VPLYV ŠUMU NA KLASIFIKÁCIU DO DVOCH TRIED

ROBUST 2010

KATARÍNA CIMERMANOVÁ

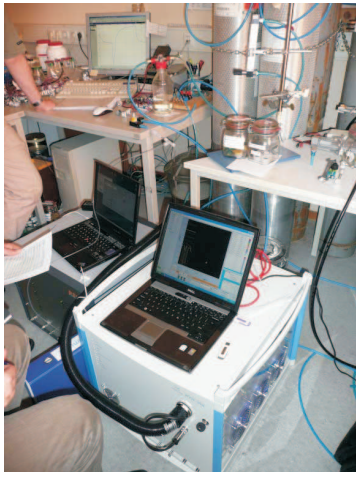
Oddelenie teoretických metód
Ústav merania
Slovenská akadémia vied

5. január 2010

Analýza dychu

- diagnostická metóda
- skorá detekcia niektorých chorôb napr.
  rakovina pľúc

BAMOD - projekt 6.RP EU



- analýza vydychovaných plynov na molekulovo orientovanú detekciu pľúcnych chorôb
  v rannom štádiu
  využitím rôznych analytických techník

Namerané dáta

- nízke koncentrácie vydychovaných plynov
  - jednotka: počet molekúl na miliardu molekúl [ppb]
- nameraný profil vydychovaného vzduchu
  - vyše 200 prchavých organických zložiek

ŠTATISTICKÁ ANALÝZA KONCENTRÁCIÍ VYDYCHOVANÝCH PLYNOV

ŠTATISTICKÁ ANALÝZA KONCENTRÁCIÍ VYDYCHOVANÝCH PLYNOV

KLASIFIKÁCIA DO DVOCH TRIED

ŠTATISTICKÁ ANALÝZA KONCENTRÁCIÍ VYDYCHOVANÝCH PLYNOV

KLASIFIKÁCIA DO DVOCH TRIED

$\{\mathbf{x}_i, y_i\}$

$\mathbf{x}_i \in R^N$
$y_i = \{+1, -1\}$
$\mathbf{x}_i \in \omega_1 \Rightarrow y_i = +1$
$\mathbf{x}_i \in \omega_2 \Rightarrow y_i = -1$

$g(\mathbf{x})$

$i = 1, \ldots, n$
N - počet vybraných prchavých organických zložiek
n - počet pozorovaných subjektov

ŠTATISTICKÁ ANALÝZA KONCENTRÁCIÍ VYDYCHOVANÝCH PLYNOV

KLASIFIKÁCIA DO DVOCH TRIED

$\{\mathbf{x}_i, y_i\}$

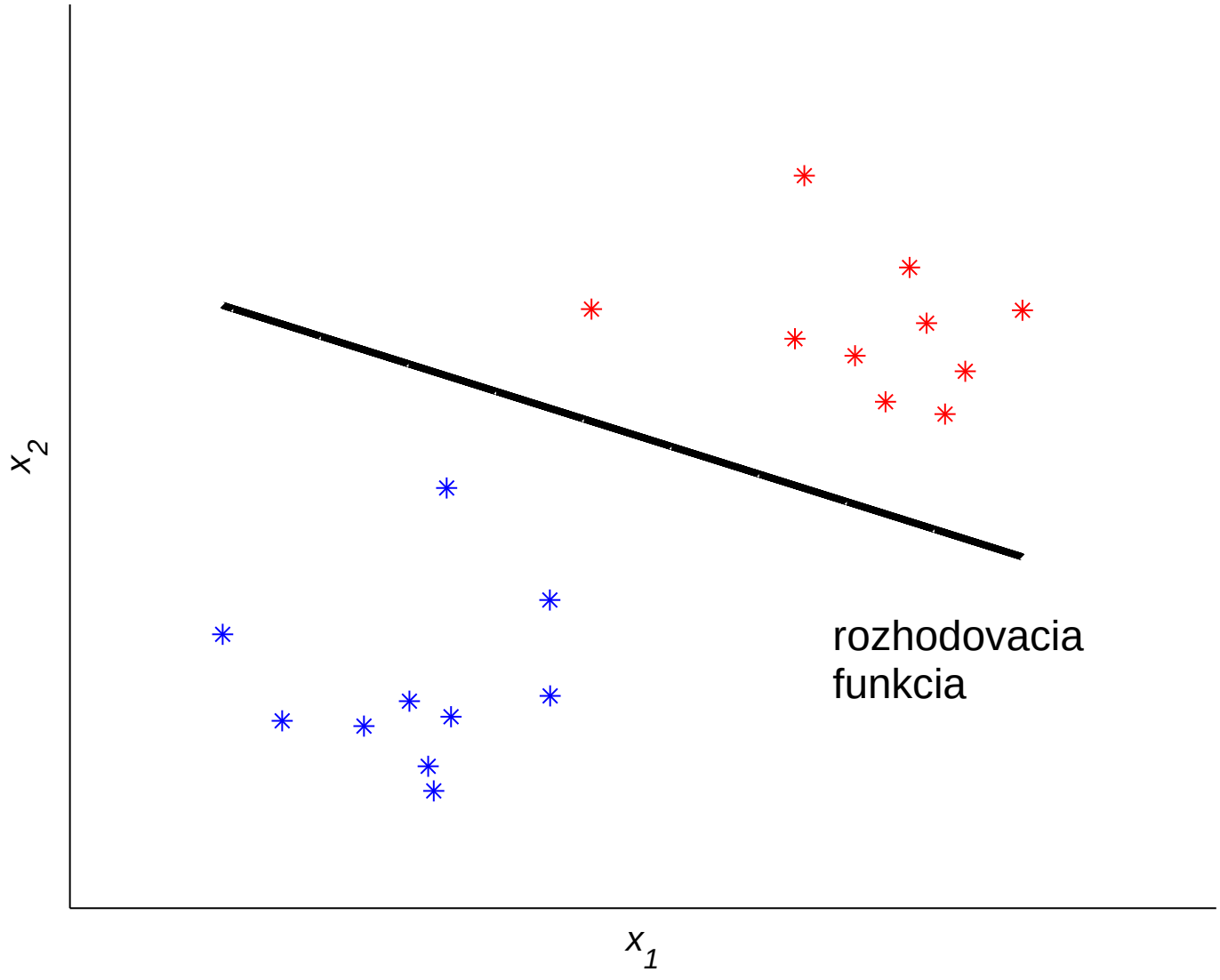
$\mathbf{x}_i \in R^N$
$y_i = \{+1, -1\}$
$\mathbf{x}_i \in \omega_1 \Rightarrow y_i = +1$
$\mathbf{x}_i \in \omega_2 \Rightarrow y_i = -1$

