Redukce dimenze příznakového prostoru, selekce příznaků

Proč?

- měření každého příznaku má určitou cenu
- vyšší počet příznaků může zvýšit chybu klasifikace
  → nezvýší, pokud je známa $P(\mathbf{x}|\omega)$
  → klasický problém přesnost vs. generalizace

A. Optimální reprezentace v prostoru nižší dimenze

- Karhunen-Loewův rozvoj (KL) = Analýza hlavních komponent (Principal component analysis, PCA)

B. Minimalizace chyby klasifikace

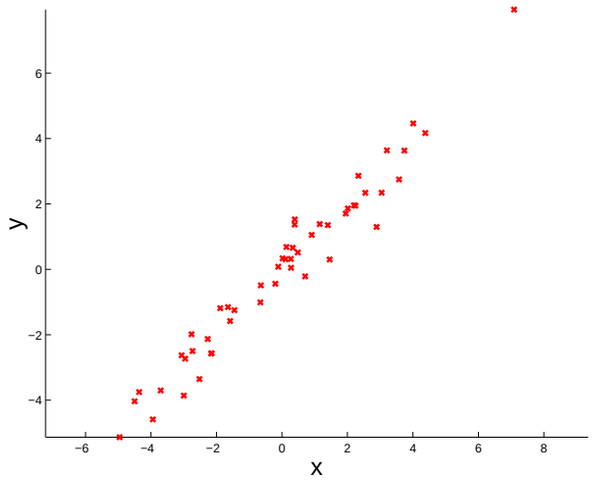
B1 Extrakce $\mathbf{X}^n \rightarrow \mathbf{X}^m$, $m < n$

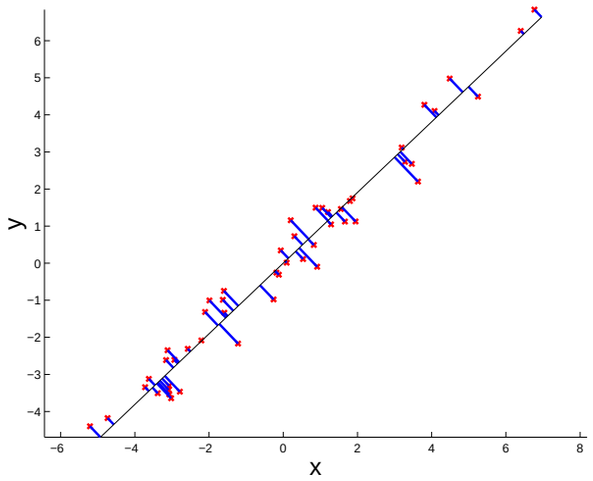
* neřeší problém ceny měřením (měřím vše)
* Fisherův lineární diskriminant (FLD)

