



Teorie kognitivních systémů

4 Robustní regresní techniky

- Problém robustní regrese
- Detekce outlierů/pákových bodů
- Least Absolute Deviations (LAD)
- Least Trimmed Squares (LTS)
- Iteratively Re-Weighted Least Squares (IRLS)
- RANSAC
- Theil-Sen Estimator





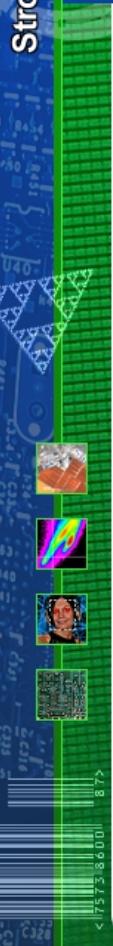
Robustní regrese

Definice problému

ROBUSTNÍ = ODOLNÝ (vůči šumu)

Je-li **metoda strojového učení** označovaná jako **robustní**, myslí se tím odolnost vůči (i) **nekvalitním**, (ii) **neúplným**, (iii) **zašuměným**, (iv) **nevhodně distribuovaným** nebo (v) vzájemně se vylučujícím vstupním datům (a samozřejmě také šumu ↑ v trénovací množině).

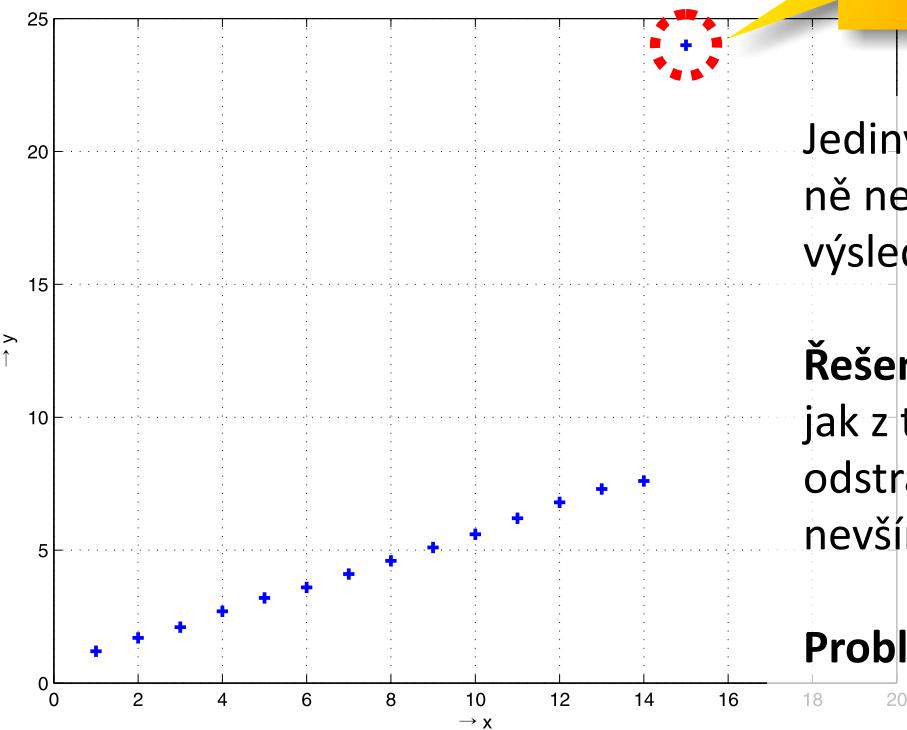
„Šum“ v případě regrese: Trénovací množina, tj. vstupní data, kterými se prokládá regresní funkce, obsahuje body, které podstatným způsobem negativně ovlivňují správnost nastavení parametrů Θ_0 až Θ_n hypotézy $h_{\Theta}(x)$.





Robustní regrese

Motivační příklad k lineární regresi



tzv. *outlier* nebo
též „pákový bod“

Jediný bod může zásadně negativně ovlivnit výsledný tvar hypotézy.