$\mathbf{x}^* \rightarrow g(\mathbf{x}^*) \rightarrow \hat{y}$

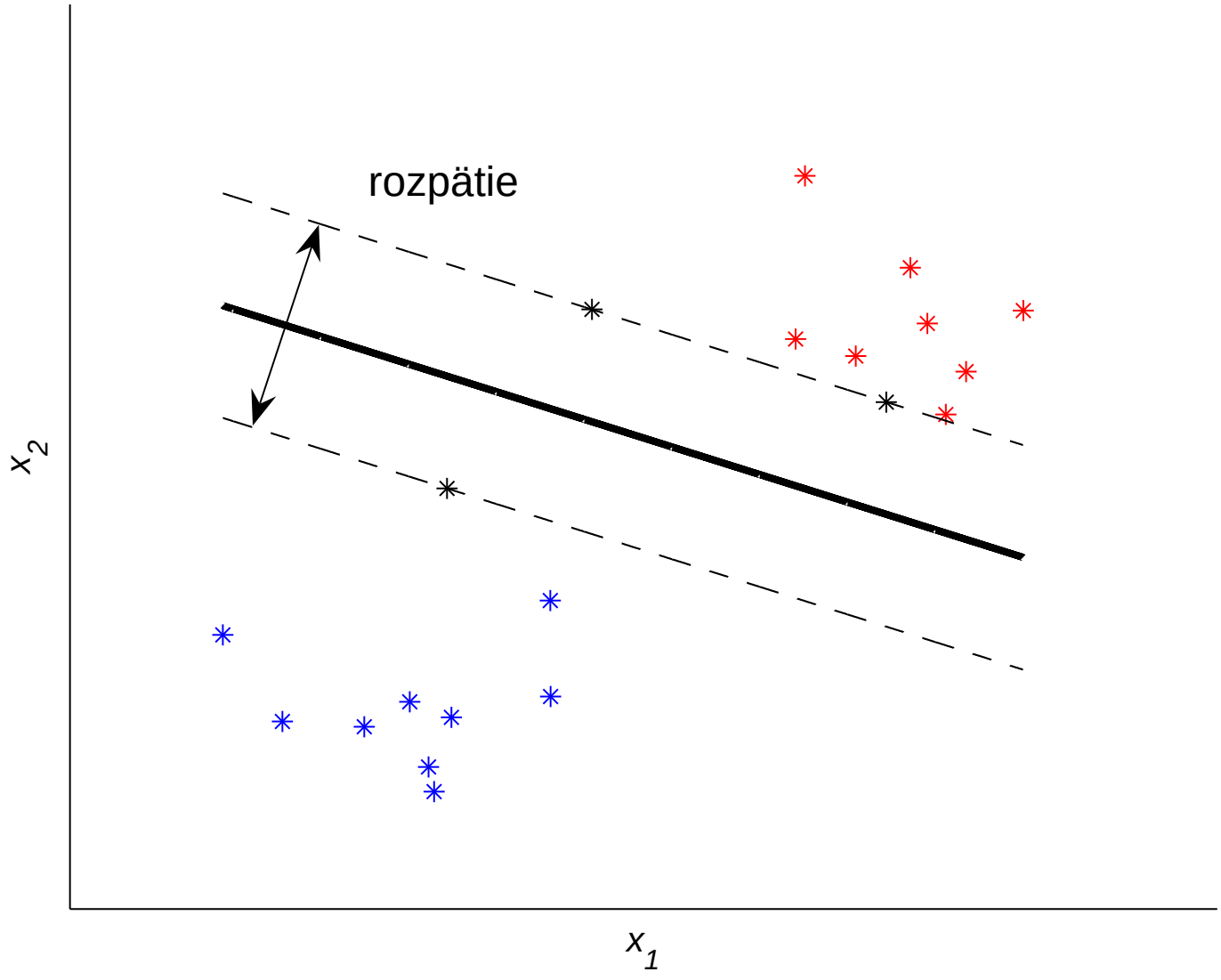
$i = 1, \ldots, n$
N - počet vybraných prchavých organických zložiek
n - počet pozorovaných subjektov

METÓDA OPORNÝCH BODOV, SVM

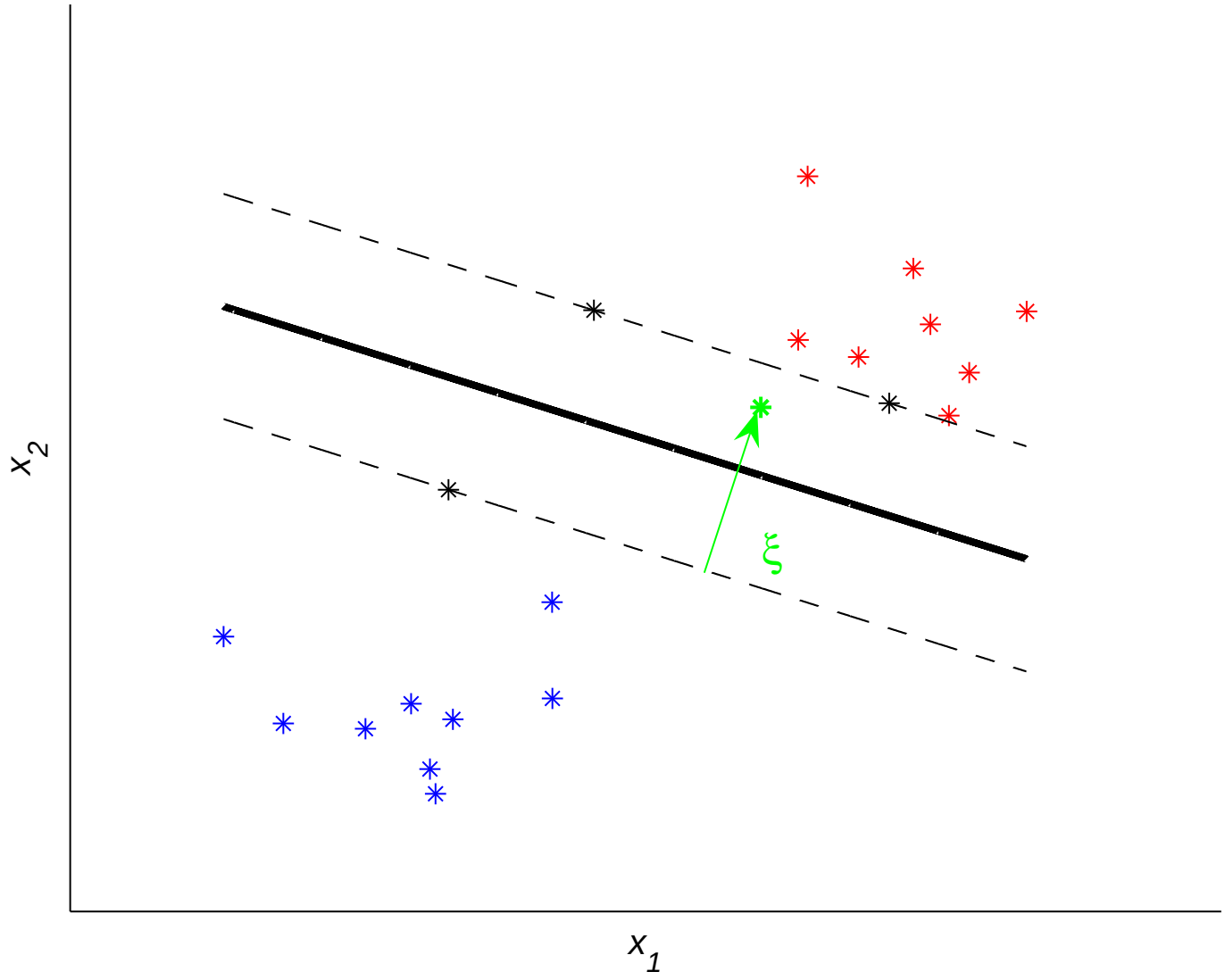
METÓDA OPORNÝCH BODOV, SVM



METÓDA OPORNÝCH BODOV, SVM



METÓDA OPORNÝCH BODOV, SVM



METÓDA OPORNÝCH BODOV, SVM

$g(\mathbf{x}) = \text{sign}(h(\mathbf{x}))$
$h(\mathbf{x}) = (\langle \mathbf{w}, \mathbf{x} \rangle + \mathbf{b})$