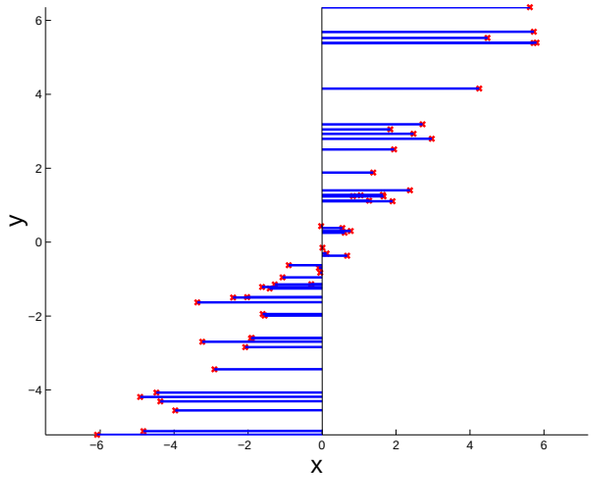
B2 Selekce: výběr podmnožiny příznaků

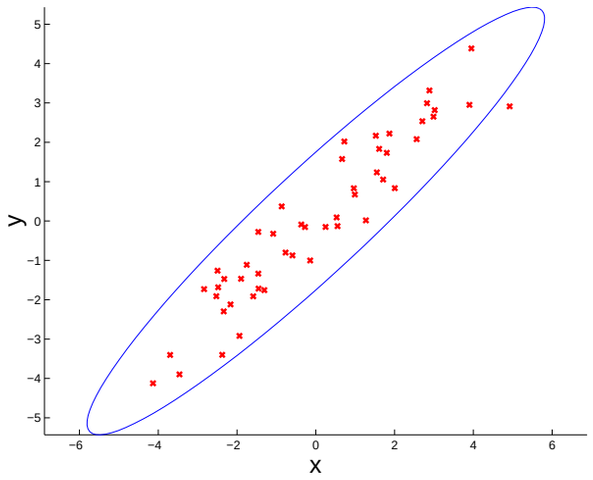
* obecně vyšší chyba klasifikace než B1
* Forward search
* Backward search
* Branch and bound

- máme: $\left\{\mathbf{X}_i \in \mathbb{R}^K, i = 1 \ldots n\right\}$
- hledáme: $\left\{\mathbf{Y}_j \in \mathbb{R}^K, j = 1 \ldots m, m < n\right\}$ tak, aby aproximace $\widetilde{\mathbf{X}}_i = \sum_{j=1}^{m} C_{ij}\mathbf{Y}_j$ byla co nejblíže původním datům









taková $\mathbf{Y}_j$ a C_{ij}, aby byla minimalizována kvadratická chyba aproximace:

$$g(\mathbf{Y}_j, C_{ij}) = \sum_{i=1}^{n} ||\mathbf{X}_i - \widetilde{\mathbf{X}}_i||^2 = \sum_{i=1}^{n} ||\mathbf{X}_i - \sum_{j=1}^{m} C_{ij}\mathbf{Y}_j||^2$$

$\rightarrow$ Minimalizuj g!

Pozorování: Bez újmy na obecnosti mohu hledat $\{\mathbf{Y}_j\}$ jako ortonormální bázi.

$$\begin{aligned} g &= \sum_{i=1}^{n} ||\mathbf{X}_i - \sum_{j=1}^{m} C_{ij}\mathbf{Y}_j||^2 \\ &= \sum_{i=1}^{n} \left(\mathbf{X}_i - \sum_{j=1}^{m} C_{ij}\mathbf{Y}_j\right)^T (\mathbf{X}_i - \sum_{k=1}^{m} C_{ik}\mathbf{Y}_k) \\ &= \sum_i \mathbf{X}_i^T\mathbf{X}_i - 2\sum_{i,j} C_{ij}\mathbf{Y}_j^T\mathbf{X}_i + \underbrace{\sum_{i,j,k} C_{ij}C_{ik}\mathbf{Y}_j^T\mathbf{Y}_k}_{\sum_{i,j} C_{ij}^2} \end{aligned}$$

Znám-li $\mathbf{Y}_j$, jaká zvolit C_{ij}?

$$0 = \frac{\partial g}{\partial C_{uv}} = -2(\mathbf{X}_u^T\mathbf{Y}_v - C_{uv}) \Rightarrow C_{uv} = \mathbf{X}_u^T\mathbf{Y}_v$$

Minimalizace g, nyní už jen jako funkce báze $\mathbf{Y}_j$:

$$g = \sum_i \mathbf{X}_i^T \mathbf{X}_i - \sum_{i,j} (\mathbf{X}_i^T \mathbf{Y}_j) = konst. - \sum_j \mathbf{Y}_j^T \underbrace{(\sum_i \mathbf{X}_i \mathbf{X}_i^T)}_{\mathbf{S}} \mathbf{Y}_j$$

Příklad. (→ průsvitka). 2D centrovaná data, chci 1D aproximaci ⇒ minimalizace $g(\mathbf{Y})$ za vedlejší podmínky $\mathbf{Y}^T\mathbf{T} = 1$:

Metoda Lagrang. multiplikátorů: $h(\mathbf{Y}, \lambda) = g(\mathbf{Y}) + \lambda(\mathbf{Y}^T\mathbf{Y} - 1)$

$$0 = \frac{\partial h}{\partial \lambda} = \mathbf{Y}^T\mathbf{Y} - 1$$

$$\mathbf{0} = \frac{\partial h}{\partial \mathbf{Y}} = -2\mathbf{S}\mathbf{Y} + 2\lambda\mathbf{Y}$$

1-D aproximace:

- $\mathbf{SY} = \lambda\mathbf{Y} \Rightarrow \mathbf{Y}$ je vlastní vektor $\mathbf{S}$.
- protože $g = konst. - \mathbf{YSY} = konst. - \lambda\mathbf{Y}^T\mathbf{Y} = konst. - \lambda$, je řešením vlastní vektor odpovídající λ_{max}.

m-D aproximace?