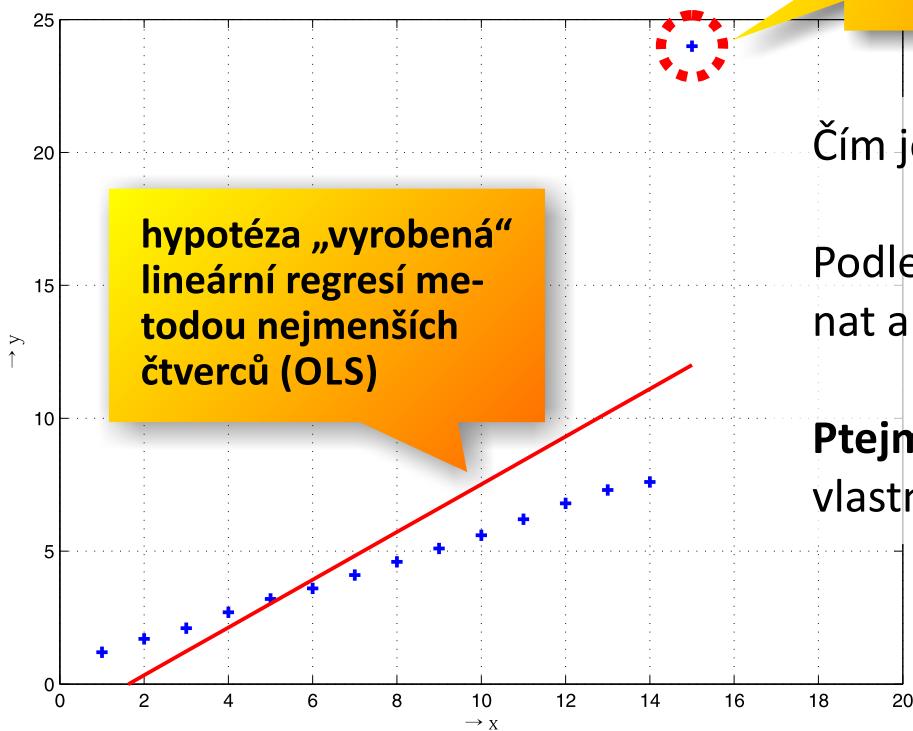
Řešení: Takový bod nějak z trénovací množiny odstranit (nebo si ho nevšímat).

Problém: Jak ho najít?



Robustní regrese

Motivační příklad k lineární regresi



hypotéza „vyrobená“
lineární regresí me-
todou nejmenších
čtverců (OLS)

tzv. *outlier* nebo
též „pákový bod“

Čím je outlier význačný?

Podle čeho by šel poz-
nat a z dat vyhodit?

Ptejme se jinak: Co to
vlastně outlier je?



Robustní regrese

Definice outlieru

- Bod v datech, který vznikl jiným procesem (např. fyzikálním) než ostatní data;
- bod, který se mezi data dostal v důsledku chyby (měření, zpracování dat, apod.);
- bod, který vznikl shodným procesem jako ostatní data, ale je zatížen náhlou nesystematickou fluktuací šumu.

Takové vlastnosti ale (snad) lze matematicky modelovat!

Tvorba modelu ale bohužel také závisí na rozpoznání (a vyloučení) outlierů.

U obyčejné metody nejmenších čtverců (**OLS** = *Ordinary Least Squares*) je hypotéza vychýlena směrem k outlieru tím více, čím „dále od zdravých“ dat se nachází → **metoda nejmenších čtverců není robustní**, tj. není vhodná na zašuměná data...



Robustní regrese

Definice bodu zhroucení (*Breakdown Point*)

- Bod zhroucení je poměrné vyjádření (obvykle v procentech) zastoupení outlierů v datech, při kterém algoritmus již přestává poskytovat věrohodné parametry hypotézy;
- je-li např. bod zhroucení 50% (u metody RANSAC), znamená to, že je-li outlierů v trénovací množině 50% a více, metoda již neposkytuje validní výsledky.

Z robustních regresních technik dostahuje nejvyššího bodu zhroucení metoda RANSAC (50%). Běžná přijatelná hodnota (aby mohl být algoritmus označen za robustní) je zhruba mezi 25% a 50%.

Vyšší hodnoty než 50% je těžko možné dosáhnout z principu: V trénovací množině je více dat, které hypotéze neodpovídají, než těch, které jí odpovídají.



Robustní regrese

Metody – alternativy k nejmenším čtvercům

- Nejmenší prořezané čtverce (*Least Trimmed Squares*)
- Nejmenší absolutní odchylky (*Least Absolute Deviations*)
- IRLS (*Iteratively Re-weighted Least Squares*)
- M-odhad (*M-estimation*)
- Theil-Sen Estimator

Přestože ve většině případů poskytují mnohem lepší výsledky než obyčejná metoda nejmenších čtverců, často nejsou v praxi příliš populární, ani používané. Důvodem může být:

- existence několika konkurenčních metod, není zřejmé kterou v dané situaci použít;
- tyto metody jsou teoreticky i výpočetně náročnější než obyč. nejmenší čtverce;
- nejsou implementovány v tradičně užívaných softwarových balících.



Least Trimmed Squares (LTS)

Nejmenší čtverce nad podmnožinou dat

Metoda nejmenších čtverců (OLS) spočívá v minimalizaci cenové funkce $J(\Theta)$, což je suma čtverců rozdílů predikce hypotézou $h_{\Theta}(x^{(i)})$ a skutečné hodnoty $y^{(i)}$ přes **všech m** bodů trén. mn., zatímco v případě LTS se sumuje **jen přes $k < m$** bodů trénovací množiny.

Tj. snažíme se vybrat $k < m$ bodů z trénovací množiny takových, že suma čtverců rozdílů bude nejmenší.