$g(\mathbf{x})$ - rozhodovacia funkcia
$\mathbf{w}, \mathbf{b}$ - parametre rozhodovacej funkcie, $|h(\mathbf{x})| \geq 1$

METÓDA OPORNÝCH BODOV, SVM

$g(\mathbf{x}) = \text{sign}(h(\mathbf{x}))$
$h(\mathbf{x}) = (\langle \mathbf{w}, \mathbf{x} \rangle + \mathbf{b})$

$g(\mathbf{x})$ - rozhodovacia funkcia
$\mathbf{w}, \mathbf{b}$ - parametre rozhodovacej funkcie, $|h(\mathbf{x})| \geq 1$

$h(\mathbf{x}^*) \geq 0 \rightarrow \hat{y} = +1$
$h(\mathbf{x}^*) < 0 \rightarrow \hat{y} = -1$

METÓDA OPORNÝCH BODOV, SVM

$$\min_{\mathbf{w},\mathbf{b},\xi} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$

s podm. $$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + \mathbf{b}) \geq 1 - \xi_i$$

$$\xi_i \geq 0,$$

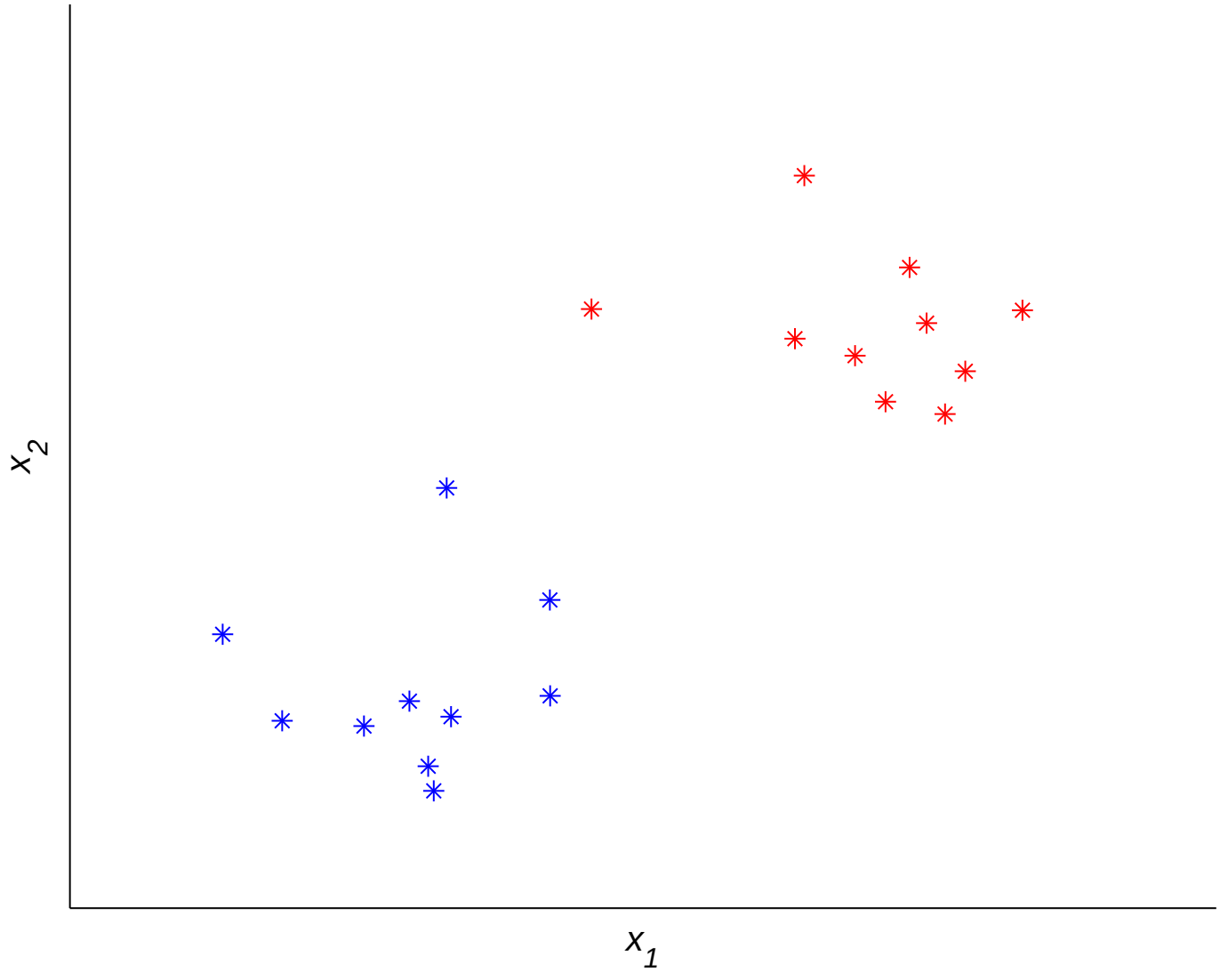
$\mathbf{x}_i$ - pozorované dáta, $\mathbf{x}_i \in R^N$
$\mathbf{w}, \mathbf{b}$ - parametre rozhodovacej funkcie
$2/||\mathbf{w}||$ - rozpätie medzi paralelnými hyperrovinami
C - regularizačná konštanta, $0 < C \leq \infty$
ξ_i - parameter straty, $\xi_i \geq 0$
$i = 1, \ldots, n$
n - počet pozorovaných subjektov
N - počet vybraných prchavých organických zložiek

ELIPSOIDÁLNY MODEL ŠUMU

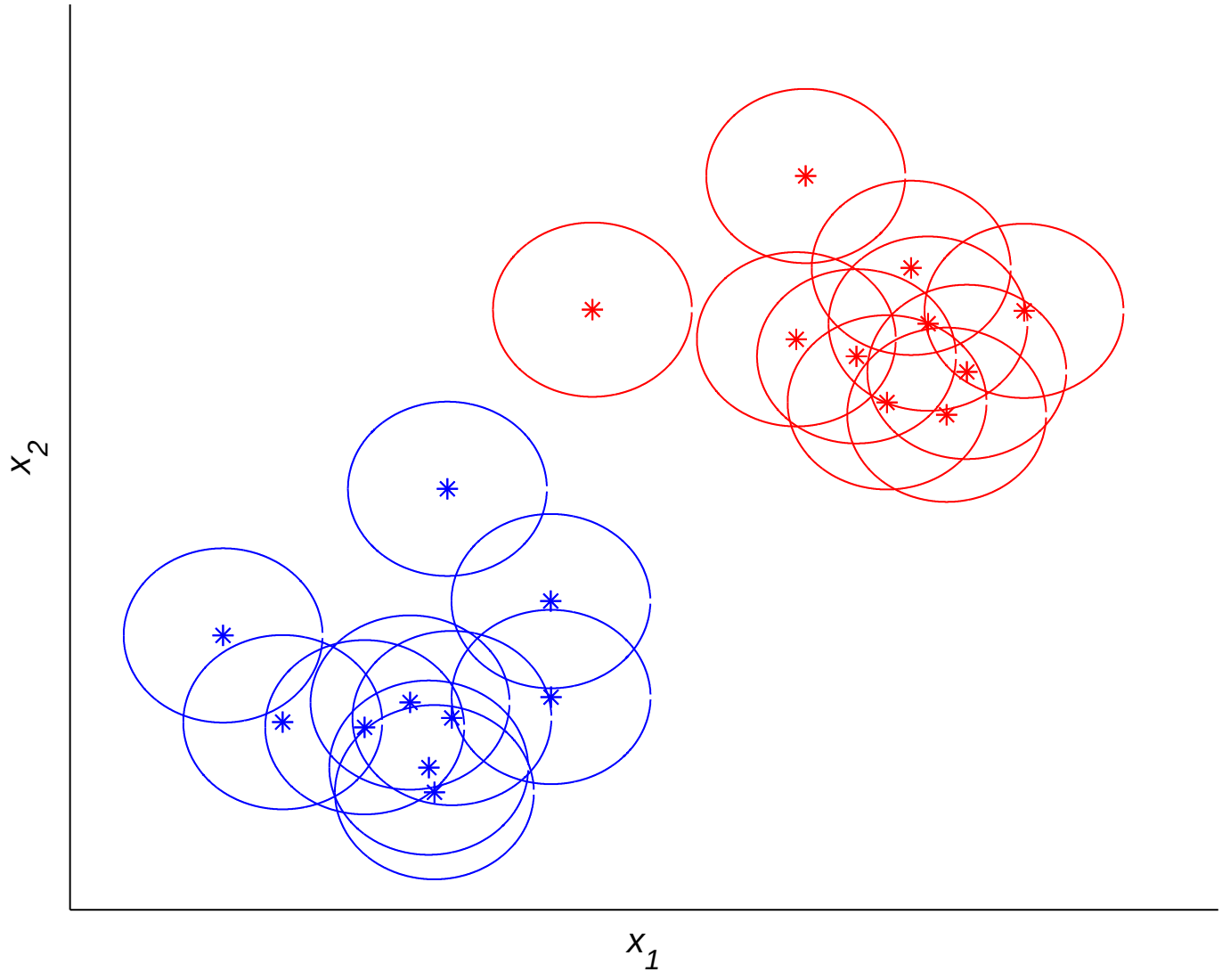
$$B(\overline{\mathbf{x}}, \mathbf{\Sigma}, \gamma) = \{\mathbf{x}_i : (\mathbf{x}_i - \overline{\mathbf{x}})' \mathbf{\Sigma}^{-1} (\mathbf{x}_i - \overline{\mathbf{x}}) \leq \gamma^2\}$$