- řešením je m vlastních vektorů $\mathbf{S}$ odpovídajících m největším vlastním číslům.

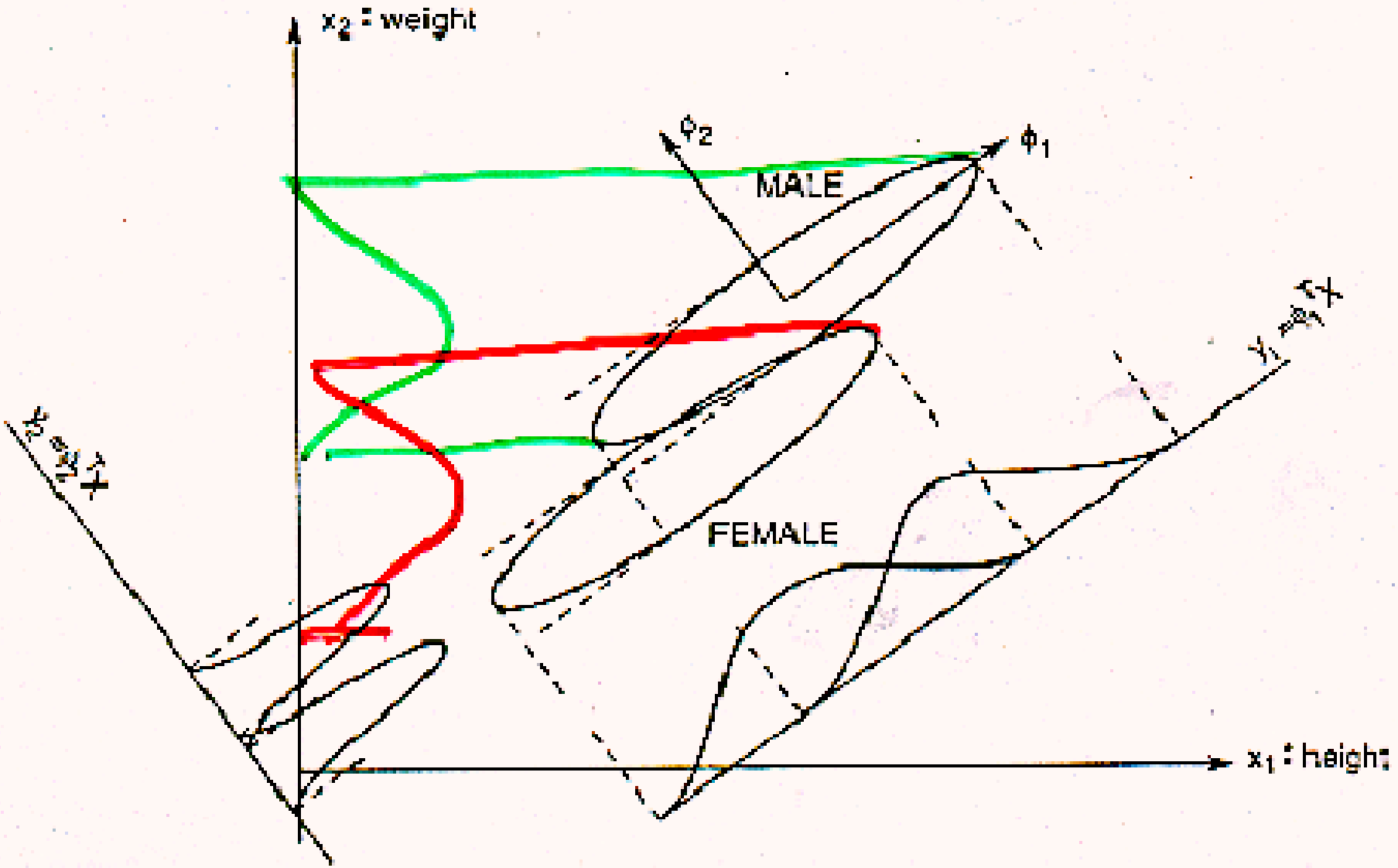


Fig. 10-1 An example of feature extraction for classification.

