pôvodné vektory.

Obrázok 1.9: Vyobrazenia vektorov $\tilde{\mathbf{f}}_1$, $\tilde{\mathbf{f}}_2$, $\tilde{\mathbf{f}}_3$

□

Úloha 1.3.1 *Zopakujte postup z príkladu* 1.3.2 *pre rôzne vektory* $\mathbf{f}$. *Nájdite taký vektor* $\mathbf{f}$, *ktorý bude najviac odlišný od svojho priemetu do* 4*-rozmerného podpriestoru generovaného vektormi* $\mathbf{w}_1$, $\mathbf{w}_2$, $\mathbf{w}_3$ *a* $\mathbf{w}_4$ *z príkladu* 1.3.1. *Pridajte vektor* $\mathbf{f}$ *do matice* $\mathbb{M}$ *pôvodných hodnôt (použite maticu* $\mathbb{M}$ *z príkladu* 1.3.1*). Vypočítajte vlastné vektory pre takto vytvorenú maticu. Zistite, koľko koeficientov stačí na presné vyjadrenie všetkých digitálnych číslic a vektora* $\mathbf{f}$.

Úloha 1.3.2 *Vyjadrite písmená slovenskej abecedy ako obrázky vytvorené z čiernych a bielych štvorčekov v sieti veľkosti* 5×5. *Každé písmeno sa dá vyjadriť ako* $25-$*zložkový vektor, ktorého zložky sú čísla* 0 *a* 1. *Nájdite pomocou metódy PCA bázu s* 6 *vektormi, v ktorej budú jednotlivé písmená najlepšie aproximované (spomedzi všetkých* $6-$*rozmerných podpriestorov* $25-$*rozmerného priestoru).*

Metódy, ktoré sme uviedli v tejto kapitole spočívajú vo vyjadrení viacrozmerných dát ako lineárnych kombinácií vlastných vektorov matice $\mathbb{R} = \mathbb{M}^{\circ T}\mathbb{M}^{\circ}$, ktorú vypočítame z matice nameraných hodnôt $\mathbb{M}$. Namiesto nameraných hodnôt matice $\mathbb{M}$ môžeme uvažovať o všetkých možných hodnotách a ich teoretických pravdepodobnostiach. V tom prípade nahradíme maticu dát náhodným procesom. Analýze náhodných procesov sa budeme venovať v ďalšej podkapitole.

1.4 Náhodné procesy

Analýza hlavných komponentov je metóda, ktorá pracuje s konkrétnymi dátami. V teoretickom modeli dáta nahradíme hodnotami odhadnutými pomocou pravdepodobnosti výskytu týchto hodnôt. Teoretická metóda zodpovedajúca PCA sa nazýva Karhunen Loēvova transformácia. Na odvodenie tejto metódy sú potrebné pojmy a tvrdenia z teórie pravdepodobnosti a náhodných procesov, ktoré tu uvedieme bez dôkazov.

1.4.1 Pojmy a tvrdenia z teórie pravdepodobnosti

V teórii pravdepodobnosti je vybudovaný aparát, ktorý prevádza reálne udalosti do reči čísel, množín, ďalších matematických objektov a vzťahov medzi nimi. Namerané hodnoty sú reprezentované teoretickými hodnotami, sú predpovedané pravdepodobnosti nastatia jednotlivých udalostí pri opakovaní pokusu s rovnakými vstupnými hodnotami. Ak pri rovnakých vstupných podmienkach nastane vždy rovnaký výstup, jedná sa o deterministický pokus. Ak pri rovnakých vstupných podmienkach môžu nastať rôzne výsledky, jedná sa o náhodný pokus.

Modelom na popis náhodného pokusu je **pravdepodobnostný priestor** náhodných udalostí $(\Omega, \mathcal{S}, P)$. Náhodným udalostiam zodpovedajú jednotlivé množiny systému $S \in \mathcal{S}$, ktoré sú zároveň podmnožinami základnej množiny Ω.

Číselná množina S zodpovedá nejakej udalosti S_ω. Pravdepodobnostná miera P priradí každej množine S číslo $P(S)$, ktoré je nazývané pravdepodobnosť udalosti S_ω.

Tvrdenie 1.4.1 (Vlastnosti pravdepodobnosti) *Nech* A *a* B *sú náhodné udalosti a nech* $P(A)$ *a* $P(B)$ *sú pravdepodobnosti týchto udalostí, potom*

$$
\begin{aligned}
&P(\emptyset) = 0,\; P(\Omega) = 1\\
&0 \leq P(A) \leq 1\\
&ak\, A \cap B = \emptyset,\; potom\, P(A \cup B) = P(A) + P(B)\\
&P(\overline{A}) = 1 - P(A)\\
&ak\ je\ množina\, A podmnožinou\ množiny\, B\ potom\, P(A) \leq P(B)
\end{aligned}
$$

Náhodná premenná je zobrazenie systému náhodných udalostí na podmnožiny množiny reálnych čísel. jednotlivým udalostiam sú priradené čísla alebo intervaly, ktoré ich reprezentujú. Zákon rozdelenia náhodnej premennej je popis, ako je rozdelená pravdepodobnosť medzi jednotlivé čísla, prípadne intervaly.