$\mathbf{x}_i$ - pozorované dáta, $\mathbf{x}_i \in R^N$
$\overline{\mathbf{x}}$ - stred elipsoidu, $\overline{\mathbf{x}} \equiv \mathbf{x}_i$
$\mathbf{\Sigma}$ - kovariančná matica, tvar šumu, $\mathbf{\Sigma} \in R^{NXN}$
γ - hladina zašumenia, $\gamma \geq 0$

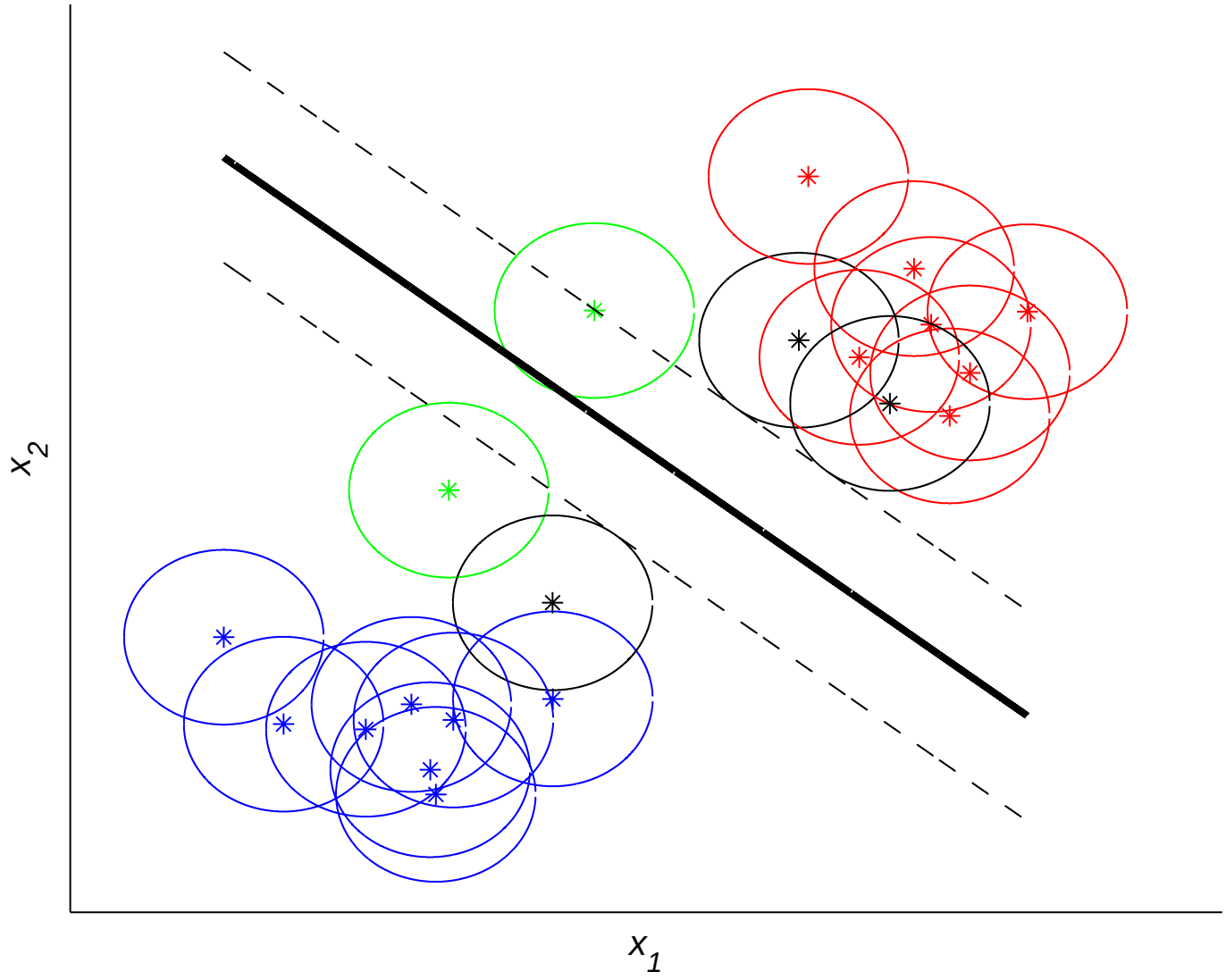
ZAŠUMENÉ DÁTA



ZAŠUMENÉ DÁTA



ZAŠUMENÉ DÁTA



ZAŠUMENÉ DÁTA

$$\min_{\mathbf{w},\mathbf{b},\xi} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$

s podm. $$y_i(\langle \mathbf{w}, \overline{\mathbf{x}}_i \rangle + \mathbf{b}) \geq 1 - \xi_i + \gamma\|\mathbf{\Sigma}_i^{\frac{1}{2}}\mathbf{w}\|$$

$$\xi_i \geq 0,$$

$\overline{\mathbf{x}}_i$ - pozorované dáta, $\overline{\mathbf{x}}_i \in R^N$
$\mathbf{w}, \mathbf{b}$ - parametre rozhodovacej funkcie
C - regularizačná konštanta, $0 < C \leq \infty$
ξ_i - parameter straty, $\xi_i \geq 0$
γ - hladina zašumenia, $\gamma \geq 0$
$\mathbf{\Sigma}$ - kovariančná matica, tvar šumu, $\mathbf{\Sigma} \in R^{NXN}$
$i = 1, \ldots, n$
n - počet pozorovaných subjektov
N - počet vybraných prchavých organických zložiek