→ kombinatorický problém –

a také kombinatorická exploze, protože teoreticky lze vyrobit $m + m(m - 1) + m(m - 1)(m - 2) + \dots + (m - 1)!$ podmnožin trénovací množiny, nad kterými by se měla minimalizovat cenová funkce $J(\Theta) \Rightarrow \text{feasibility}$



Least Absolute Deviations (*LAD*)

Jednoduché řešení se zajímavými vlastnostmi

Také známé jako **Least Absolute Errors (LAE)**, **Least Absolute Value (LAV)** či **Least Absolute Residual (LAR)**.

Minimalizujeme cenovou funkci v tomto tvaru:

$$J(\Theta) = \sum_{i=1}^m |h_\Theta(\mathbf{x}^{(i)}) - y^{(i)}|.$$

tz. místo kvadrátu rozdílu odpovědi predikované hypotézou a odpovědi učitele se používá **absolutní hodnota**...

V důsledku toho:

- neexistuje analytická metoda řešení (vs normální rovnice);
- může existovat více řešení;
- řešení není stabilní.



Least Absolute Deviations (*LAD*)

Iterační postup nalezení řešení

- LAD lze řešit **iteračně**: v každém kroku iteračního postupu vypočteme matici chyby predikce, takto:

$$J(\Theta) = \sum_{i=1}^m \left| h_\Theta(\mathbf{x}^{(i)}) - y^{(i)} \right|.$$

$$\epsilon(h_{\Theta_k}(\mathbf{x}^{(i)})) = \epsilon_i$$

$$\mathbf{E}(\Theta_k) = \text{diag} \begin{pmatrix} \epsilon_1 & 0 & \cdots & 0 \\ 0 & \epsilon_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \epsilon_m \end{pmatrix}^{-1}$$





Least Absolute Deviations (*LAD*)

Iterační postup nalezení řešení

- Parametry hypotézy se na začátku iterace nastaví na 0 a pak se rekurentně upravují až do splnění kritéria konvergence:

$$\Theta_0 = 0$$

$$\Theta_1 = \left(\mathbf{X}^T \mathbf{E}(\Theta_0) \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{E}(\Theta_0) \mathbf{y}$$

$$\Theta_2 = \left(\mathbf{X}^T \mathbf{E}(\Theta_1) \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{E}(\Theta_1) \mathbf{y}$$

$\text{Th} = \text{pinv}(\mathbf{X}' * \mathbf{E} * \mathbf{X}) * \mathbf{X}' * \mathbf{E} * \mathbf{y}$

\vdots

$$\Theta_{k+1} = \left(\mathbf{X}^T \mathbf{E}(\Theta_k) \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{E}(\Theta_k) \mathbf{y}$$

\vdots

výsledné parametry

$$\Theta = \lim_{k \rightarrow \infty} \Theta_k$$



M-estimator

Zcela obecný přístup k řešení problému

- M- znamená **Maximum Likelihood**, tj. nejvyšší **věrohodnost**;
- nejmenší čtverce (OLS) jsou spec. případem M-estimátoru;
- **minimalizuje se suma funkcí dat** – takový proces se nazývá **M-odhad (M-estimation)**;
- funkce odhadu věrohodnosti dat je obvykle derivací nějaké jiné statistické funkce nad daty;
- obecný tvar cenové funkce je:

$$\arg \min_{\Theta} J(\Theta) = \arg \min_{\Theta} \left(- \sum_{i=1}^m \log \left(f(\mathbf{x}^{(i)}, \Theta) \right) \right)$$

Peter Huber (1964) navrhl zobecnění odhadu nejvyšší věrohodnosti ve tvaru:

$$\arg \min_{\Theta} J(\Theta) = \arg \min_{\Theta} \left(\sum_{i=1}^m \rho(\mathbf{x}^{(i)}, \Theta) \right)$$



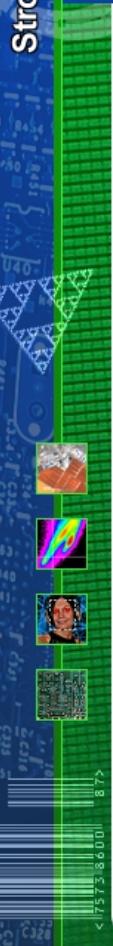
M-estimator

Zcela obecný přístup k řešení problému

$$\arg \min_{\Theta} J(\Theta) = \arg \min_{\Theta} \left(\sum_{i=1}^m \rho(\mathbf{x}^{(i)}, \Theta) \right)$$

- Při splnění určitých (nepříliš omezujících) požadavků na tvar funkce ρ je taková minimalizace vždy proveditelná, ať iteračně nebo nalezením kořene derivace položené rovno 0 (což ale může být velice matematicky komplikované).
- Existuje několik typů M-estimátorů, také tvaru funkce ρ může existovat mnoho...

Výpočetní postup, který dovoluje minimalizovat cenové funkce v tomto obecném tvaru se nazývá **Iteratively Re-Weighted Least Squares (IRLS)**.



