Randomizované metody nelineární regrese

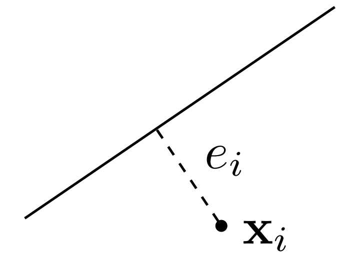
1. Připomenutí obyčejné regrese
2. RANSAC (Random Sample Consensus) [Fishler & Bolles 1981]
3. MLESAC/MAPSAC
   (Maximum Likelihood RANSAC, Maximum Aposteriori Probability RANSAC)
4. LMS (Least Median of Squares) [Rousseeuw 1984]

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

Rovnice přímky: $\mathbf{a}^\top \mathbf{x} + b = 0$, parametry $\theta = (\mathbf{a}, b)$.



Algebraická vzdálenost bodu $\mathbf{x}_i$ od přímky:

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

Regrese je nalezení θ řešením optimalizačního problému

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

$$\Big| \underbrace{[a^\top, b]}_{\theta} \cdot \tilde{x} \Big| = |a^\top x + b|$$

Řešení: zavedeme $\tilde{\mathbf{x}} = [\mathbf{x}, 1]$, pak

$$\sum_i (x_i^\top \theta)^\top (x_i^\top \theta) = \sum_i \theta^\top x_i x_i^\top \theta$$

$$\sum_i (\tilde{\mathbf{x}}_i^\top \theta)^2 = \cdots = \theta^\top \left(\sum_i \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top \right) \theta = \theta^\top \mathbf{M}\, \theta, \qquad \|\theta\| = c > 0$$

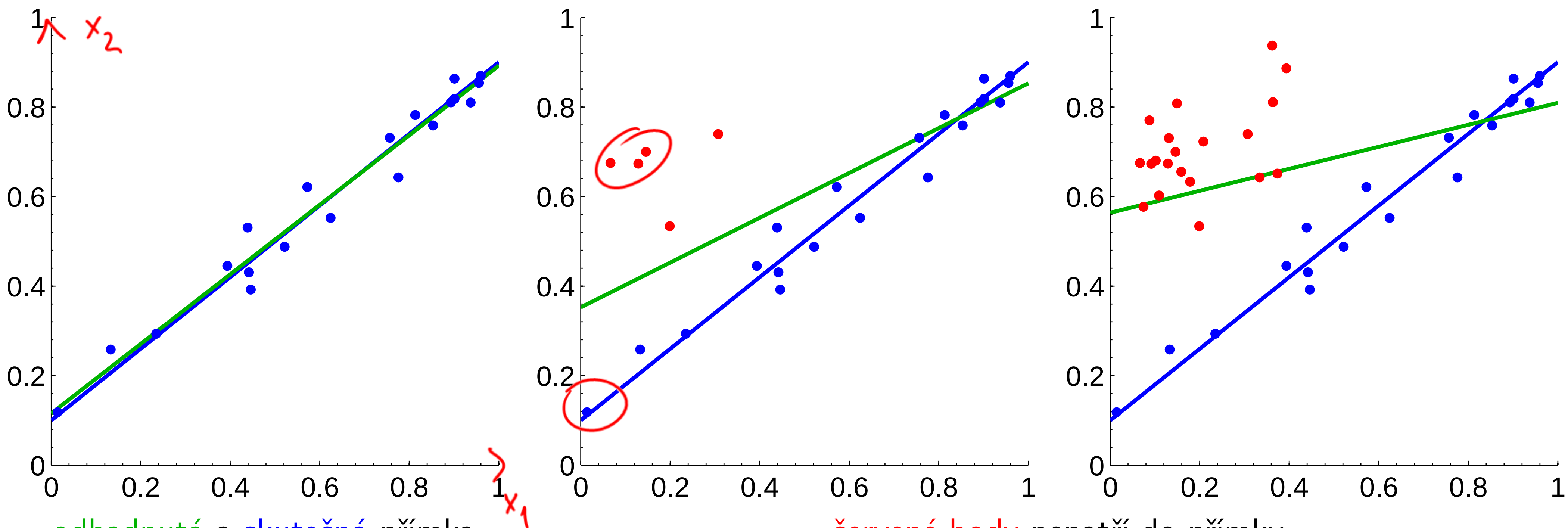
$$m_i^\top m_j = \begin{cases} 1 & i = j \\ 0 & \text{jinak} \end{cases}$$