Príklad 1.4.1 *Predpis pre náhodnú premennú (prevod medzi udalosťami a číslami) popisujúci výsledky pri hode kockou je*

$$\begin{aligned} &\mathbb{X} : \Omega \to \{1, 2, 3, 4, 5, 6\} \\ &\mathbb{X}\,(\textit{na kocke padlo číslo}\, k) \mapsto k \end{aligned}$$

Pre pravdepodobnosť výsledku $k \in \{1, 2, 3, 4, 5, 6\}$ platí $P(\mathbb{X} = k) = \frac{1}{6}$. □

Na popis pravdepodobnosti hodnôt náhodnej premennej sa používa **distribučná funkcia**. je definovaná ako reálna funkcia na celej množine reálnych čísel predpisom:

$$\begin{aligned} &F : R \to R \\ &F(x) = P(\mathbb{X} < x) \end{aligned}$$

Ak je známa distribučná funkcia náhodnej premennej, je táto úplne popísaná. Na jednoduchší popis náhodnej premennej sa používajú **číselné charakteristiky** náhodnej premennej. Pre naše potreby budeme používať strednú hodnotu a disperziu.

Nech $x_k \in \mathbb{X}$ sú hodnoty náhodnej premennej a p_k ich pravdepodobnosti. **Strednou hodnotou** náhodnej premennej $\mathbb{X}$ nazveme číslo

$$E(\mathbb{X}) = \sum_{x_k \in \mathbb{X}} x_k \cdot p_k$$

Disperziou náhodnej premennej $\mathbb{X}$ nazveme číslo

$$D(\mathbb{X}) = E(\mathbb{X} - E(\mathbb{X}))^2 = \sum_{x_k \in \mathbb{X}} x_k^2 \cdot p_k - \left(\sum_{x_k \in \mathbb{X}} x_k \cdot p_k \right)^2$$

Kovariancia dvoch náhodných premenných $\mathbb{X}, \mathbb{Y}$ je definovaná

$$cov(\mathbb{X}, \mathbb{Y}) = E\{(\mathbb{X} - E(\mathbb{X}))(\mathbb{Y} - E(\mathbb{Y}))\} = E(\mathbb{X}\,\mathbb{Y}) - E(\mathbb{X})E(\mathbb{Y})$$

S pojmom kovariancia úzko súvisí pojem korelácia.
Korelácia dvoch náhodných premenných $\mathbb{X}, \mathbb{Y}$ je číslo

$$r(\mathbb{X}, \mathbb{Y}) = E\,(\mathbb{X}\,\mathbb{Y})$$

Dve náhodné premenné $\mathbb{X}, \mathbb{Y}$ nazveme **nekorelované**, ak platí

$$r(\mathbb{X}, \mathbb{Y}) = 0$$

Procesy, ktorým sme sa venovali v predchádzajúcich kapitolách boli deterministické procesy. V **deterministickom procese** je pre každý časový okamih t_k určená jediná hodnota f_k.

$$\mathbf{f} = (f_0, f_1, \ldots, f_{N-1})$$

V tejto kapitole budeme vyšetrovať vlastnosti náhodných procesov.
V **náhodnom procese** je každému časovému okamihu t_k priradená náhodná hodnota $f(\omega,\, t_k)$.

$$\mathbf{f} = (f(\omega,\, t_0), f(\omega,\, t_1), \ldots, f(\omega,\, t_{N-1}))$$

Náhodná premenná popisuje náhodné udalosti tak, že výsledku náhodného pokusu priradí číslo. je to číselná hodnota reprezentujúca konkrétnu udalosť.
Náhodný proces popisuje náhodné udalosti tak, že výsledku náhodného pokusu priradí funkciu času $f_\omega(t_k)$, táto funkcia je jedna **realizácia** náhodného procesu:

$$f_\omega(t_k) = f(\omega = \omega_0, t_k), \quad k = 0, 1, \ldots, N-1$$

Všetky realizácie náhodného procesu v nejakom fixnom čase $t_k = t^*$, teda **rez náhodného procesu** v čase, tvorí náhodnú premennú:

$$f(\omega, t_k = t^*) = f_{t^*}(\omega), \quad \omega \in \mathcal{S}$$

Náhodný proces môžeme chápať ako vektor, ktorého zložky sú náhodné premenné:

$$\mathbf{f}(\omega, t_k) = (f(\omega, t_0), f(\omega, t_1), \ldots, f(\omega, t_{N-1}))$$

Distribučnou funkciou náhodného procesu nazveme funkciu

$$F : R^N \to R$$
$$F(x_0, x_1, \ldots, x_{N-1}) = P(f(\omega, t_0) < x_0, f(\omega, t_1) < x_1, \ldots, f(\omega, t_{N-1}) < x_{N-1})$$

Hovoríme, že dva náhodné procesy **sa rovnajú** ak sa rovnajú ich distribučné funkcie. Ak existujú parciálne derivácie distribučnej funkcie podľa všetkých premenných môžeme zaviesť pojem hustoty náhodného procesu.
Hustotou náhodného procesu nazveme funkciu N premenných

$$\varphi(x_0, x_1, \ldots, x_{N-1}) = F'(x_0, x_1, \ldots, x_{N-1}) = \left(\frac{\partial F}{\partial x_0}, \frac{\partial F}{\partial x_1}, \ldots, \frac{\partial F}{\partial x_{N-1}}\right)$$