ZAŠUMENÉ DÁTA

$$\min_{\mathbf{w},\mathbf{b},\xi} \quad \sum_{i=1}^{n} \xi_i$$

s podm. $$y_i(\langle \mathbf{w}, \overline{\mathbf{x}}_i \rangle + \mathbf{b}) \geq 1 - \xi_i + \gamma \|\mathbf{\Sigma}_i^{\frac{1}{2}} \mathbf{w}\|$$

$$||\mathbf{w}|| \leq W$$

$$\xi_i \geq 0,$$

$\overline{\mathbf{x}}_i$ - pozorované dáta, $\overline{\mathbf{x}}_i \in R^N$
$\mathbf{w}, \mathbf{b}$ - parametre rozhodovacej funkcie
W - regularizačná konštanta, $0 < W \leq \infty$
ξ_i - parameter straty, $\xi_i \geq 0$
γ - hladina zašumenia, $\gamma \geq 0$
$\mathbf{\Sigma}$ - kovariančná matica, tvar šumu, $\mathbf{\Sigma} \in R^{N \mathrm{X} N}$
$i = 1, \ldots, n$
n - počet pozorovaných subjektov
N - počet vybraných prchavých organických zložiek

ODHAD TVARU ŠUMU

ODHAD TVARU ŠUMU

1. $$\Sigma^{11} = \Sigma^{jj} = \Sigma^{NN} = \min_j\{\max_i(x_{ij}) - \min_i(x_{ij})\}$$

$i = 1, \ldots, n, j = 1, \ldots, N$
n - počet pozorovaných subjektov
N - počet vybraných prchavých organických zložiek

ODHAD TVARU ŠUMU

1. $$\Sigma^{11} = \Sigma^{jj} = \Sigma^{NN} = \min_j\{\max_i(x_{ij}) - \min_i(x_{ij})\}$$

2. $$\Sigma^{jj} = a_j = \{\max_i x_{ij} - \min_i x_{ij}\}$$

$i = 1, \ldots, n, j = 1, \ldots, N$
n - počet pozorovaných subjektov
N - počet vybraných prchavých organických zložiek

ODHAD TVARU ŠUMU

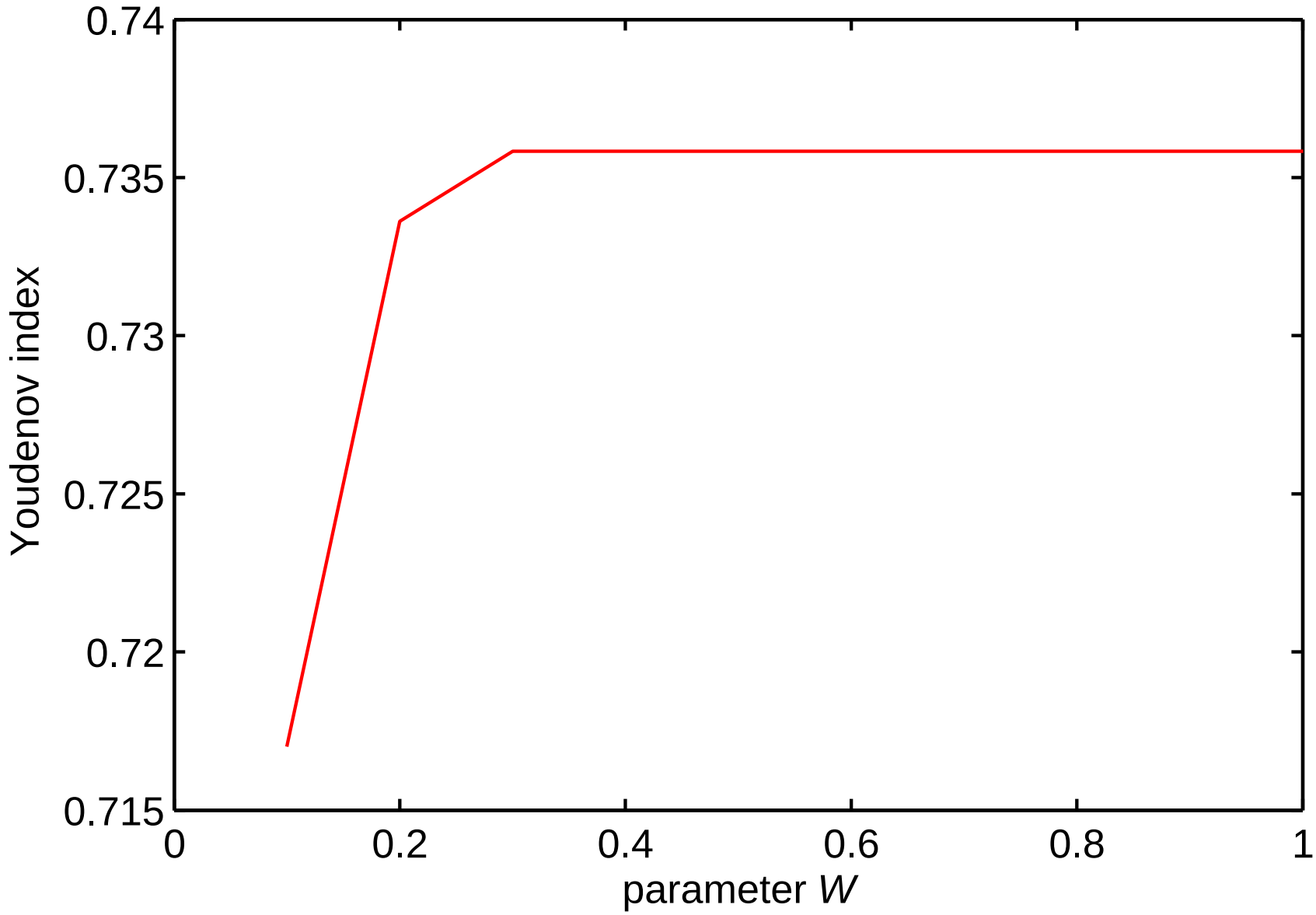
1. $$\Sigma^{11} = \Sigma^{jj} = \Sigma^{NN} = \min_j\{\max_i(x_{ij}) - \min_i(x_{ij})\}$$