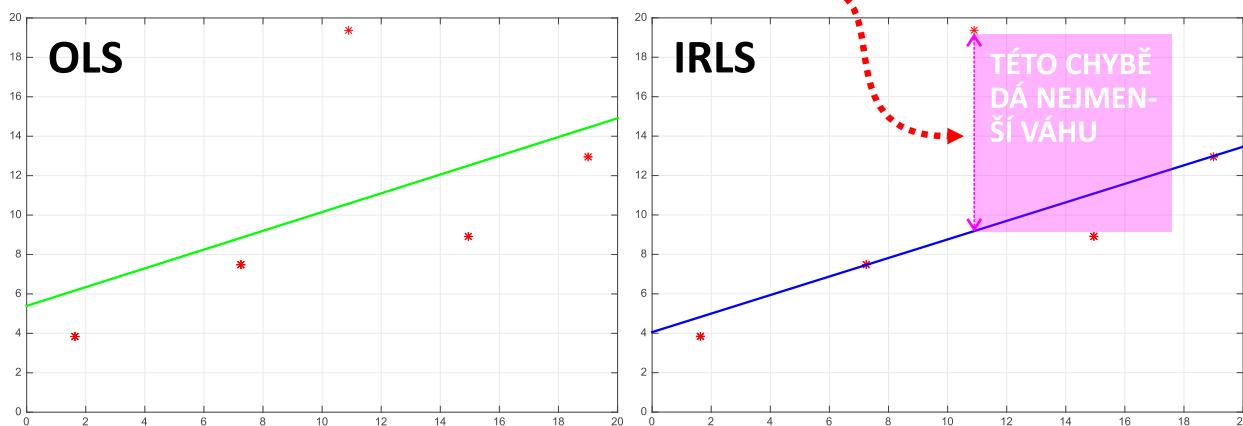


Iteratively Re-weighted Least Squares (IRLS)

Nejmenší čtverce s váhou chyby odhadu

Základní myšlenka: Každý sčítanec celkové chyby hypotézy může mít nějakou váhu, tzn. optimalizační proces si může některých chyb „všímat“ více a některých méně, tedy:

$$J(\Theta) = \sum_{i=1}^m w^{(i)} (h_\Theta(\mathbf{x}^{(i)}) - y^{(i)})^2$$



Jak ale vektor vah $w^{(i)}$ najít?



Iteratively Re-weighted Least Squares (IRLS)

Nalezení vektoru vah chyby odhadu

Iterační algoritmus:

- (1) začneme s váhami nastavenými na 1, $w^{(i)} = 1$ pro $i \in \langle 1, m \rangle$
tj. první odhad Θ odpovídá výsledku OLS
- (2) provedeme optimalizaci (nalezení Θ , pro které nabývá $J(\Theta)$ minima, např. via normální rovnice) dle kritéria:

$$\Theta_{(t+1)} = \arg \min_{\Theta} \sum_{i=1}^m w^{(i)}(\Theta_{(t)}) |h_{\Theta}(\mathbf{x}^{(i)}) - y^{(i)}|^2$$

- (3) vypočteme s hodnotami Θ získanými v kroku (2) váhu i-tého odhadu jako:

$$w^{(i)}(\Theta_{(t)}) = |\Theta_{(t)} \mathbf{x}^{(i)} - y^{(i)}|^{p-2}$$

$$w^{(i)}(\Theta_{(t)}) = \frac{1}{|\Theta_{(t)} \mathbf{x}^{(i)} - y^{(i)}|}$$

volba L^p
normy

$p = 1$: LAD
 $p = 2$: OLS

- (4) pokud $|\Theta_{(t+1)} - \Theta_{(t)}| < \epsilon$, pak konec, jinak goto (2)



Iteratively Re-weighted Least Squares (IRLS)

Algoritmus v MATLABu/Octave

```
w = ones([1 tss]); W = diag(w);
LastTheta = [0 0]';

while true
    Theta = inv(X' * W * X) * X' * W * y;

    for i = 1:tss
        hdiff = y(i) - X(i, :) * Theta;
        W(i, i) = 1 / abs(hdiff);
    end

    nr = norm(LastTheta - Theta);
    if nr < 0.1
        break
    end

    LastTheta = Theta;
end
```

