Randomizované metody nelineární regrese

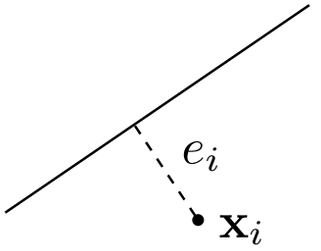
1. Připomenutí obyčejné regrese
2. RANSAC (Random Sample Consensus) [Fishler & Bolles 1981]
3. MLESAC/MAPSAC
   (Maximum Likelihood RANSAC, Maximum Aposteriori Probability RANSAC)
4. LMS (Least Median of Squares) [Rousseeuw 1984]

Přednášky k tématu

[1] 25 Years of RANSAC. Workshop in conjunction with CVPR 2006.
http://cmp.felk.cvut.cz/ransac-cvpr2006/

Na úvod: přímková regrese

Normální rovnice přímky: $\mathbf{a}^\top \mathbf{x} + b = 0, \quad \|\mathbf{a}\| = 1, \quad$ parametry $\theta = (\mathbf{a}, b)$.



Vzdálenost bodu $\mathbf{x}_i$ od přímky v normálním tvaru:

$$e_i = \|\mathbf{a}^\top \mathbf{x}_i + b\| \qquad \|a\|^2 = a^\top a = 1$$

Regrese je nalezení θ řešením optimalizačního problému za omezující podmínky $\|\mathbf{a}\| = 1$

$$\theta^* = \arg\min_\theta \sum_i e_i^2(\theta)$$

3×3

Řešení: zavedeme $\tilde{\mathbf{x}} = [\mathbf{x}, 1]$, pak $(\tilde{x}_i \theta)^\top (\tilde{x}_i \theta) = \theta^\top (\tilde{x}_i^\top \tilde{x}_i) \theta$

$\theta^* = \arg\min_\theta \theta^\top M \theta$

$[v, D, v] = \text{svd}(X * X')$

$v(:, \text{end})$

$$\sum_i (\tilde{\mathbf{x}}_i^\top \theta)^2 = \cdots = \theta^\top \left(\sum_i \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top \right) \theta = \theta^\top \mathbf{M} \theta$$

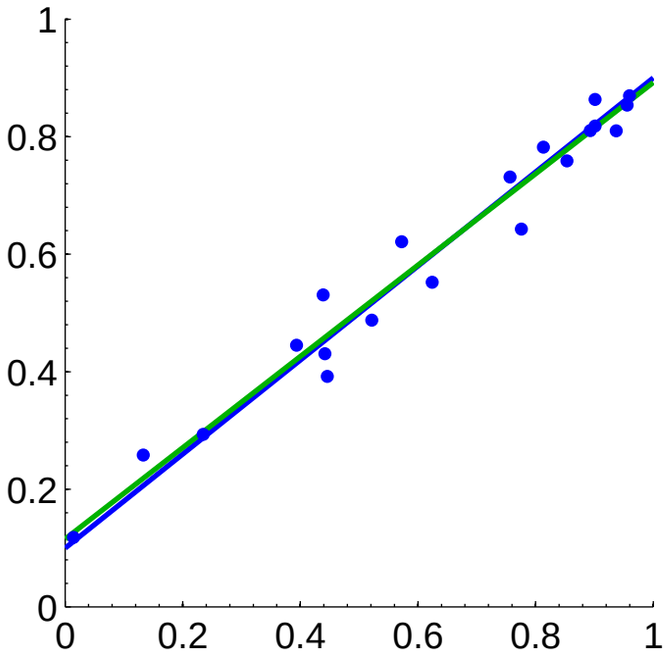
$M = \sigma_1^2 m_1 m_1^\top + \sigma_2^2 m_2 m_2^\top + \sigma_3^2 m_3 m_3^\top$