Nástin na jednoduchém příkladě → průsvitka.

2 třídy. Chci najít (1-D) směr, na který se vzorky promítnou a v rámci těchto projekcí budou 2 třídy co nejlépe separované.

Projekce: $y_i = \mathbf{w}^T \mathbf{X}_i; \qquad (\mathbf{w}^T \mathbf{w} = 1)$

Zvolím si míru "separovanosti" → rozumnou volbou je:

$J(\mathbf{w}) = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2}$, kde

$m_i = \frac{1}{n_i} \sum_{y \in \mathcal{E}_i} y$

($\mathcal{E}_i$ je množina promítnutých vzorků příslušných i-té třídě a n_i počet prvků v ní)

$s_i^2 = \sum_{y \in \mathcal{E}_i} (y - m_i)^2$

To si vyjádřím pomocí původních dat:

$$(m_1 - m_2)^2 = \mathbf{w}^T(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T\mathbf{w} = \mathbf{w}^T\mathbf{S}_B\mathbf{w}$$

$$(s_1^2 + s_2^2) = \mathbf{w}^T\mathbf{S}_W\mathbf{w}$$

kde

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2$$

$$\mathbf{S}_i = \sum_{x\in\mathcal{D}_i}(\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T \qquad (\mathbf{m}_i = \tfrac{1}{n_i}\sum_{x\in\mathcal{D}_i}\mathbf{x})$$

($\mathcal{D}_i$ je množina vzorků příslušných i-té třídě)

$$\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T$$

takže $J(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{S}_B\mathbf{w}}{\mathbf{w}^T\mathbf{S}_W\mathbf{w}} \Rightarrow$ najdi maximum $J(\mathbf{w})$!

$J(\mathbf{w})$ má tvar Rayleighova kvocientu a jeho extremalizace vede na zobecněný problém vlastních čísel:

$$\mathbf{S}_B\mathbf{w} = \lambda\mathbf{S}_W\mathbf{w}$$

$$\Rightarrow {\mathbf{S}_W}^{-1}\mathbf{S}_B\mathbf{w} = \lambda\mathbf{w}$$

Protože $\mathbf{S}_B\mathbf{x} \sim (\mathbf{m}_1 - \mathbf{m}_2)$ (je násobkem) pro kterékoli $\mathbf{x}$, řešení $\mathbf{w}$ musí být násobkem ${\mathbf{S}_W}^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$:

$$\mathbf{w} \sim {\mathbf{S}_W}^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

→ průsvitka

LINEÁRNÍ DISKRIMINANTY

Intuitivně je zřejmé, že pravděpodobnost chyby je nepřímo úměrná poměru

$$\frac{\text{ROZPTYL MEZI TŘÍDAMI}}{\text{ROZPTYL UVNITŘ TŘÍD}}$$

Dáno:

$\bar{x} \in \mathcal{X}^m$

$\omega_r, \; r = 1, 2, \ldots, R$

$\{P(\omega_r), \; P(x/\omega_r)\}$

ROZPTYL MEZI TŘÍDAMI:

$$B(x) = \sum_{r=1}^{R-1} \sum_{s=r+1}^{R} P(\omega_r) P(\omega_s) \mu_{rs} \mu_{rs}^T =$$

$$= \sum_{r=1}^{R} P(\omega_r) (\mu_r - \mu_0)(\mu_r - \mu_0)^T$$

kde

$$\mu_r = \int x \, p(x/\omega_r) \, dx$$

$$\mu_{rs} = \mu_r - \mu_s$$

$$\mu_0 = \sum_{r=1}^{R} P(\omega_r) \mu_r = \int_{\mathcal{X}^m} x p(x) \, dx$$

ROZPTYL UVNITŘ TŘÍD

$$x^T(x) = \sum_{r=1}^{R} P(\omega_r) \int_{\mathcal{X}^m} (x - \mu_r)^T (x - \mu_r) p(x/\omega_r) \, dx$$

POMĚR ROZPTYLŮ

$$\rho(x) = tr\left((\mathcal{X}(x))^{-1} B(x)\right)$$

$$= \sum_{r=1}^{R} P(\omega_r)(\mu_r - \mu_0)(\mathcal{X}(x))^{-1}(\mu_r - \mu_0)$$

hledáme $\bar{y} = A\bar{x}$, tak aby $\rho(x)$ bylo maximální!