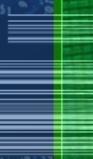
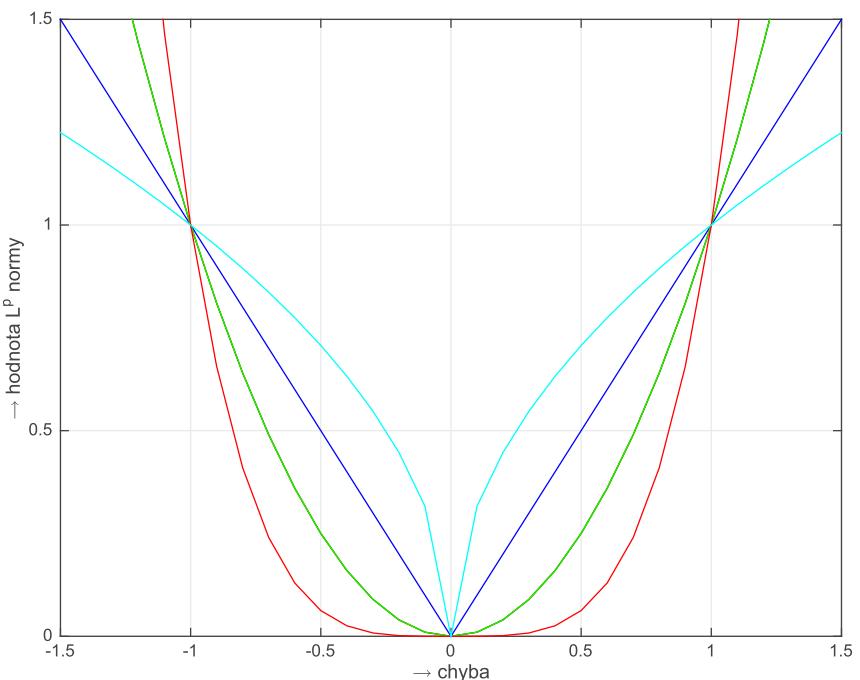
velikost trénovací množiny (training set size)





Iteratively Re-weighted Least Squares (IRLS)

Různé L^p normy





RANSAC

Shoda náhodných výběrů

RANSAC (RANdom Sample And Consensus) – metodu uvedli roku 1981 Fischler a Bolles. Zvládá **bod zhroucení (Breakdown Point)** větší než 50%, tzn. více než polovina dat jsou outliersy.

Algoritmus RANSAC — iterativně se opakují 2 kroky:

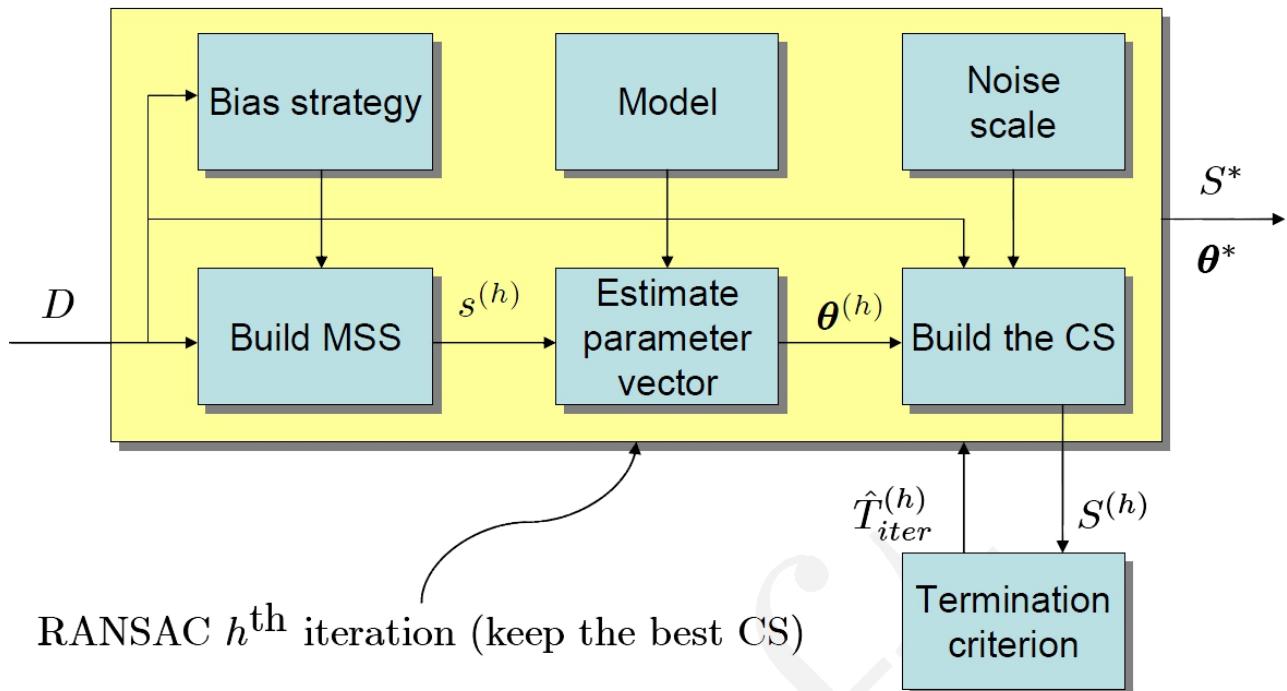
- (1) **Tvorba hypotézy** — z trénovací množiny se náhodně vybere minimální množina vzorků (tzn. jen tolik vzorků z trénovací mn., kolik je třeba k jednoznačnému určení parametrů hypotézy, tj. pro lineární regresi 2), vypočte se optimální $h_{\Theta}(x)$ (např. LinR).
- (2) **Test** — testuje se, kolik vzorků z trénovací množiny neobsažených v náhodném výběru ↑, vyhovuje vytvořené hypotéze ↑. Množina takových vzorků se nazývá *Consensus Set (CS)*.