Ak je známa distribučná funkcia náhodného procesu, je proces jednoznačne popísaný, poznáme ho celý. Takýto popis je však komplikovaný a preto často stačí ak je známa **dvojrozmerná distribučná funkcia** procesu:

$$F(x_i, x_j) = P(f(\omega, i) < x_i, f(\omega, j) < x_j), x_i, x_j \in R \quad i, j \in \{0, 1, \ldots, N-1\}$$

Popisom procesov, pri ktorom sa používa iba dvojrozmerná distribučná funkcia sa zaoberá **korelačná teória**.
Základné charakteristiky, ktoré sme uviedli pre náhodnú premennú sú stredná hodnota, disperzia , kovariancia. Tieto pojmy zavedieme aj pre náhodné procesy:
Strednou hodnotou náhodného procesu nazveme deterministický vektor

$$\mathbf{m} = (m_0, m_1, \ldots, m_{N-1}), \quad \text{kde} \quad m_k = E\left(f(\omega, k)\right), \quad k = 0, 1, \ldots, N-1$$

Disperziou náhodného procesu nazveme deterministický vektor

$$\mathbf{d} = (d_0, d_1, \ldots, d_{N-1}), \quad \text{kde} \quad d_k = E\left(f^2(\omega, k)\right) - m_k^2 \quad k = 0, 1, \ldots, N-1$$

Kovariančnou maticou náhodného procesu nazveme maticu

$$\mathbf{C} = [c_{ij}], \quad \text{kde} \quad c_{ij} = E\Big((f(\omega, i) - m_i) \, \overline{(f(\omega, j) - m_j)} \Big)$$

Na určenie strednej hodnoty a disperzie stačí použiť jednorozmernú distribučnú funkciu. Na určenie kovariančnej matice je potrebná dvojrozmerná distribučná funkcia.
Centrovaným náhodným procesom sa nazýva proces, ktorého stredná hodnota je rovná nule.

$$\mathbf{f}^\circ(\omega, t) = \mathbf{f}(\omega, t) - \mathbf{m} = (f(\omega, 0) - m_0, f(\omega, 1) - m_1, \ldots, f(\omega, N-1) - m_{N-1})$$

Kvôli zjednodušeniu a bez ujmy na všeobecnosti budeme v ďaľšom texte študovať len centrované náhodné procesy.

Tvrdenie 1.4.2 *Pre centrovaný náhodný proces majú korelačná a kovariančná matica rovnaký tvar*

$$\mathbf{C} = [c_{ij}] = \mathbf{R} = [r_{ij}] = \Big[E\left(f^\circ(\omega, i) \, \overline{f^\circ(\omega, j)} \right) \Big] \quad \textit{kde} \quad i, j = 0, 1, \ldots, N-1$$

Tvrdenie 1.4.3 *Pre prvky korelačnej matice platí*

$$r_{ij} = \overline{r_{ji}}$$

Úloha 1.4.1 *Ak je proces popísaný strednou hodnotou, disperziou a korelačnou maticou, môžeme ho rovnako dobre popísať len dvomi z týchto charakteristík. Ktorými a prečo?*

1.4.2 Priestor náhodných vektorov

Definujme N-rozmerný vektorový priestor náhodných vektorov. Prvkami sú **vektory**, ktorých jednotlivé zložky sú náhodné premenné:

$$\mathbf{f} = (f(\omega, 0), f(\omega, 1), \ldots, f(\omega, k), \ldots, f(\omega, N-1))$$

Skaláry v tomto priestore môžu byť buď reálne čísla, alebo reálne náhodné premenné.
Súčtom dvoch nezávislých náhodných vektorov **f** a **g** nazveme vektor **z**, ktorého hustotu dostaneme ako konvolúciu hustôt vektorov **f** a **g**.

$$\mathbf{z} = \mathbf{f} + \mathbf{g}$$

$$\mathbf{z} \quad \text{má hustotu} \quad \varphi_{\mathbf{z}}(\mathbf{y}) = \int\limits_{\mathbf{x} \in R^N} \varphi_{\mathbf{f}}(\mathbf{y} - \mathbf{x}) \varphi_{\mathbf{g}}(\mathbf{x}) d\mathbf{x}$$

Opačným vektorom k vektoru

$$\mathbf{f} = (f(\omega, 0), f(\omega, 1), \ldots, f(\omega, N-1))$$

je vektor

$$-\mathbf{f} = (-f(\omega, 0), -f(\omega, 1), \ldots, -f(\omega, N-1))$$

Neutrálnym prvkom vektorového priestoru je nulový náhodný vektor

$$\mathbf{0} = (0, 0, \ldots, 0)$$

Súčin vektora so skalárom je definovaný

$$c \cdot \mathbf{f} = (c \cdot f(\omega, 0), c \cdot f(\omega, 1), \ldots, c \cdot f(\omega, N-1))$$

Bez ujmy na všeobecnosti budeme v ďalšom texte uvažovať iba centrované náhodné procesy. Skalárny súčin môžeme definovať viacerými spôsobmi, výsledok môže byť náhodný, alebo nenáhodný.