$$\mathbf{M} = \sigma_1\, \mathbf{m}_1 \mathbf{m}_1^\top + \sigma_2\, \mathbf{m}_2 \mathbf{m}_2^\top + \sigma_3\, \mathbf{m}_3 \mathbf{m}_3^\top, \quad \sigma_1 \geq \sigma_2 \geq \sigma_3 \geq 0$$ toto je SVD

- θ je násobkem vlastního vektoru $\mathbf{v}_0$ matice $\mathbf{M}$, který odpovídá nejmenšímu vlastnímu číslu λ_0 $\circledast$1 za důkaz
- tj. $\theta = \alpha\, \mathbf{v}_0$, α 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

$a_1 x_1 + a_2 x_2 + b = 0$

- 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

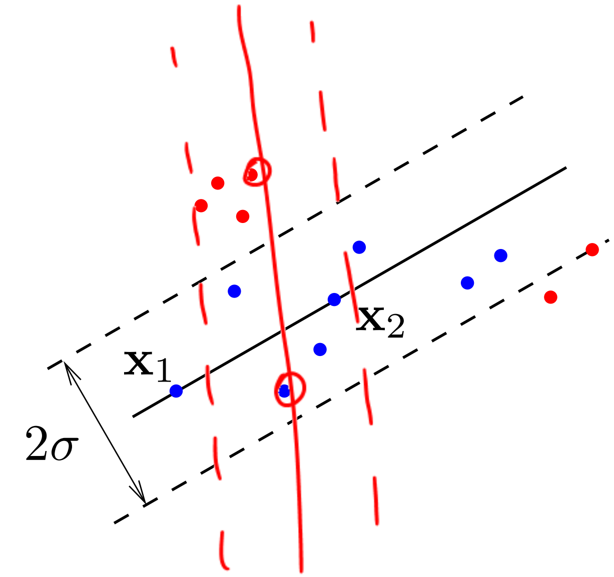
Problém:

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, tzv. ‚měřítko'
3. počet pokusů n