↓ (str. 156, Adaptivní a Učící se systémy →

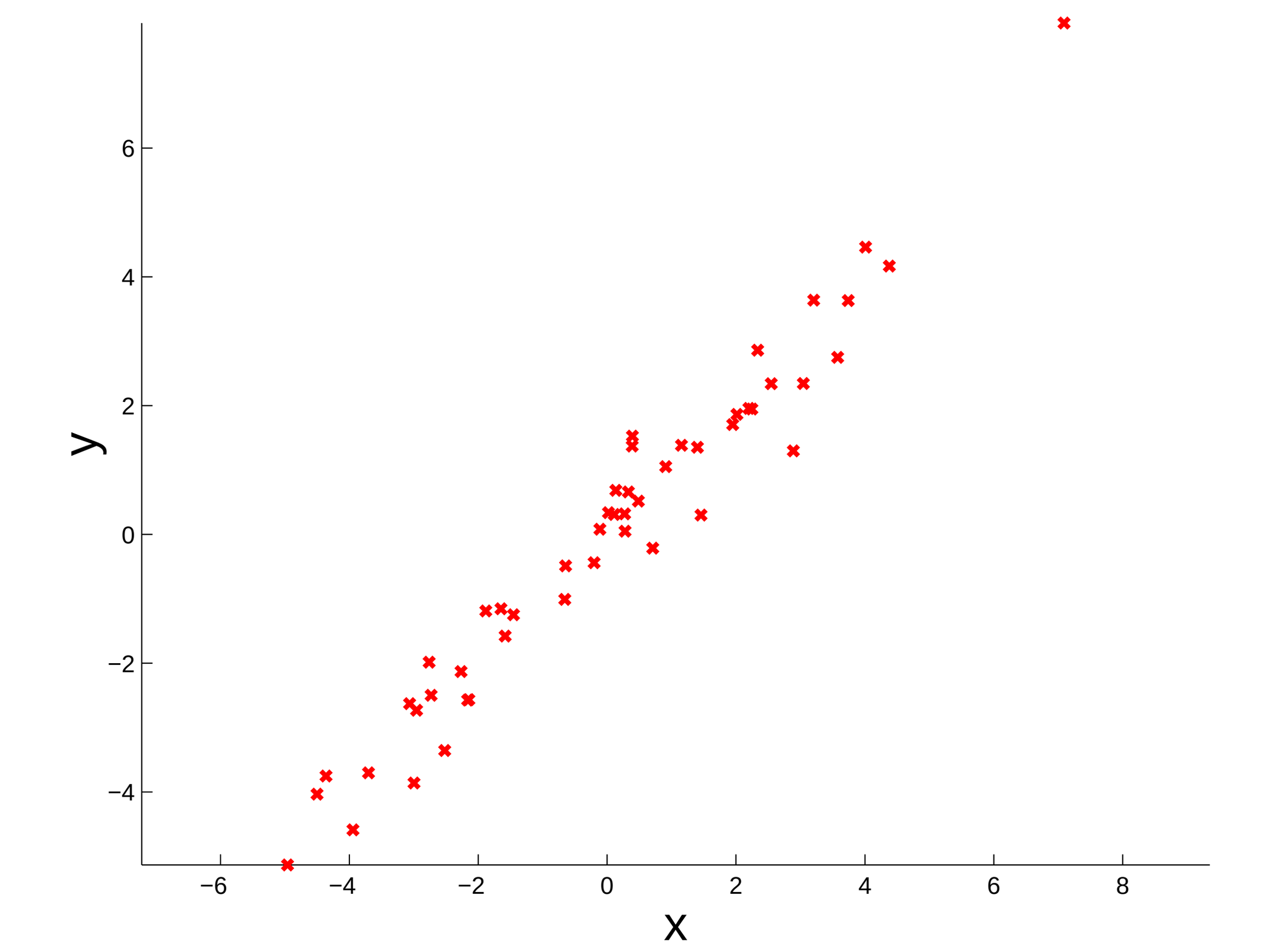
[A sestavíme z charakteristických vektorů $(\mathcal{X}^{-1}(x))^{-1} \cdot B(x)$]