Konec iterace: Nelze-li najít lepší (větší) CS.



RANSAC

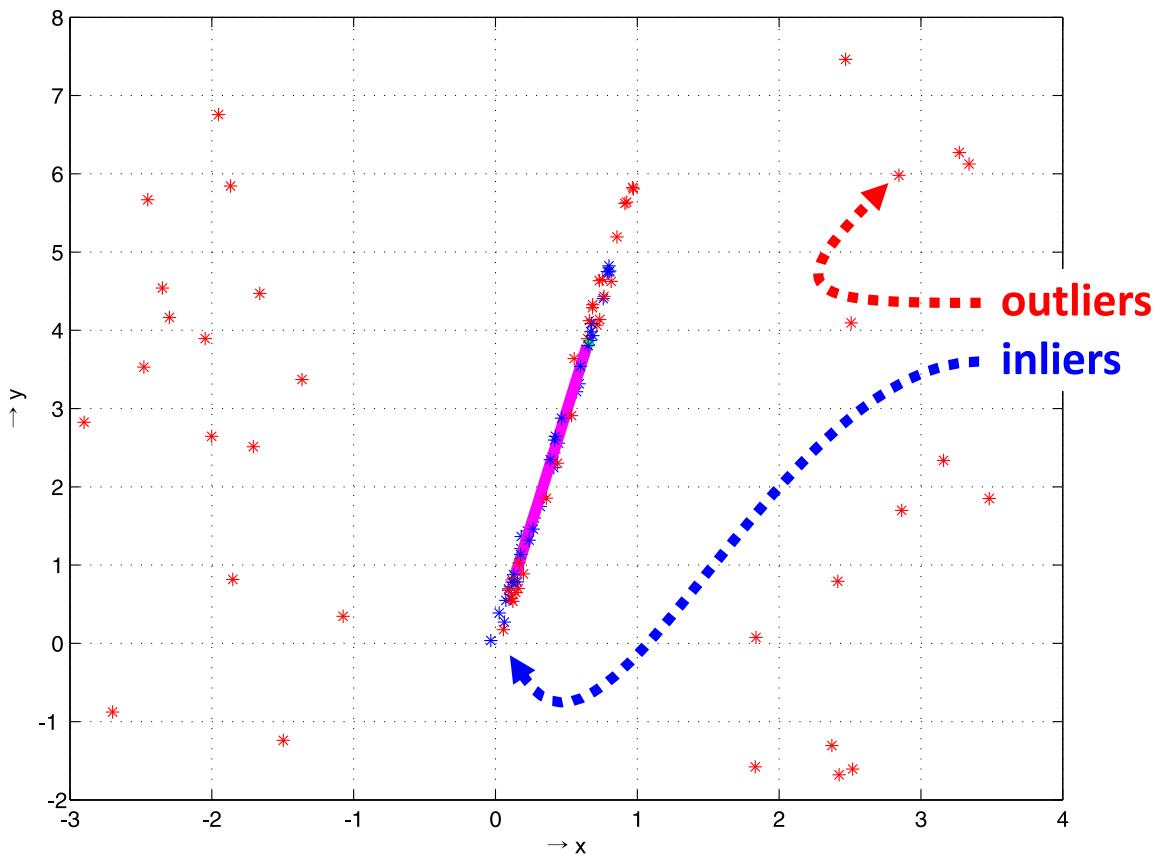
Blokové schéma obecné podoby algoritmu





RANSAC

Výsledek na velmi nekvalitních datech





Theil–Sen Estimator

Geometrická metoda robustní lineární regrese

Navržena Henri Theilem (1950) a vylepšena Pranabem Senem (1968), také známá jako ***Sen's Slope Estimator***, ***Slope Selection***, ***Single Median Method***, ***Kendall Robust Line-Fit Method*** nebo ***Kendall–Theil Robust Line***.

- Efektivní a snadno naprogramovatelný výpočet, **$O(n^2)$** ;
- málo citlivý na šum v datech, **breakdown point $\approx 29,3\%$** ($1 - 1 / \sqrt{2}$), tzn. až 29,3% dat mohou být outliersy, aniž by došlo ke snížení přesnosti odhadu hypotézy;
- tradiční použití v astronomii, biofyzice, DPZ (odhad plochy listů z data odrazivosti povrchu), v IT se používá k odhadům trendů stárnutí software.



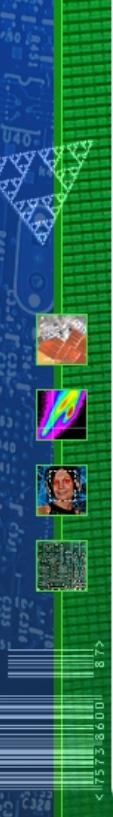


Theil–Sen Estimator

Popis algoritmu

- Vytvoříme $m \times (m - 1)$ hypotéz (přímek) ve tvaru $y = kx + q$ (nebo chcete-li $h_{\Theta}(x) = \Theta_0 + \Theta_1 x$) tak, že vezmeme každý bod z trénovací množiny a proložíme jím přímky procházející všemi ostatními body;
- z takto vytvořených hypotéz pak vypočteme průměrnou hypotézu jako medián – výpočtu mediánu se ale smí zúčastnit jen ty hypotézy, které mají koeficient k (tedy sklon přímky) nenulový, tj. nejsou konstantami.