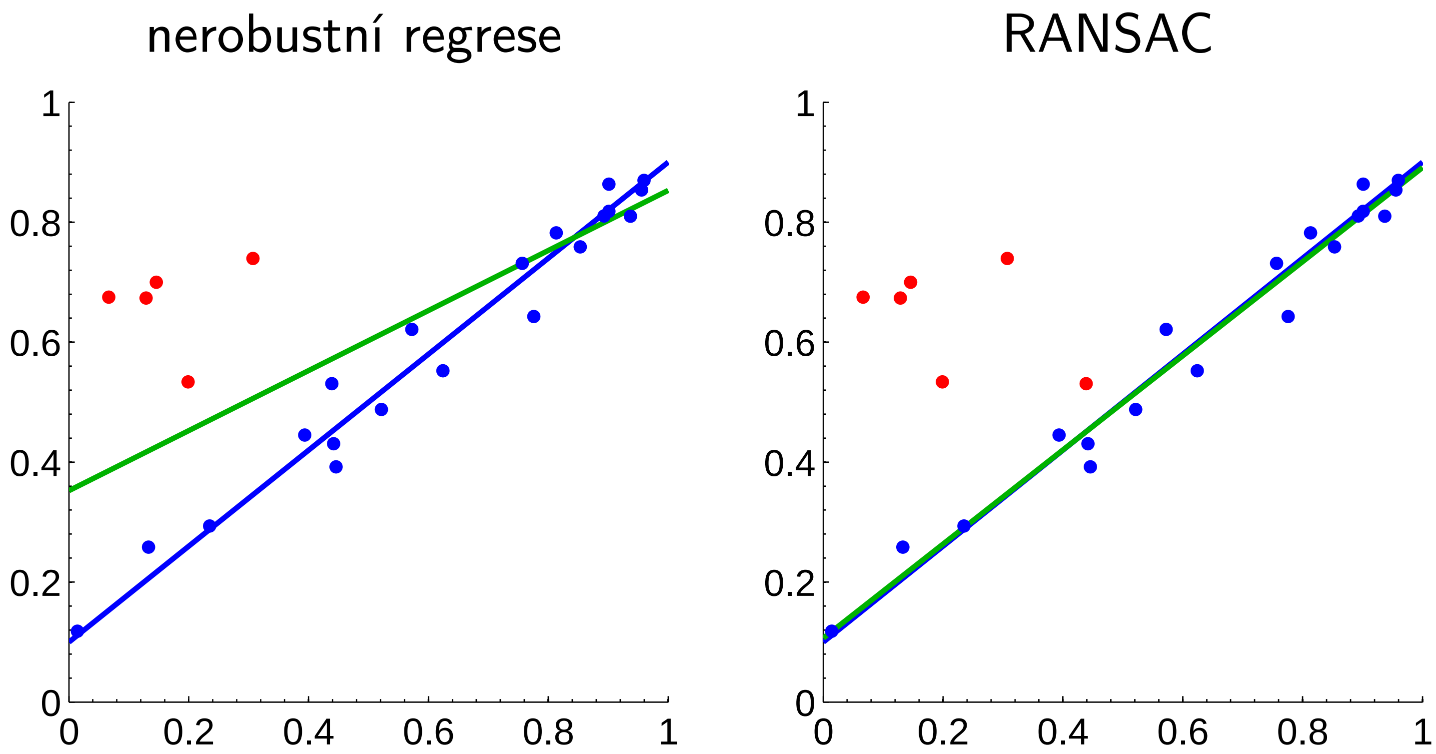
Procedura:

$\binom{k}{2}$

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
   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



- použito $\sigma = 0.06$
- 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

příklad s přímkou: $s = 2$

	$s=2$		$s=10$	
	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	$1.6\cdot10^7$	$4.5\cdot10^7$

$(1-w)^s =$ 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

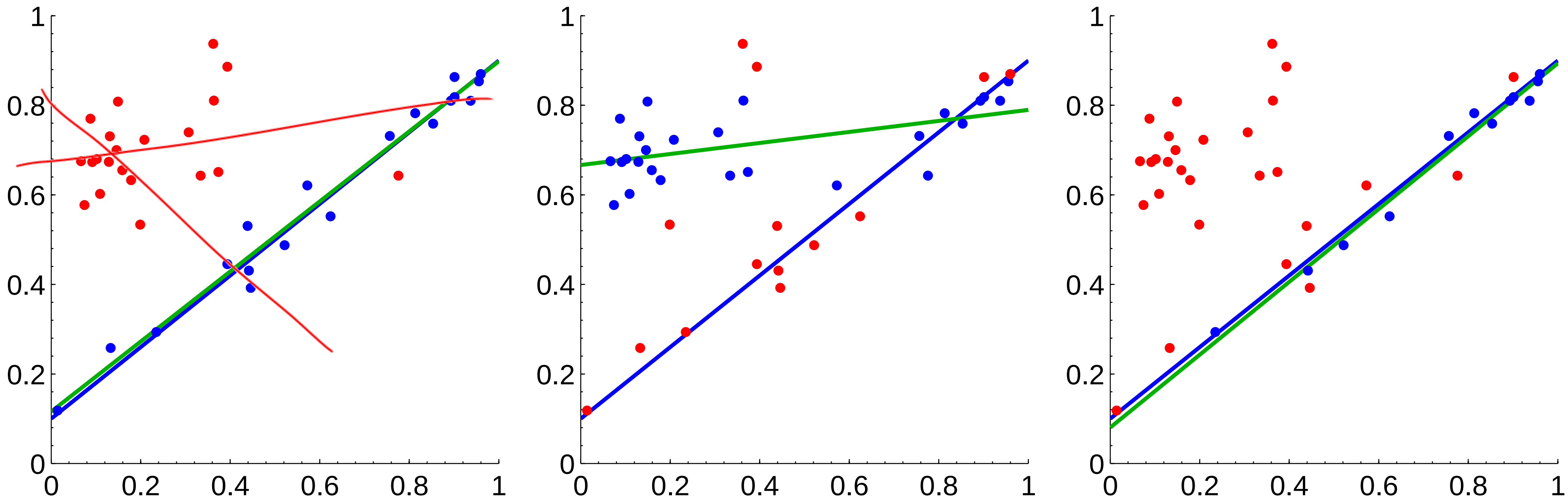
1. $n := \infty$, $i := 0$
2. Dokud $i < n$ opakuj:
   a. proveď výběr a spočti $|C_j|$
   b. pokud $|C_j| > |C_{\max}|$
      i. $C_{\max} := C_j$
      ii. $w := 1 - \frac{|C_{\max}|}{|P|}$
      iii. $n := \frac{\log(1-p)}{\log(1-(1-w)^s)}$
3. $i := i + 1$