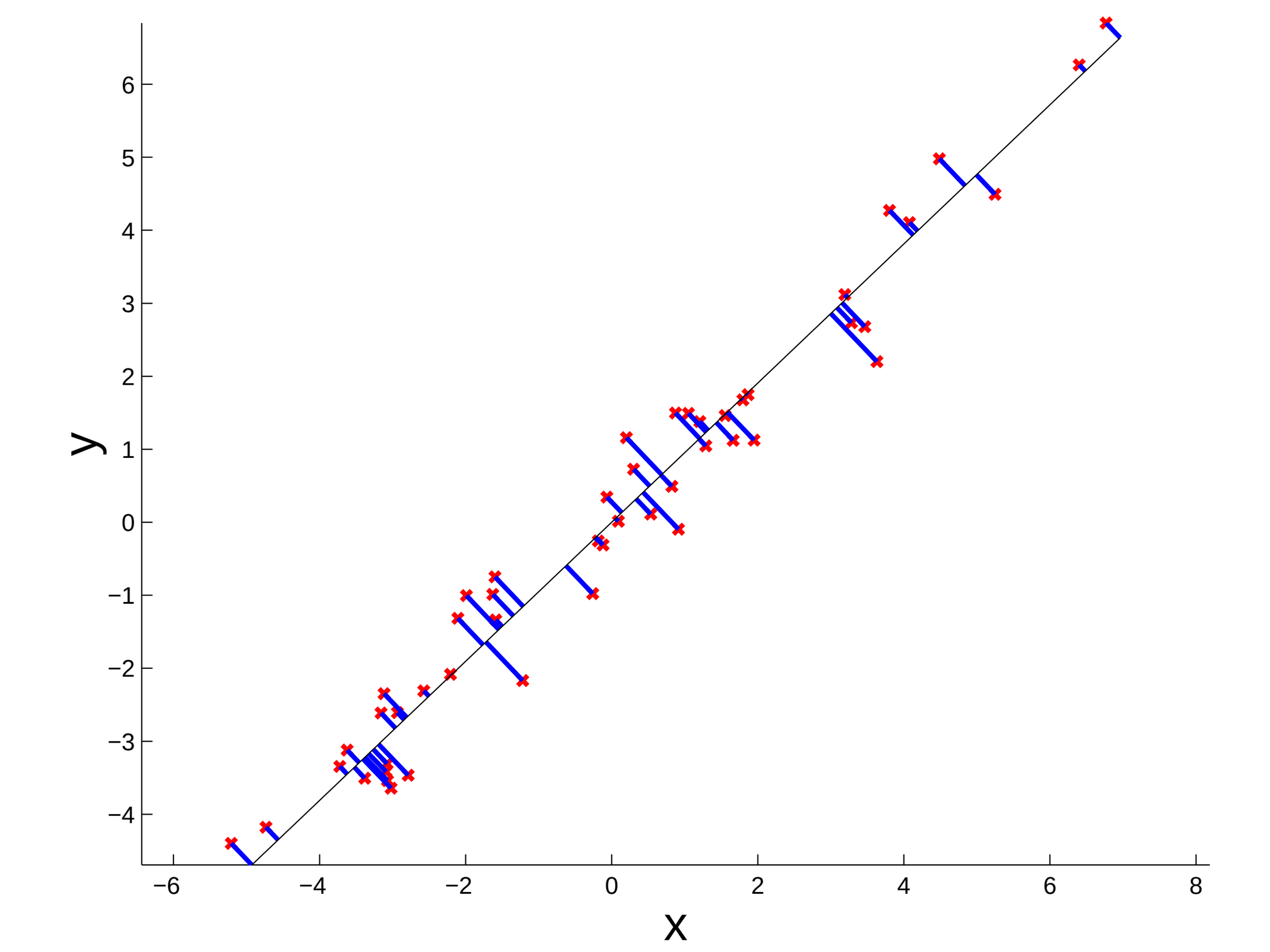
Při aplikaci:

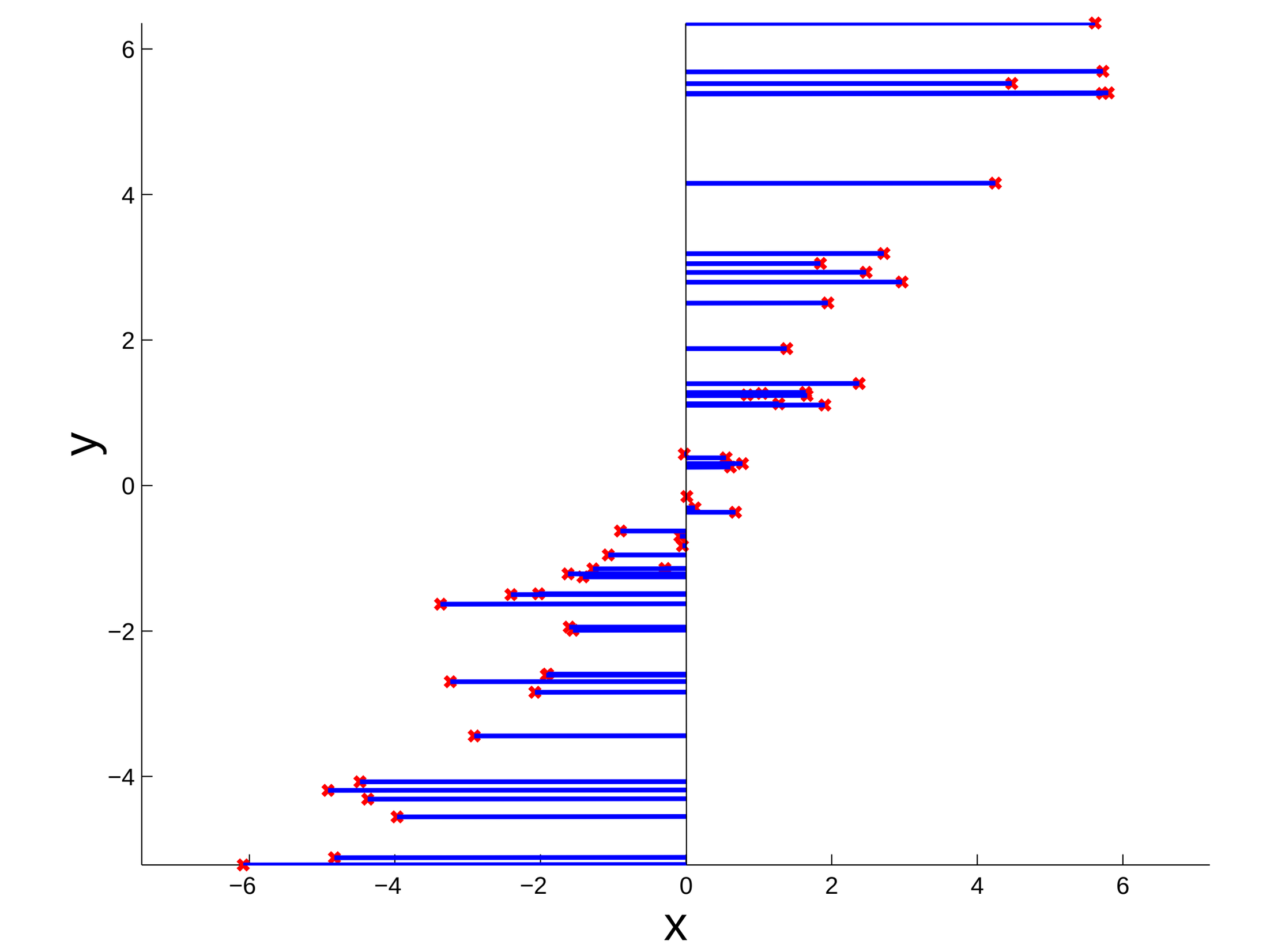
1. použijeme VÝBĚROVOU DISPERZNÍ MATICI, VÝBĚROVÉ STŘEDNÍ HODNOTY