Konstanty jsou vyloučeny proto, že jejich příspěvek k průměru je nulový, ale zvýšíly by počet průměrovaných položek, takže by výsledná průměrná hodnota „sklonu“ hypotézy byla menší, než by měla být...





Theil–Sen Estimator

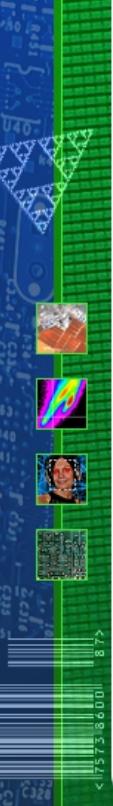
Algoritmus v pseudokódu

Algoritmus 1 Theil–Sen Estimator

```

1: ▷ trénovací množina  $T = \{(x^{(1)}, y^{(1)}), \dots, (x^{(m)}, y^{(m)})\}$ 
2: func THEILSEN(trénovací množina T)
3:   var pole hypotéz  $\mathbf{H}[m \cdot (m - 1)]$ 
4:   var čítač  $c \leftarrow 0$ 
5:   for  $i \in \langle 1, m \rangle :$ 
6:     for  $j \in \langle 1, m \rangle :$ 
7:       if  $i \neq j$  then
8:          $k = \frac{y^{(j)} - y^{(i)}}{x^{(j)} - x^{(i)}}; q = -kx^{(i)} + y^{(i)}$ 
9:          $c \leftarrow c + 1$ 
10:         $\mathbf{H}[c] = (k, q)$ 
11:   for  $i \in \langle 1, m \cdot (m - 1) \rangle :$ 
12:     if  $\mathbf{H}[i].k = 0$  then
13:       odstraň  $\mathbf{H}[i]$ 
14:    $k_{ret} \leftarrow \text{median}(\mathbf{H}[\cdot].k)$ 
15:    $q_{ret} \leftarrow \text{median}(\mathbf{H}[\cdot].q)$ 
16:   return ( $k_{ret}, q_{ret}$ )
    }
```