Uvažujme najprv vektorový priestor, v ktorom je definovaný **nenáhodný skalár**. **Skalárny súčin** dvoch náhodných vektorov je skalár vektorového priestoru, teda nenáhodné číslo.

$$(\mathbf{f}, \mathbf{g}) = E\left(\sum_{n=0}^{N-1} f(\omega, n)\overline{g(\omega, n)}\right)$$

Dva náhodné vektory nazveme **ortogonálne** ak je ich skalárny súčin rovný 0:

$$\mathbf{f} \perp \mathbf{g} \quad \Leftrightarrow \quad (\mathbf{f}, \mathbf{g}) = 0$$

Veľkosť náhodného vektora definujeme:

$$||\mathbf{f}|| = \sqrt{(\mathbf{f}, \mathbf{f})}$$

Vzdialenosť dvoch náhodných vektorov definujeme:

$$\mathrm{d}\,(\mathbf{f}, \mathbf{g}) = ||\mathbf{f} - \mathbf{g}|| = \sqrt{(\mathbf{f} - \mathbf{g}, \mathbf{f} - \mathbf{g})}$$

Rovnosť dvoch náhodných vektorov definujeme:

$$\mathbf{f} = \mathbf{g} \quad \Leftrightarrow \quad ||\mathbf{f} - \mathbf{g}|| = 0$$

V inom vektorovom priestore môžeme definovať skalár, ktorý je **náhodná premenná**.
Skalárny súčin dvoch náhodných vektorov je skalár vektorového priestoru, teda náhodná premenná:

$$(\mathbf{f}, \mathbf{g}) = \sum_{k=0}^{N-1} f(\omega, k)\overline{g(\omega, k)}$$

V tomto prietore sú dva náhodné vektory **ortogonálne**, ak je ich skalárny súčin rovný náhodnej premennej so strednou hodnotou 0:

$$\mathbf{f} \perp \mathbf{g} \quad \Leftrightarrow \quad E(\mathbf{f}, \mathbf{g}) = 0$$

Veľkosť náhodného vektora definujeme:

$$||\mathbf{f}|| = \sqrt{E(\mathbf{f}, \mathbf{f})}$$

Vzdialenosť dvoch náhodných vektorov definujeme:

$$\mathrm{d}\,(\mathbf{f}, \mathbf{g}) = ||\mathbf{f} - \mathbf{g}|| = \sqrt{E(\mathbf{f} - \mathbf{g}, \mathbf{f} - \mathbf{g})}$$

Rovnosť dvoch náhodných vektorov definujeme:

$$\mathbf{f} = \mathbf{g} \quad \Leftrightarrow \quad ||\mathbf{f} - \mathbf{g}|| = 0$$

1.4.3 Vlastnosti priestorov s náhodnými vektormi

Tvrdenie 1.4.4 *Pre náhodné vektory platí Pytagorova veta.*

Dôkaz:
Nech sú $\mathbf{f}$ a $\mathbf{g}$ navzájom kolmé náhodné vektory, potom

$$||\mathbf{f}+\mathbf{g}||^2 = E\left(\sum_{n=0}^{N-1}(f(\omega,n)+g(\omega,n))\overline{(f(\omega,n)+g(\omega,n))}\right) =$$

$$= E\left(\sum_{n=0}^{N-1} f(\omega,n)\overline{f(\omega,n)}\right) + E\left(\sum_{n=0}^{N-1} g(\omega,n)\overline{g(\omega,n)}\right) = ||\mathbf{f}||^2 + ||\mathbf{g}||^2$$

Teda platí

$$E\left(\sum_{n=0}^{N-1}|f(\omega,n)+g(\omega,n)|^2\right) = E\left(\sum_{n=0}^{N-1}|f(\omega,n)|^2\right) + E\left(\sum_{n=0}^{N-1}|g(\omega,n)|^2\right)$$

To znamená, že stredná energia súčtu dvoch navzájom kolmých procesov sa rovná súčtu stredných energií týchto procesov □

Podmienku kolmosti môžeme nahradiť podmienkou nekorelovanosti.

Tvrdenie 1.4.5 *Stredná energia súčtu dvoch nekorelovaných procesov sa rovná súčtu stredných energií týchto procesov.*