2. $$\Sigma^{jj} = a_j = \{\max_i x_{ij} - \min_i x_{ij}\}$$

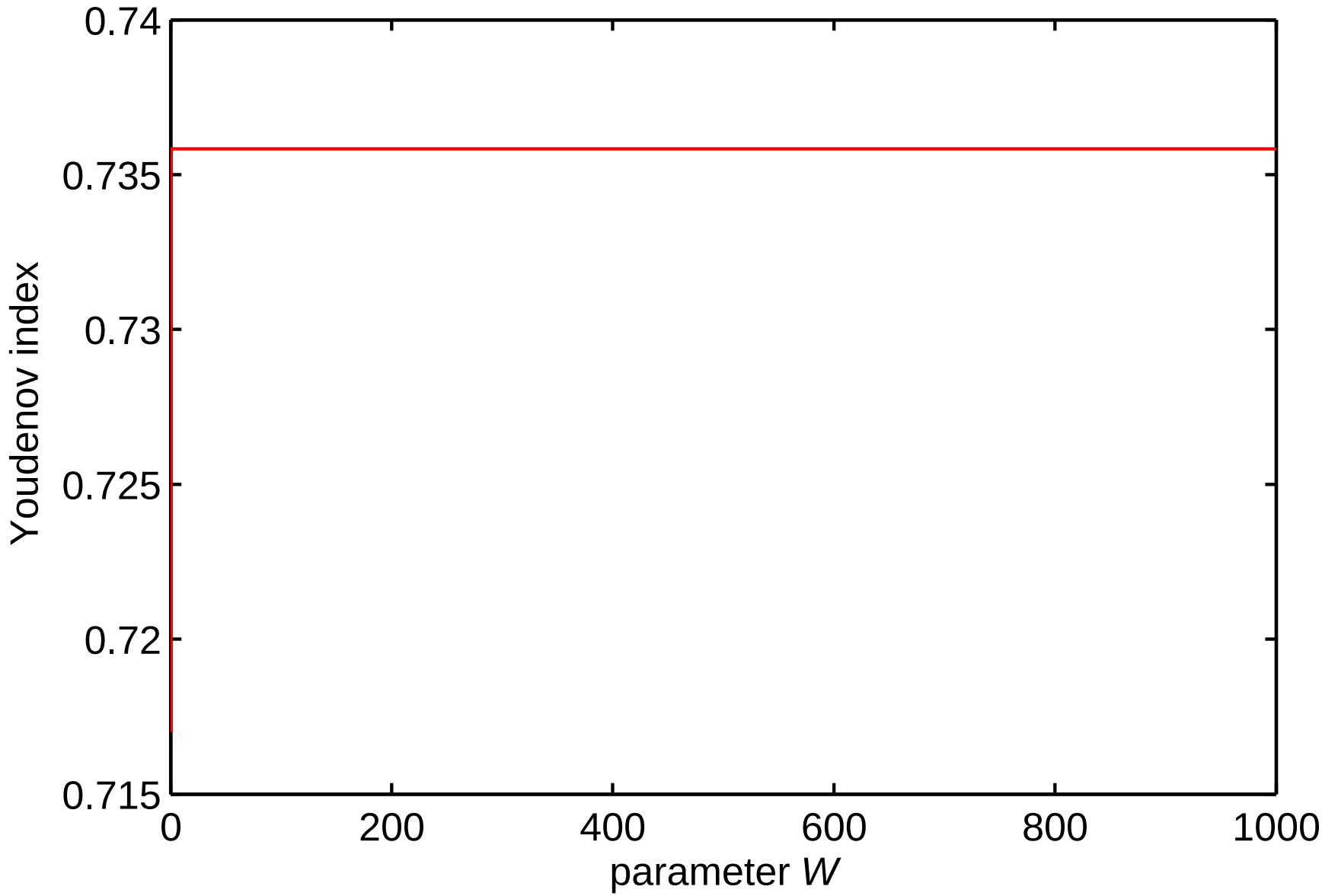
3. $$\Sigma^{jj} = \sigma_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij})^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_{ij}\right)^2$$

$i = 1, \ldots, n, j = 1, \ldots, N$
n - počet pozorovaných subjektov
N - počet vybraných prchavých organických zložiek