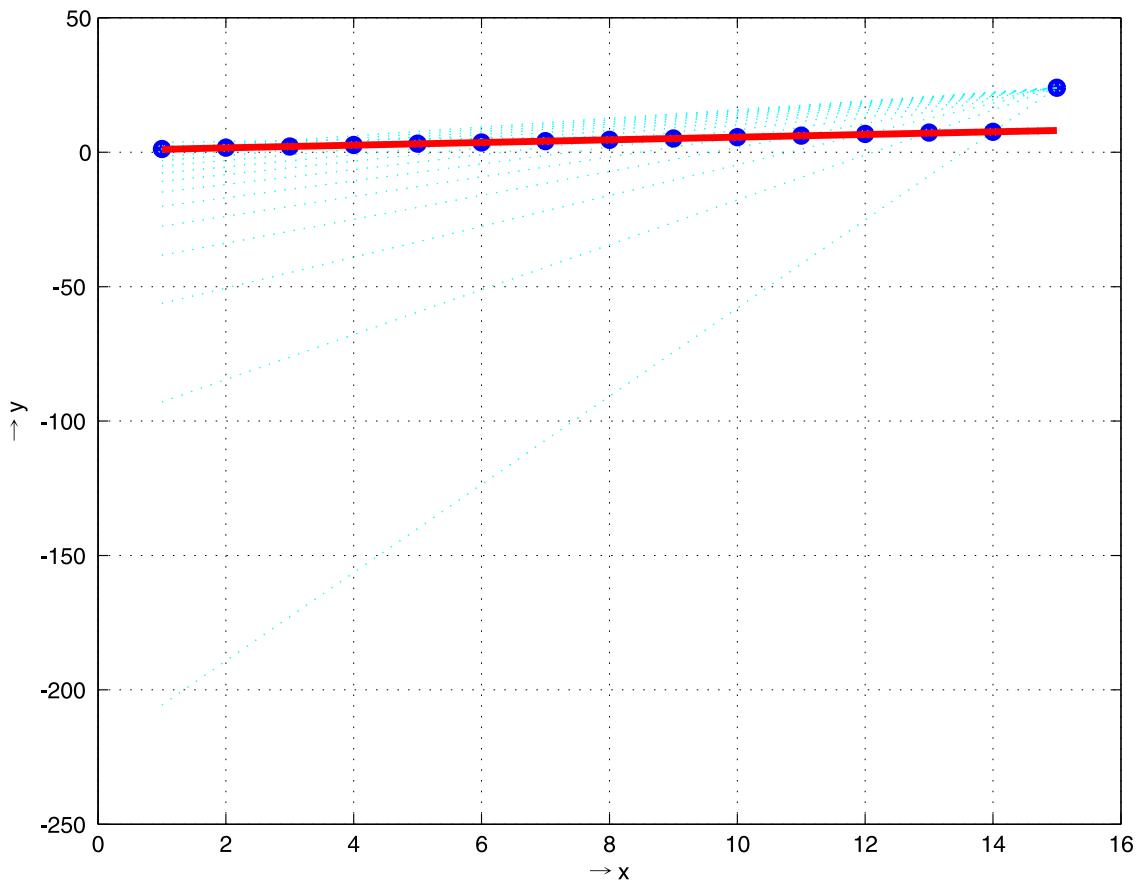
Senova úprava pův.
Theilova algoritmu





Theil–Sen Estimator

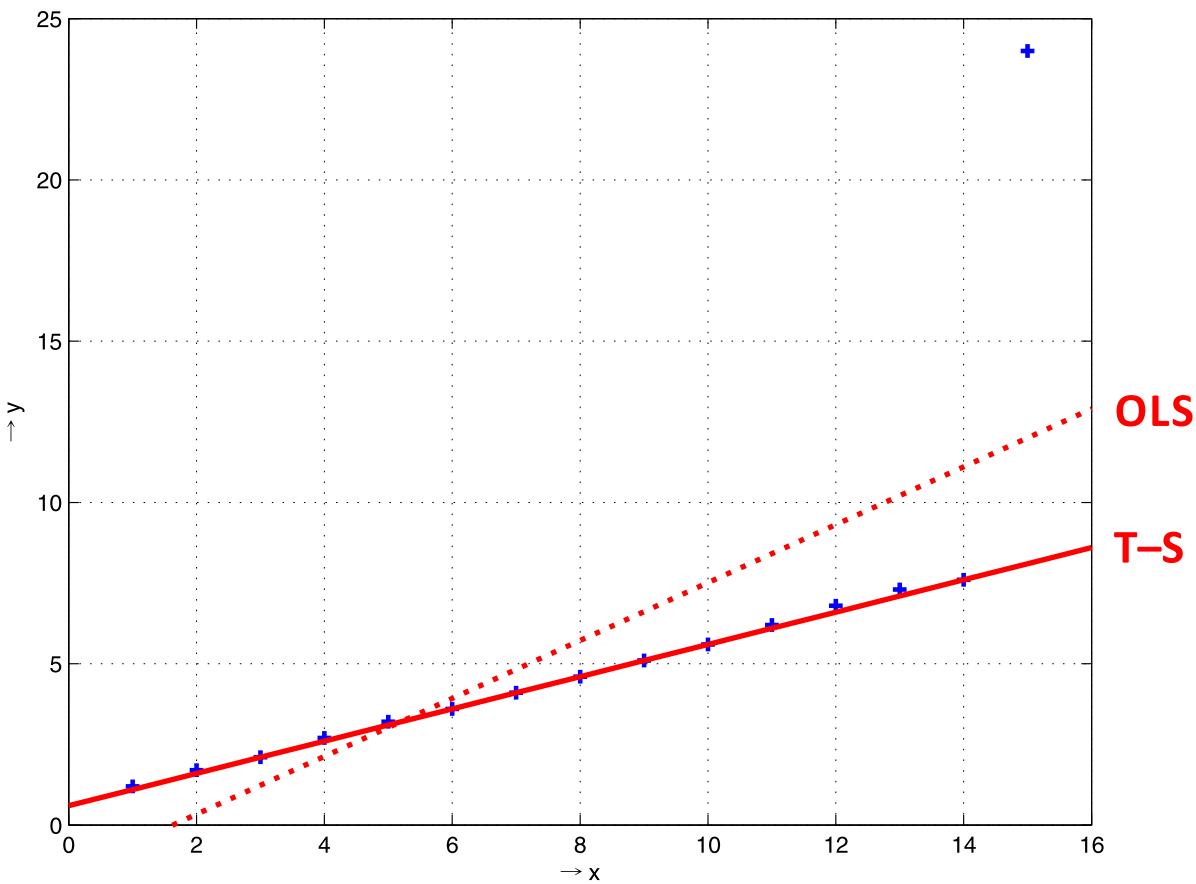
Zobrazení principu odhadu hypotézy





Theil–Sen Estimator

Výkon: Theil–Sen vs nejmenší čtverce (OLS)





Theil–Sen Estimator

Úpravy základního algoritmu

- **Siegel (1982)** – vypočítá se medián všech hypotéz procházejících jedním bodem, celková výsledná hypotéza se počítá jako medián z těchto mediánů.
- **Párování hypotéz** – hypotézy se párují podle řádu x-ové souřadnice (hypotéza s nejnižší hodnotou x je v páru s první hypotézou nad mediánem), celková výsledná hypotéza se počítá jako medián z těchto párů.
- **Vážené mediány** – větší váhu dostávají páry vzorků, jejichž x-ové souřadnice se od sebe více liší.