Systém vektorov $\mathcal{B} = \{\mathbf{b_n}, n = 0, 1, \ldots, N-1\}$ s N prvkami tvaru

$$\mathbf{b_n} = (b_n(\omega,0), b_n(\omega,1), \ldots, b_n(\omega,N-1))$$

nazveme **bázou** vektorového priestoru náhodných procesov, ak sa každý náhodný proces $\mathbf{f}$ vektorového priestoru dá jednoznačne napísať ako lineárna kombinácia vektorov bázy:

$$\mathbf{f} = \sum_{n=0}^{N-1} c_n \mathbf{b_n}$$

Tvrdenie 1.4.6 *Ak pre každé dva vektory $\mathbf{b_n}, \mathbf{b_m}$ $n \neq m$ bázy $\mathcal{B}$, platí, že sú ortogonálne*

$$E(\mathbf{b_n}, \mathbf{b_m}) = 0$$

potom sú aj nekorelované a pre koeficienty c_n rozkladu $\mathbf{f} = \sum_{n=0}^{N-1} c_n \mathbf{b_n}$ platí:

a) v prípade, že skaláre sú deterministické (reálne, alebo komplexné čísla)

$$c_n = \frac{(\mathbf{f}, \mathbf{b_n})}{(\mathbf{b_n}, \mathbf{b_n})} = \frac{E\left(\sum_{k=0}^{N-1} f(\omega,k)\overline{b_n(\omega,k)}\right)}{E\left(\sum_{k=0}^{N-1}|b_n(\omega,k)|^2\right)} \qquad n = 0, 1, \ldots, N-1$$

b) v prípade, že skaláre sú náhodné (reálne, alebo komplexné náhodné premenné)

$$c_n(\omega) = \frac{(\mathbf{f}, \mathbf{b_n})}{(\mathbf{b_n}, \mathbf{b_n})} = \frac{\sum_{k=0}^{N-1} f(\omega,k)\overline{b_n(\omega,k)}}{\sum_{k=0}^{N-1}|b_n(\omega,k)|^2} \qquad n = 0, 1, \ldots, N-1$$

V prípade, že vo vektorovom priestore sú skaláre náhodné čísla, tak sa dá ukázať, že ortogonálnu bázu môžeme nahradiť deterministickou bázou. Výhodou rozkladu do deterministickej bázy je, že do rovnakej bázy je možné rozložiť nielen proces, ale aj každú jeho realizáciu.

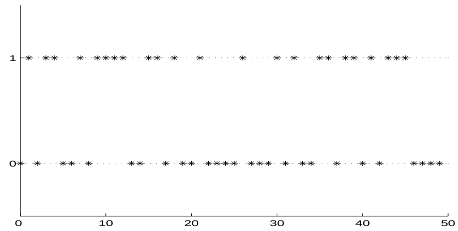
Tvrdenie 1.4.7 (Brown (1970)) *Vzdialenosť medzi procesom* $\mathbf{f}$ *a jeho rozkladom do deterministickej bázy sa rovná nule práve vtedy, ak je táto báza ortogonálna.*

$$\mathrm{d}\,(\mathbf{f}, \sum_{n=0}^{N-1} c_n(\omega) \cdot \mathbf{b}_n) = 0$$

Ak je daná ortogonálna báza tvorená deterministickými vektormi. Ak zoberieme ku vektorom tejto bázy náhodnú amplitúdu, vznikne náhodný proces.

1.5 Karhunen Loẽvov rozklad

Príklad 1.5.1 *Uvažujme proces, ktorý vznikne ako výsledok opakovaného hodu mincou. Ak padne hlava hodnota procesu bude 0, ak padne znak, hodnota procesu bude 1. Na obrázku vidíme 50 hodnôt takéhoto procesu. Žiadnou z doteraz uve-*



Obrázok 1.10: hod mincou: znak 1, hlava 0

dených metód analýzy procesov nevieme odhadnúť ďalšiu hodnotu tohoto procesu, pretože medzi hodnotami procesu nie je žiadna súvislosť a hodnoty nezávisia od času.

Zopakujme si postup pri analýze procesu $\mathbf{f}$:

- Odstránenie časovej závislosti (všetky druhy lineárnej regresie, vrátane aproximácie pomocou harmonickej bázy)

$$\tilde{\mathbf{f}}(k) = \sum_{n=0}^{M-1} c_n \varphi_n(k) \quad \text{pre} \quad k = 0, 1, \ldots, N-1$$

  Dostaneme proces $\mathbf{g} = \mathbf{f} - \tilde{\mathbf{f}}$.

- Odstránenie súvislosti medzi hodnotami (modely kĺzavých súčtov)

$$\tilde{\mathbf{g}}(t_k) = \sum_{n=0}^{L-1} d_n \mathbf{g}(k-n) \quad \text{pre} \quad k = 0, 1, \ldots, N-1$$

  Dostaneme proces $\mathbf{z} = \mathbf{g} - \tilde{\mathbf{g}}$.

Proces $\mathbf{z}$ by pri dobre urobenej analýze mal mať charakter podobný ako proces z príkladu 1.5.1, v ktorom sa nedá predpovedať žiadna hodnota. Ak sme ako prediktor použili model využívajúci jednu predchádzajúcu hodnotu

$$\tilde{f}_{k+1} = c \cdot f_k$$

Koeficient c vypočítame pomocou vzorca $c = \frac{(\mathbf{z},\mathbf{b})}{(\mathbf{b},\mathbf{b})}$. Ak je hodnota $c = 0$, tak bola predikcia urobená dobre. Hodnotu c môžeme považovať za koeficient korelácie medzi nasledujúcimi hodnotami. Ak je $c = 0$, tak hodnoty sú nekorelované. Ak je dobre odhadnutá časová zložka, tak aritmetický priemer hodnôt procesu je rovný nule. Po odstránení časovej závislosti a súvislosti medzi hodnotami ostáva ešte jedna zákonitosť, a to štatistická. Táto zákonitosť sa v príklade 1.5.1 prejaví tak, že pri veľkom počte pokusov padne rovnako veľa 0 a 1.