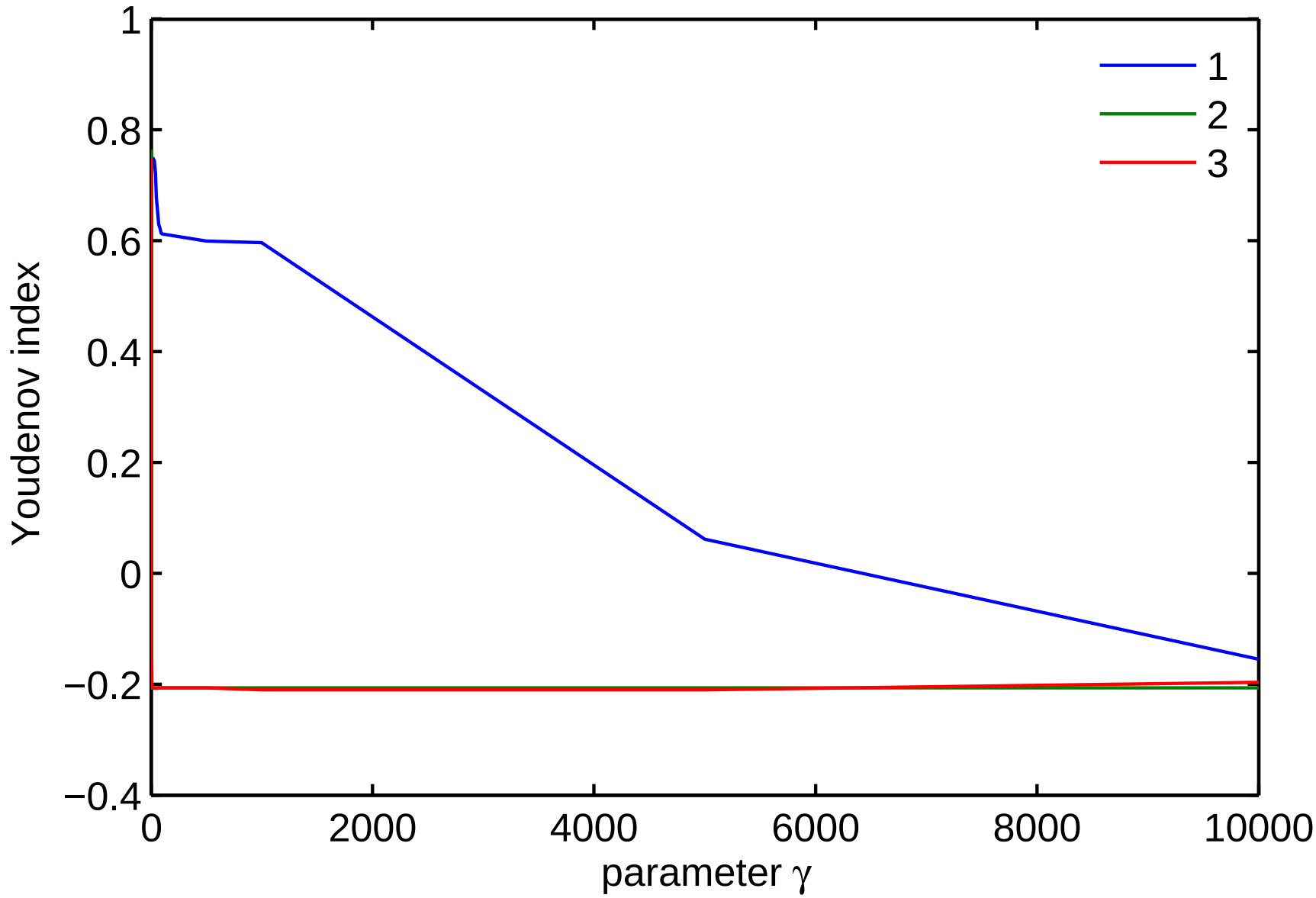
VÝSLEDKY PRE $N = 210$



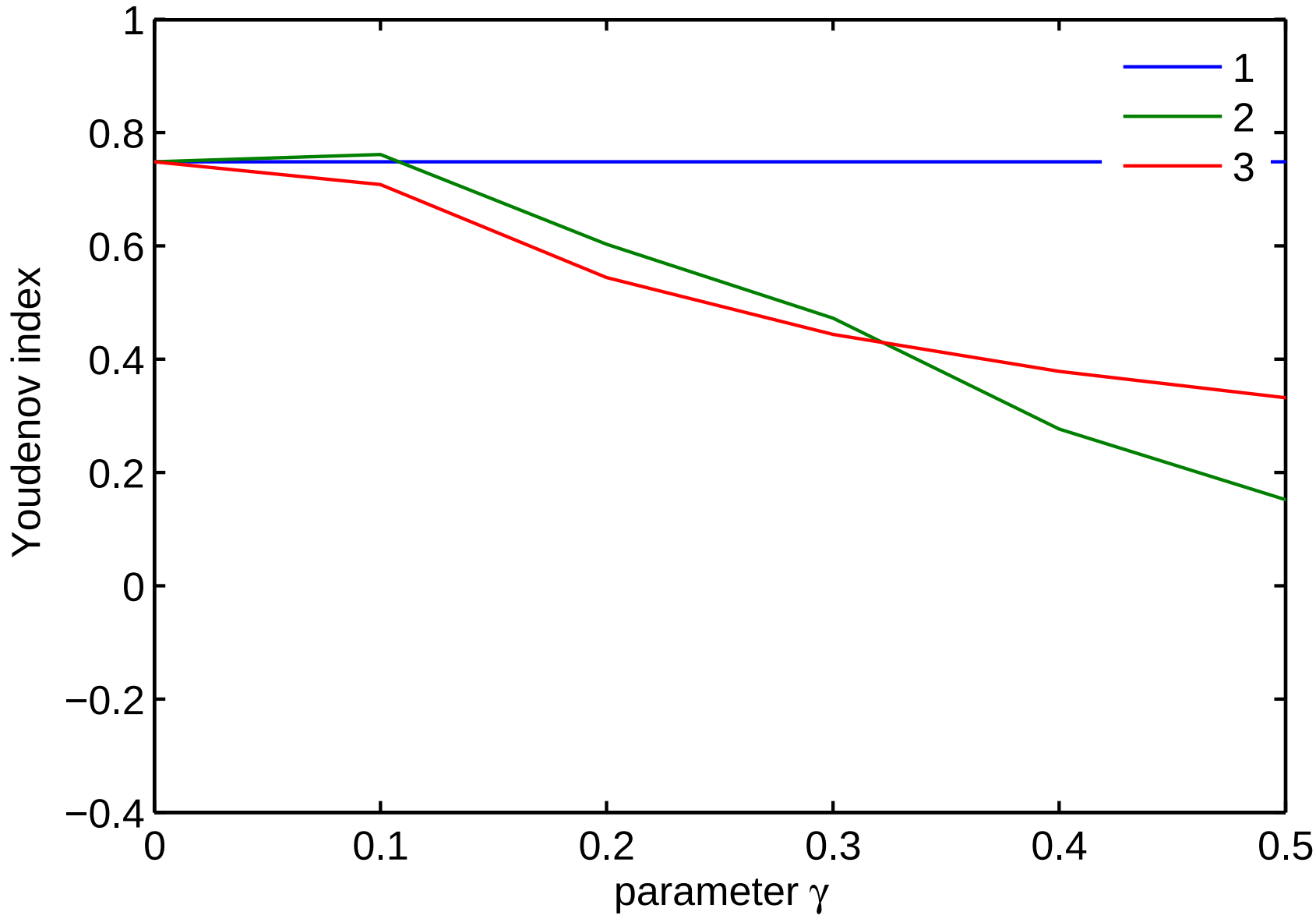
VÝSLEDKY PRE $N = 210$



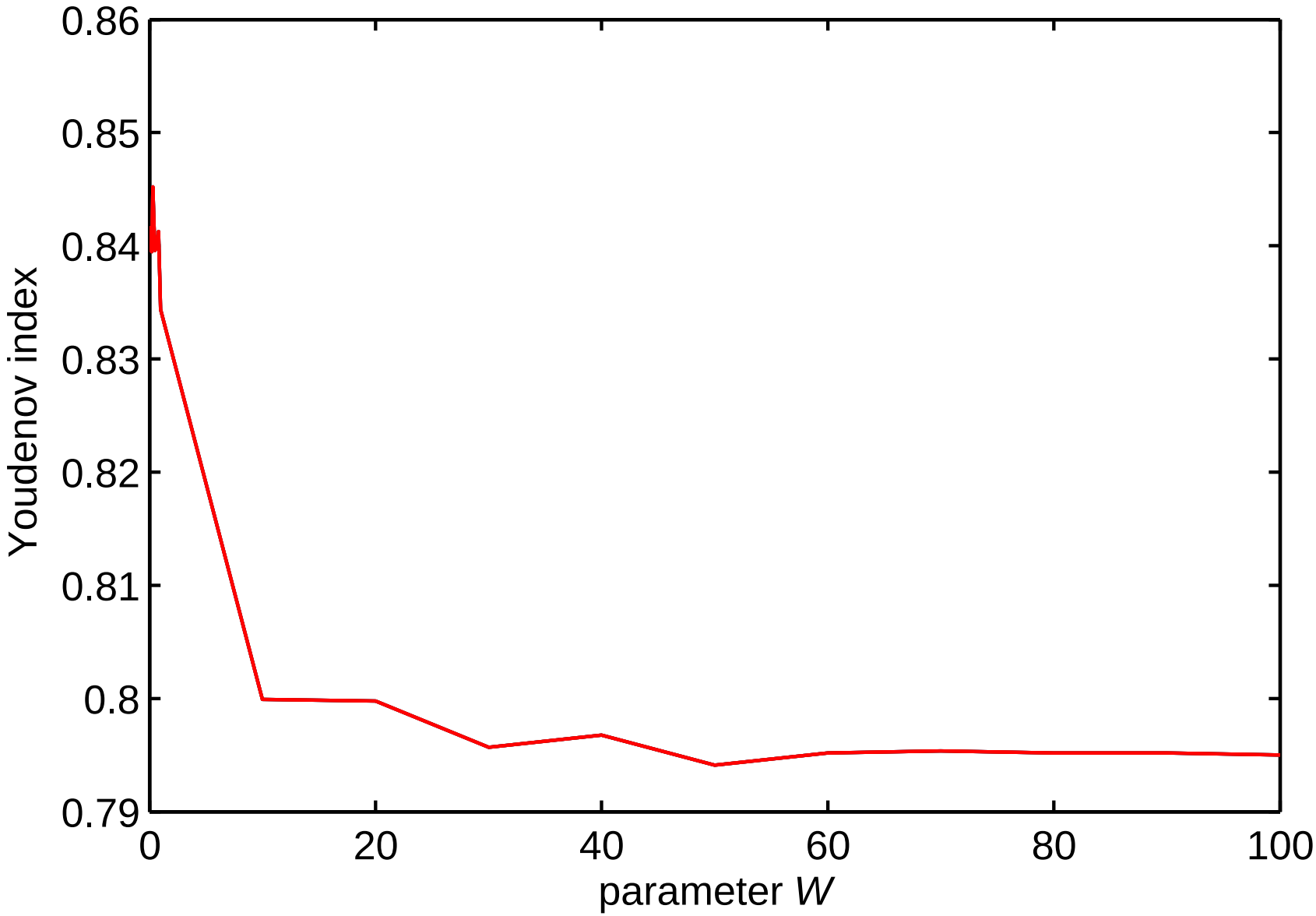
VÝSLEDKY PRE $N = 210$



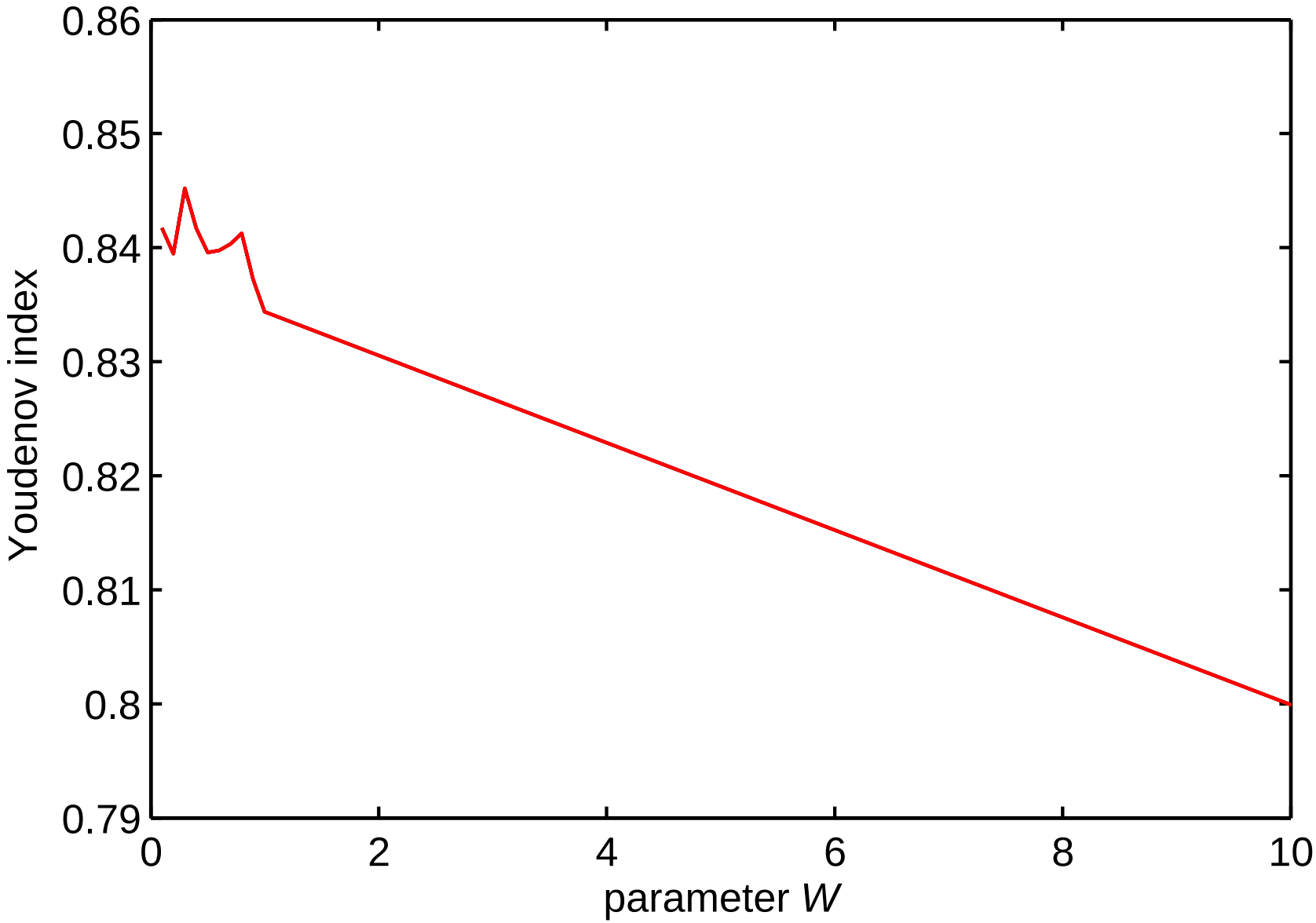
VÝSLEDKY PRE $N = 210$