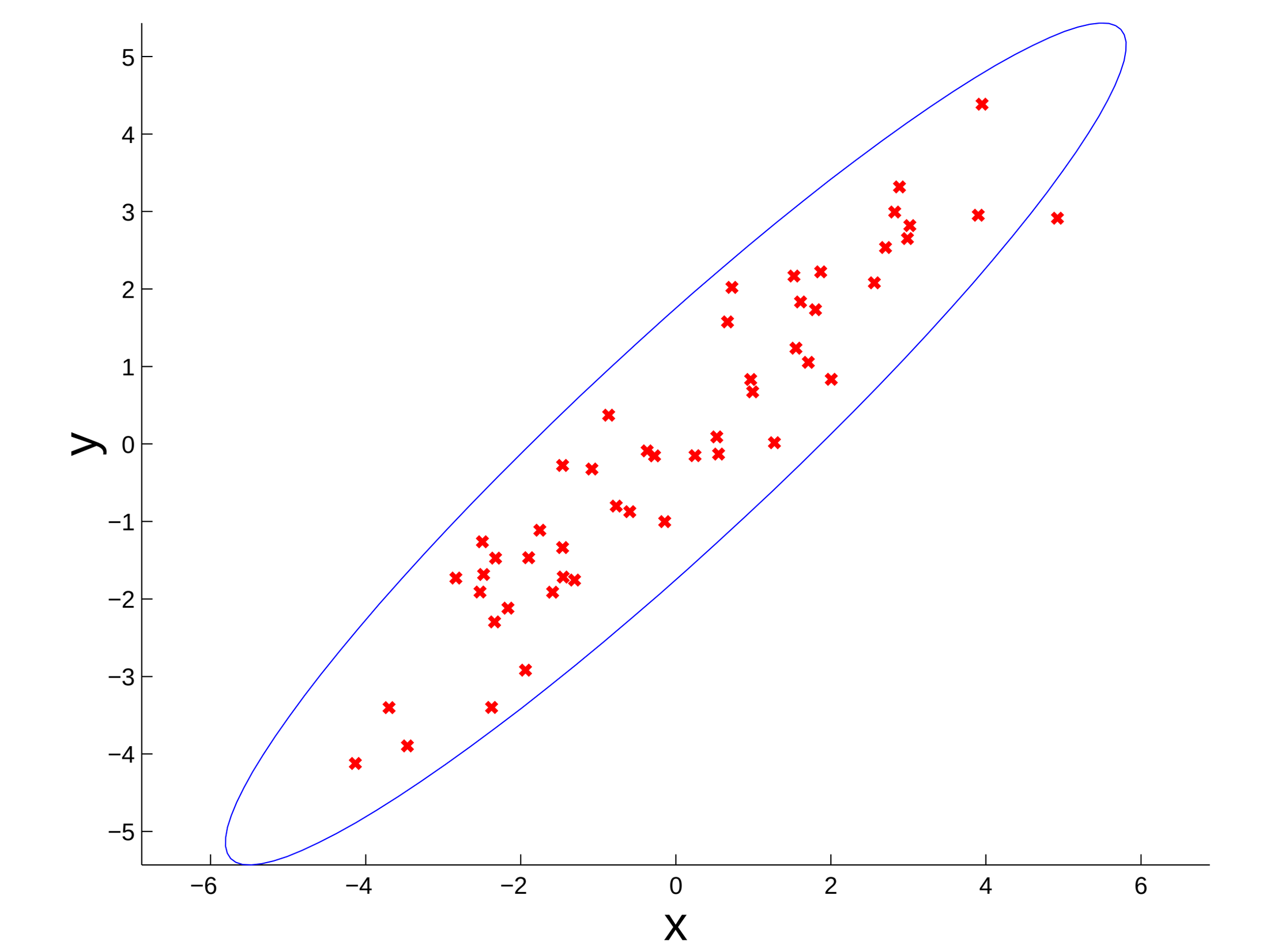
2. ČASTO NEJPRVE POUŽIJEME KAR.–LOEVE A PAK TEPRVE FLD!

3. POZOR NA SINGULARITU!


6
4
2
0
−2
−4
y
−6
−4
−2
0
2
4
6
8
X


y
x
6
5
4
3
2
1
0
−1
−2
−3
−4
−6
−4
−2
0
2
4
6
8


6
4
2
0
−2
−4
y
−6
−4
−2
0
2
4
6
X


5
4
3
2
1
0
−1
−2
−3
−4
−5
y
−6
−4
−2
0
2
4
6
X