Ak hodnoty procesu nezávisia na čase, tak nemusíme popisovať na časovej osi, ale môžeme výsledky pokusu nanášať v jednom čase. V príklade s mincou dostaneme 7 krát hodnotu 1 a 13 krát hodnotu 0. Dostali sme vlastne 20 realizácií náhodnej premennej. Ak by ostala v procese $\mathbf{z}$ nejaká periodická časová závislosť, naskladáme na seba časové periódy, ktoré proces vykazuje a dostaneme náhodný vektor. Tento budeme analyzovať ako náhodný proces.

Tvrdenie 1.5.1 (Karhunen Loẽvov rozklad) *Nech* $\mathbf{f}$ *je centrovaný náhodný vektor z vektorového priestoru s náhodným skalárom. Potom existuje jeho rozklad do deterministickej bázy* $\mathcal{B} = \{\mathbf{b}_n, n = 0, 1, \ldots, N-1\}$*:*

$$\mathbf{f} = \sum_{n=0}^{N-1} c_n(\omega)\mathbf{b}_n$$

taký, že každé pre $n \neq m$ *sú vektory* $\mathbf{b}_n, \mathbf{b}_m$ *na seba kolmé. Navyše sú koeficienty* $c_n(\omega), c_m(\omega)$ *pre* $n \neq m$ *nekorelované.*
Bázové vektory $\mathbf{b}_n$ *sú vlastné vektory kovariančnej matice procesu* $\mathbf{f}$*, ktoré zodpovedajú vlastným hodnotám* λ_n*. Pre* λ_n *platí*

$$\lambda_n = ||\mathbf{b}_n||^2 E\left(|c_n(\omega)|^2\right) = ||\mathbf{b}_n||^2\sigma_n^2$$

Dôkaz:
Nech je proces rozložený do bázy podľa tvrdenia vety:

$$\mathbf{f} = \sum_{n=0}^{N-1} c_n(\omega)\mathbf{b}_n$$

a nech platí $(\mathbf{b}_n, \mathbf{b}_m) = 0$ a nech $c_n(\omega), c_m(\omega)$ sú nekorelované.
Vypočítame energiu n-tého bázického vektora $\mathbf{b}_n$ a označíme σ_n^2 disperziu náhodného skalára $c_n(\omega)$.

$$(\mathbf{b_n}, \mathbf{b_n}) = ||\mathbf{b}_n||^2 = \sum_{k=0}^{N-1} |b_n(k)|^2 \tag{1.5.1}$$

Z predpokladu nekorelovanosti $c_n(\omega)$, $c_m(\omega)$ dostaneme

$$E\left(c_n(\omega) \cdot \overline{c_m(\omega)}\right) = \begin{cases} 0 & \text{pre} \quad n \neq m \\ \sigma_n^2 & \text{pre} \quad n = m \end{cases} \tag{1.5.2}$$

Koeficienty $c_n(\omega)$ procesu $\mathbf{f}$ vzhľadom na ortogonálnu deterministickú bázu vypočítame:

$$c_n(\omega) = \frac{(\mathbf{f}, \mathbf{b_n})}{(\mathbf{b_n}, \mathbf{b_n})} = \frac{1}{||\mathbf{b}_n||^2} \sum_{k=0}^{N-1} f(\omega, k)\overline{b_n(k)}, \quad n = 0, \dots, N-1$$

Číslo komplexne združené ku $c_m(\omega)$ je

$$\overline{c_m(\omega)} = \frac{1}{||\mathbf{b}_m||^2} \sum_{k=0}^{N-1} \overline{f(\omega, k)}\, b_m(k) \quad m = 0, \dots, N-1$$

Ak dosadíme $c_n(\omega), \overline{c_m(\omega)}$ do vzťahu (1.5.2) dostaneme

$$E\left(c_n(\omega) \cdot \overline{c_m(\omega)}\right) = E\left(\frac{1}{||\mathbf{b}_n||^2} \sum_{k=0}^{N-1} f(\omega, k)\overline{b_n(k)}\, \frac{1}{||\mathbf{b}_m||^2} \sum_{k=0}^{N-1} \overline{f(\omega, k)} b_m(k)\right) =$$

$$= \begin{cases} 0 & \text{pre} \quad n \neq m \\ \sigma_n^2 & \text{pre} \quad n = m \end{cases}$$