$\sigma_1 \geq \sigma_2 \geq \sigma_3 \qquad m_i^\top m_j = \delta(i,j)$

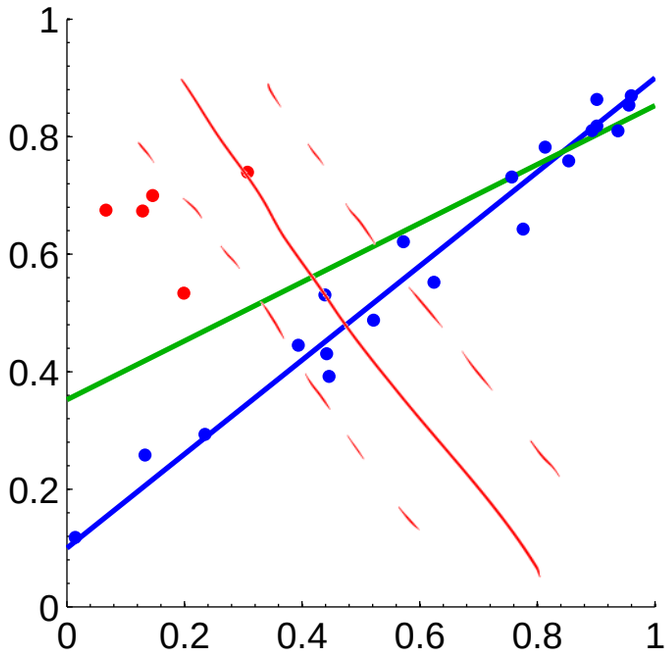
- θ je násobkem vlastního vektoru $\mathbf{v}_0$ matice $\mathbf{M}$, který odpovídá nejmenšímu vlastnímu číslu λ_0 ⊛2 za důkaz
- tj. $\theta = \alpha\, \mathbf{v}_0, \quad \alpha$ určíme z podmínky $\|\mathbf{a}\| = \|\theta_{1:2}\| = 1$

Co se stane s klasickou regresí při kontaminaci dat jiným procesem?

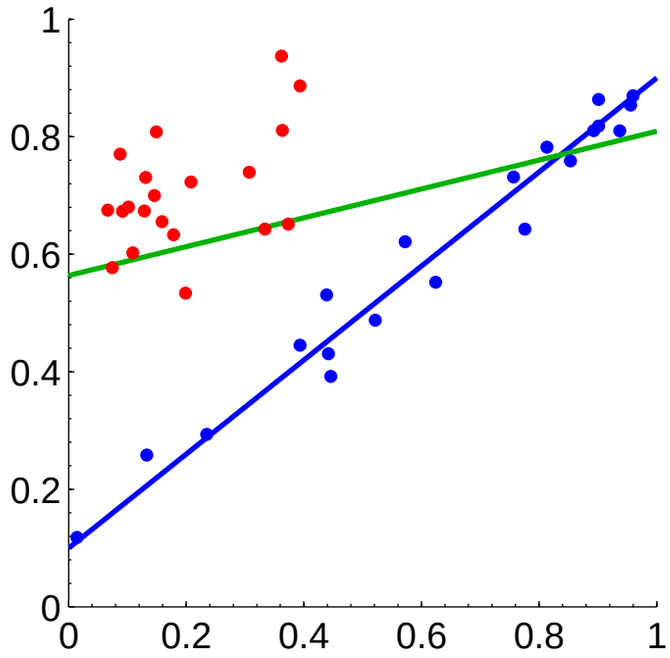
20 nekontaminovaných bodů



5 kontaminujících bodů



20 kontaminujících bodů



odhadnutá a skutečná přímka

červené body nepatří do přímky

- přímka: $\mathbf{a}^\top \mathbf{x} + b + \nu(0; 0.03) = 0$
- další proces: $\mathbf{x} = \mu([0.2, 0.7]^\top; 0.1)$
- ν, μ – normální rozdělení

Problém:

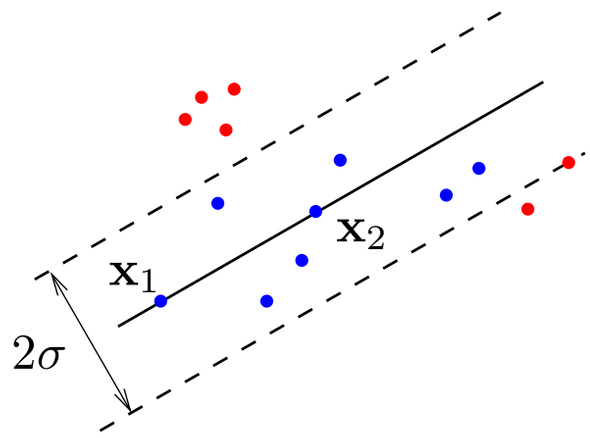
1. pro každý bod zjistit hodnotu binární proměnné $c(\mathbf{x}_i)$
2. nalézt parametry přímky $\theta = (\mathbf{a}, b)$ pro body $c(\mathbf{x}_i) = 1$

Základní forma RANSACu

Dáno:

1. množina bodů $P = \{\mathbf{x}_i \mid i = 1, 2, \ldots, k\}$
2. toleranční práh σ odhad rozptylu šumu
3. počet pokusů n

$\binom{k}{2} = O(k^2)$



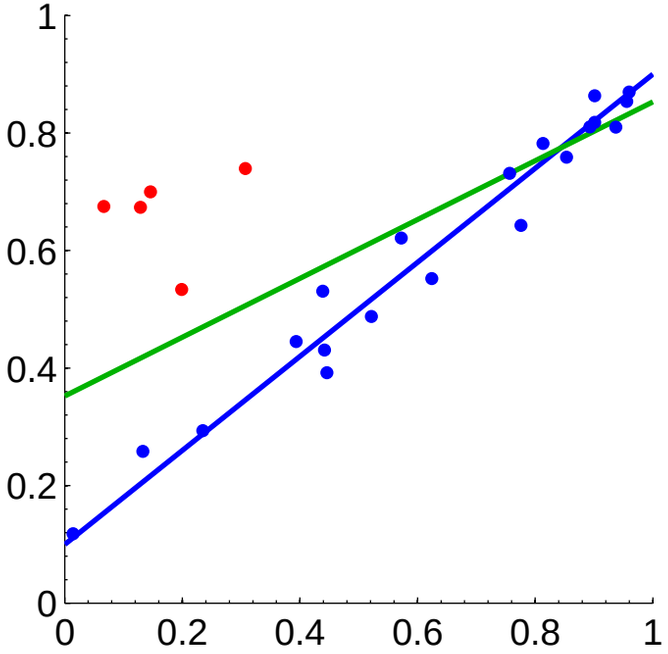
Procedura:

1. Inicializuj $C_{\max} := \emptyset$.
2. Opakuj pro $j = 1, 2, \ldots, n$:
   a. Náhodně vyber dvojici bodů $\mathbf{x}_1$, $\mathbf{x}_2$ z P $x_1 \neq x_2$
   b. Z $\mathbf{x}_1$, $\mathbf{x}_2$ vypočti parametry přímky $\theta_j := (\mathbf{a}_j, b_j)$
   c. Vypočti vzdálenost $e_1, e_2, \ldots e_k$ všech bodů $\mathbf{x}_i$ vzhledem k θ_j
   d. Nalezni *množinu konsensu*
   $$C_j = \left\{\mathbf{x}_i \mid \frac{|e_i|}{\sigma} < 1,\ i = 1, 2, \ldots, k\right\}$$
   e. Pokud je C_j větší než $C_{\max}$, potom $C_{\max} := C_j$
3. Vypočti $\theta^* := (\mathbf{a}, b)$ z bodů v $C_{\max}$ obyčejnou regresí