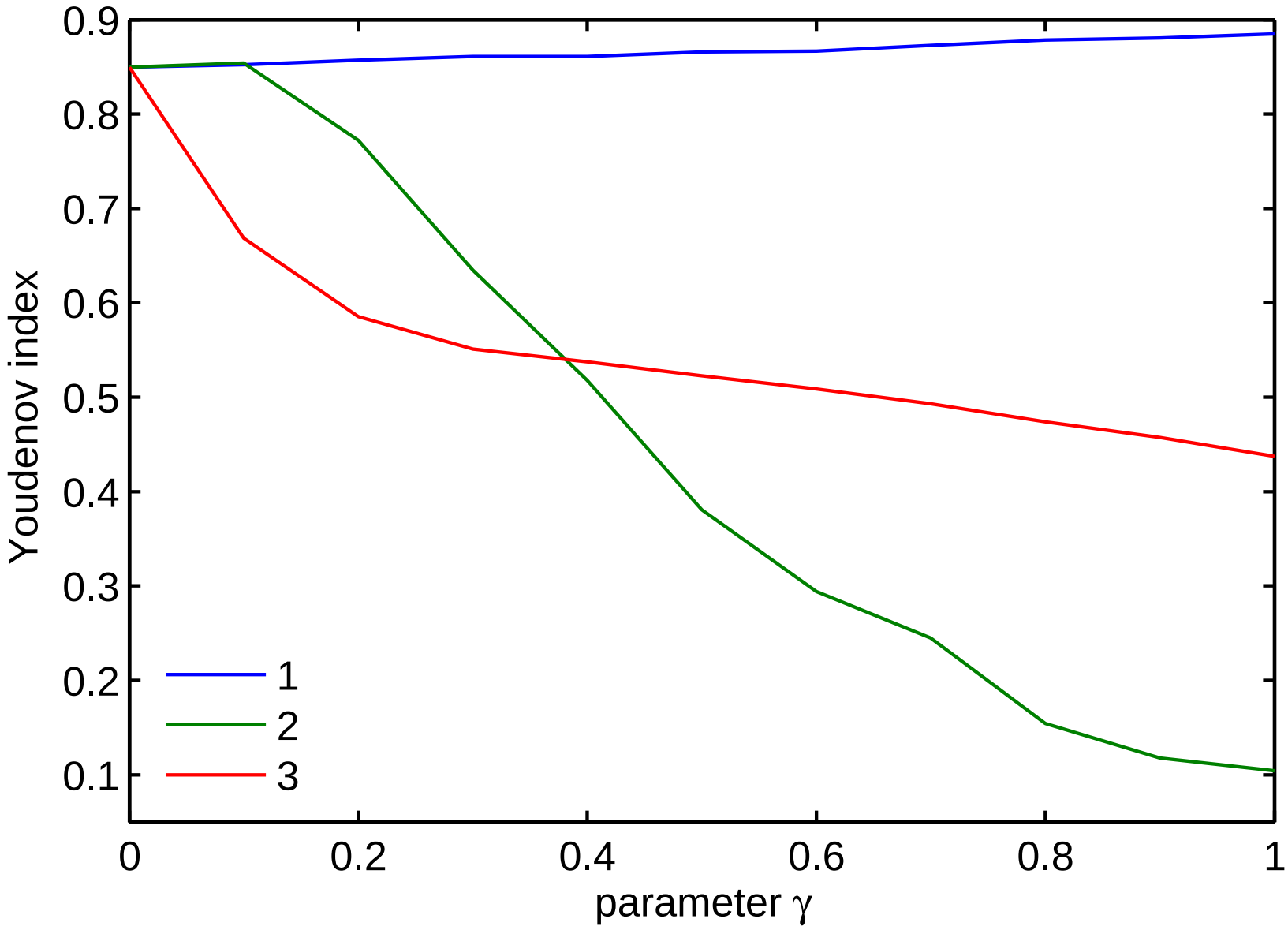
VÝSLEDKY PRE $N = 12$



VÝSLEDKY PRE $N = 12$



VÝSLEDKY PRE $N = 12$



VÝSLEDKY

- Predpokladom, že vtupné dáta sú zašumené, sa zvýšila efektívnosť klasifikácie
- ako aj vhodným odhadom parametrov šumu.

Ďakujem za pozornosť!