Po ďalších úpravách dostaneme

$$\frac{1}{||\mathbf{b}_n||^2 ||\mathbf{b}_m||^2} \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} E\left(f(\omega, k)\overline{f(\omega, l)}\right) \overline{b_n(k)}\, b_m(l) = \begin{cases} 0 & \text{pre} \quad n \neq m \\ \sigma_n^2 & \text{pre} \quad n = m \end{cases}$$

$$\frac{1}{||\mathbf{b}_n||^2\, ||\mathbf{b}_m||^2} \sum_{k=0}^{N-1} \overline{b_n(k)} \sum_{l=0}^{N-1} r_{kl}\, b_m(l) = \begin{cases} 0 & \text{pre} \quad n \neq m \\ \sigma_n^2 & \text{pre} \quad n = m \end{cases}$$

nakoniec

$$\sum_{k=0}^{N-1} \overline{b_n(k)} \sum_{l=0}^{N-1} r_{kl} b_m(l) = \begin{cases} 0 & \text{pre} \quad n \neq m \\ \sigma_n^2\, ||\mathbf{b}_n||^2\, ||\mathbf{b}_m||^2 & \text{pre} \quad n = m \end{cases} \tag{1.5.3}$$

pre $n \neq m$ je pravá strana rovná 0, teda ostávajú len $n = m$, pričom platí

$$||\mathbf{b}_n||^2 = \sum_{k=0}^{N-1} |b_n(k)|^2 = \sum_{k=0}^{N-1} \overline{b_n(k)} b_n(k)$$

poslednú rovnosť rozšírime výrazom $||\mathbf{b}_n||^2\, \sigma_n^2$, dostaneme

$$||\mathbf{b}_n||^2\, ||\mathbf{b}_n||^2\, \sigma_n^2 = \sum_{k=0}^{N-1} \overline{b_n(k)} ||\mathbf{b}_n||^2 \sigma_n^2 b_n(k)$$

porovnaním poslednej rovnosti s (1.5.3) pre $n = m$ dostaneme

$$\sum_{k=0}^{N-1} \overline{b_n(k)} \sum_{l=0}^{N-1} r_{kl}\, b_n(l) = \sum_{k=0}^{N-1} \overline{b_n(k)}\, ||\mathbf{b}_n||^2\, \sigma_n^2\, b_n(k) \tag{1.5.4}$$

Rovnosť (1.5.4) je splnená napríklad pre také $\mathbf{b}_n$ a σ_n^2, pre ktoré platí

$$\sum_{l=0}^{N-1} r_{kl}\, b_n(l) = ||\mathbf{b}_n||^2\, \sigma_n^2\, b_n(k)$$

označme $\lambda_n = ||\mathbf{b}_n||^2 \, \sigma_n^2$

$$\sum_{l=0}^{N-1} r_{kl} b_n(l) = \lambda_n \, b_n(k)$$

alebo ak $\mathbf{b}_n = (b_n(0), b_n(1), \ldots, b_n(N-1))$ a $\mathbf{R} = [r_{k,l}]$

$$\mathbf{R}\mathbf{b}_n = \lambda_n \, \mathbf{b}_n$$

Pre nekorelované $c_n(\omega)$ dostaneme, že bázické vektory $\mathbf{b}_n$ budú vlastné vektory kovariančnej matice $\mathbf{R}$ náhodného procesu $\mathbf{f}$.
Vlastné čísla λ_n kovariančnej matice $\mathbf{R}$ určujú strednú energiu n-tej zložky $c_n(\omega)\mathbf{b}_n$ rozkladu náhodného procesu $\mathbf{f}$.

$$\lambda_n = ||\mathbf{b}_n||^2 \, \sigma_n^2 = (\mathbf{b}_n, \, \mathbf{b}_n) \, E\left(|c_n\,(\omega)|^2\right)$$

Ak zoradíme vlastné vektory podľa veľkosti vlastných čísel, ktoré im prislúchajú

$$\lambda_0 \geq \lambda_1 \geq \cdots \geq \lambda_{N-1}$$

najlepšou projekciou do M-rozmerného podpriestoru bude proces

$$\tilde{\mathbf{f}} = \sum_{n=0}^{M-1} c_n(\omega)\mathbf{b}_n$$

a chyba aproximácie bude

$$||\mathbf{f} - \tilde{\mathbf{f}}||^2 = \sum_{n=M}^{N-1} \lambda_n$$

Nekorelovanosť koeficientov $c_n(\omega)$ zabezpečuje, že pre zadané $M < N$ má odhad v Karhunen-Loẽvovej báze najmenšiu chybu.

Príklad 1.5.2 *V tabuľke* 1.9 *sú váhy piatich novorodencov a váhy týchto detí v deň ich prvých narodenín. Nech je náhodný vektor* $\mathbf{f}(\omega)$ *modelom pre dáta,*

dieťa	váha pri narodení	váha v 1.roku
D1	3	11
D2	2.96	9
D3	3.2	10
D4	3.6	12
D5	3.65	11.3

Tabuľka 1.9: Váha piatich detí pri narodení a v 1.roku

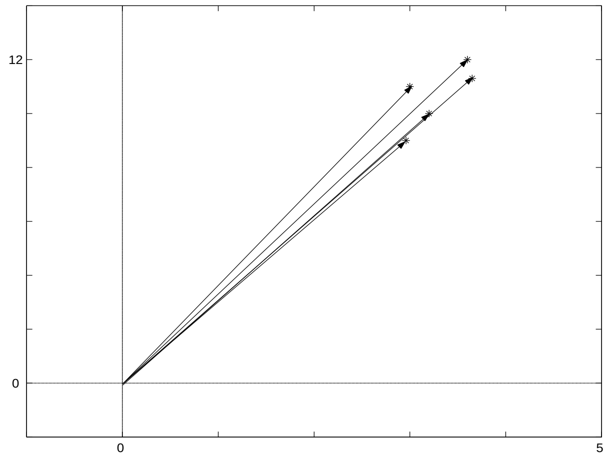
predstavujúce váhy detí pri narodení (prvá zložka vektora) a v deň prvých narodenín (druhá zložka vektora). Nájdite ortogonálnu bázu, v ktorej náhodný vektor $\mathbf{f}$ *bude mať nekorelované koeficienty* $c_0(\omega)$, $c_1(\omega)$.