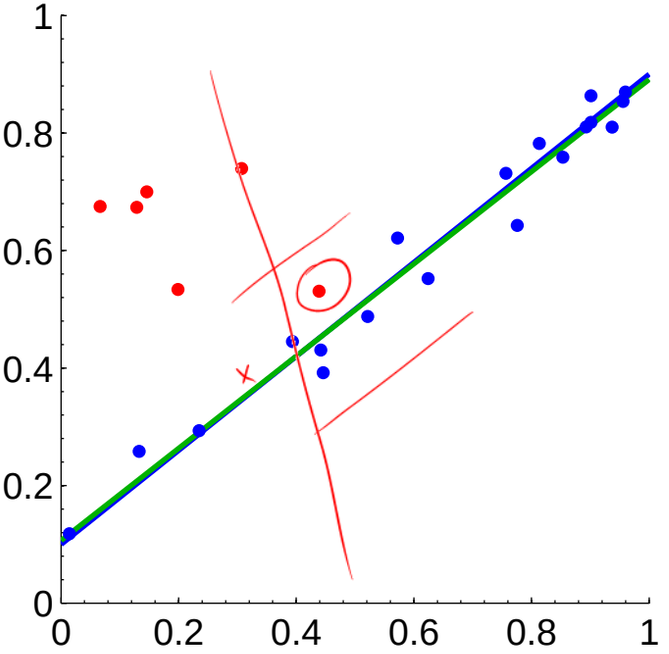
Výsledek regresí z $C_{\max}$

5 kontaminujících bodů

nerobustní regrese



RANSAC



- $\sigma = 0.06$
- identifikováno 19 bodů z 20

Kolik pokusů n?

- RANSAC je randomizovaný algoritmus
- není zaručeno, že nalezne správnou množinu $C_{\max}$
- můžeme jen požadovat, aby pravděpodobnost, že nalezne správnou $C_{\max}$ byla p
- vysoké $p \Rightarrow$ dlouhá doba běhu

Abychom dosáhli předepsané p, počet pokusů n musí být

$$n \geq \frac{\log(1-p)}{\log(1-(1-w)^s)}$$

- p – pravděpodobnost, že alespoň jedna selekce je správná
- w – procento kontaminujících bodů v P
- s – velikost minimálního výběru nutného pro výpočet θ

	$s=2$		$s=10$	
	p		p	
w	0.8	0.99	0.8	0.99
0.2	2	5	15	41
0.5	6	17	1648	4714
0.8	40	113	$1.6 \cdot 10^7$	$4.5 \cdot 10^7$

$(1-w)^s =$ selekce neobsahuje kontaminující bod

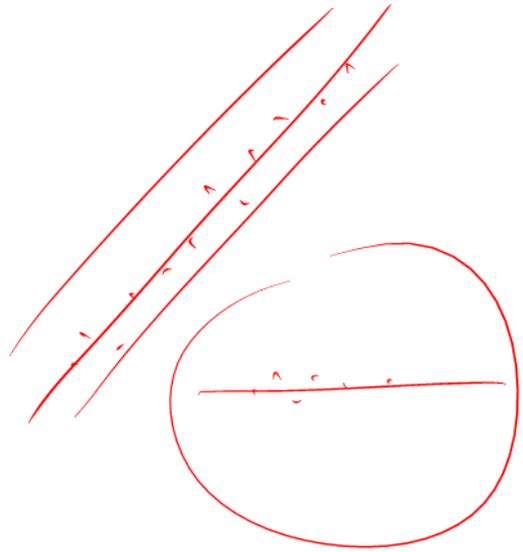
$1-(1-w)^s =$ selekce obsahuje alespoň jeden kontaminující bod

$1-p =$ všechny selekce obsahovaly kontaminující bod $= (1-(1-w)^s)^n$

Odhad n při neznámém w

Odhad w a n v průběhu algoritmu

1. $n := \infty$, $i := 0$
2. Dokud $i < n$ opakuj:
   a. proveď výběr a spočti $|C_{\max}|$
   b. $w := 1 - \frac{|C_{\max}|}{|P|}$
   c. $n := \frac{\log(1-p)}{\log(1-(1-w)^s)}$
   d. $i := i + 1$



- na našem problému s přímkou s 5 kontaminujícími body provede v průměru 8.14 kroků při $\sigma = 0.06$, $p = 0.99$ (100 pokusů)