Riešenie:

Jednotlivé realizácie náhodného procesu $\mathbf{f}(\omega)$ sú reprezentované maticou $\mathbf{M}$.

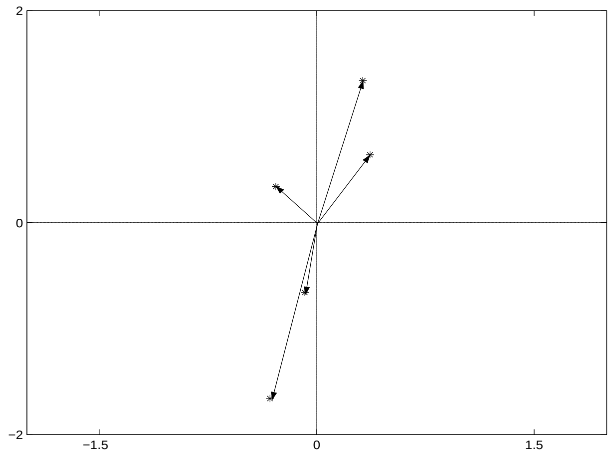
$$\mathbf{M} = \begin{pmatrix} 3 & 11 \\ 2.96 & 9 \\ 3.2 & 10 \\ 3.6 & 12 \\ 3.65 & 11.3 \end{pmatrix}$$

Vektory, reprezentujúce jednotlivé realizácie sú zakreslené na obrázku 1.11.



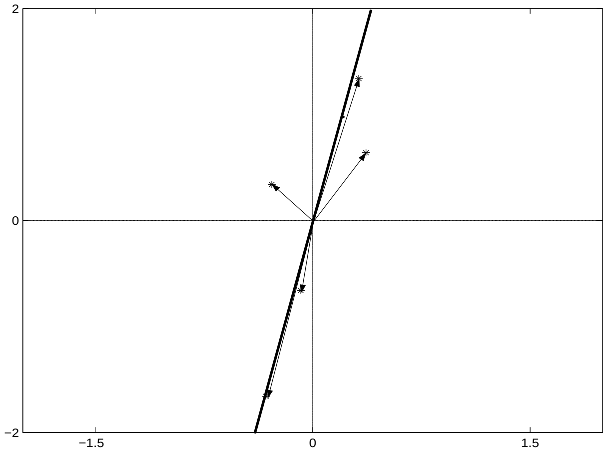
Obrázok 1.11: Vektory zodpovedajúce matici $\mathbf{M}$

Riadky centrovanej matice $\mathbf{M}^\circ$ sú zakreslené ako vektory na obrázku 1.12.



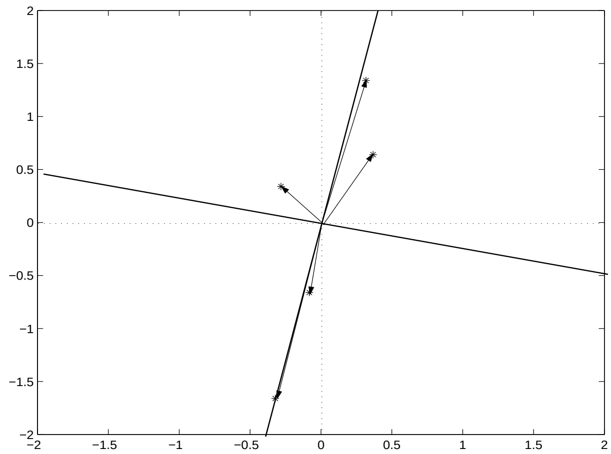
Obrázok 1.12: Vektory zodpovedajúce matici $\mathbf{M}^\circ$

Na obrázku 1.13 je zakreslená priamka predstavujúca bázový vektor v smere najväčšieho rozptylu náhodného vektora $\mathbf{f}(\omega)$.



Obrázok 1.13: Smer najväčšieho rozptylu náhodného vektora $\mathbf{f}(\omega)$

Na obrázku 1.14 je okrem prvej bázovej priamky nakreslená aj druhá bázová priamka, ktorá je na prvú kolmá. Náhodná premenná $c_1(\omega)$ predstavuje koeficient priemetu vektora $\mathbf{f}(\omega)$ na túto priamku. Koeficient $c_1(\omega)$ je nekorelovaný s koeficientom $c_0(\omega)$.



Obrázok 1.14: Dva kolmé smery, v ktorých má náhodný vektor $\mathbf{f}(\omega)$ nekorelované koeficienty

□