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

Problémy základního RANSACu

použito $p = 0.99$

správné $\sigma = 0.06$ | velké $\sigma = 0.12$ | malé $\sigma = 0.03$



● nesprávný model

● malé zhodnocení dat

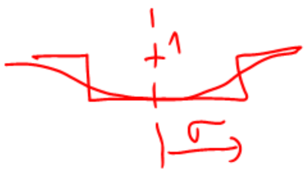
● velká chyba

Řešení

1. Robustní odhadování of σ [Wang & Suter IEEE Trans PAMI 26(11):1459-1474]
2. MLESAC [Torr CVIU 78(1):138-156]

MLESAC a MAPSAC

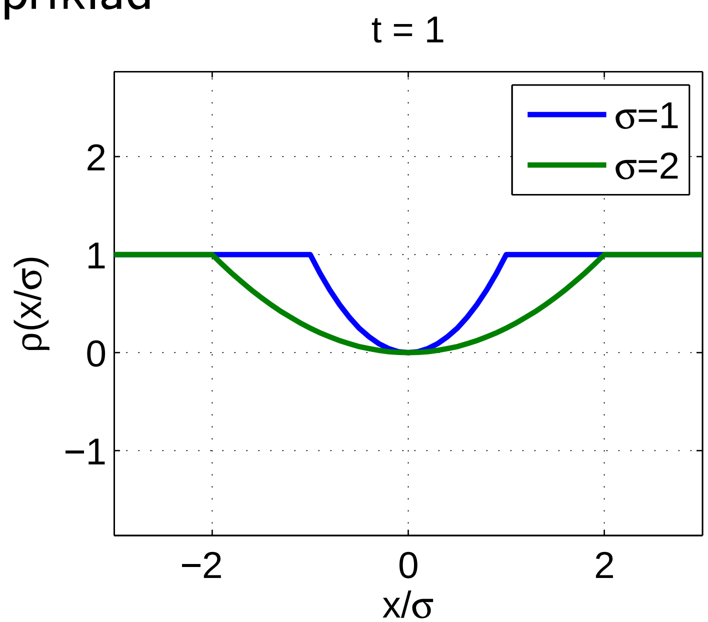
- místo maximalizace $|C_{\max}|$ lze minimalizovat počet detekovaných kontaminujících bodů:



$$|P| - |C_{\max}| = \sum_i \mathbf{1}(e_i^2 > \sigma^2), \qquad \mathbf{1}(x) = \begin{cases} 1 & \text{když } x \text{ platí} \\ 0 & \text{jinak} \end{cases}$$

- RANSAC může minimalizovat obecnou chybu, například MLESAC

$$E = \sum_i \rho(e_i), \quad \rho(x) = \begin{cases} \frac{x^2}{\sigma^2} & \frac{x}{\sigma} < t \\ t^2 & \frac{x}{\sigma} \geq t \end{cases}$$



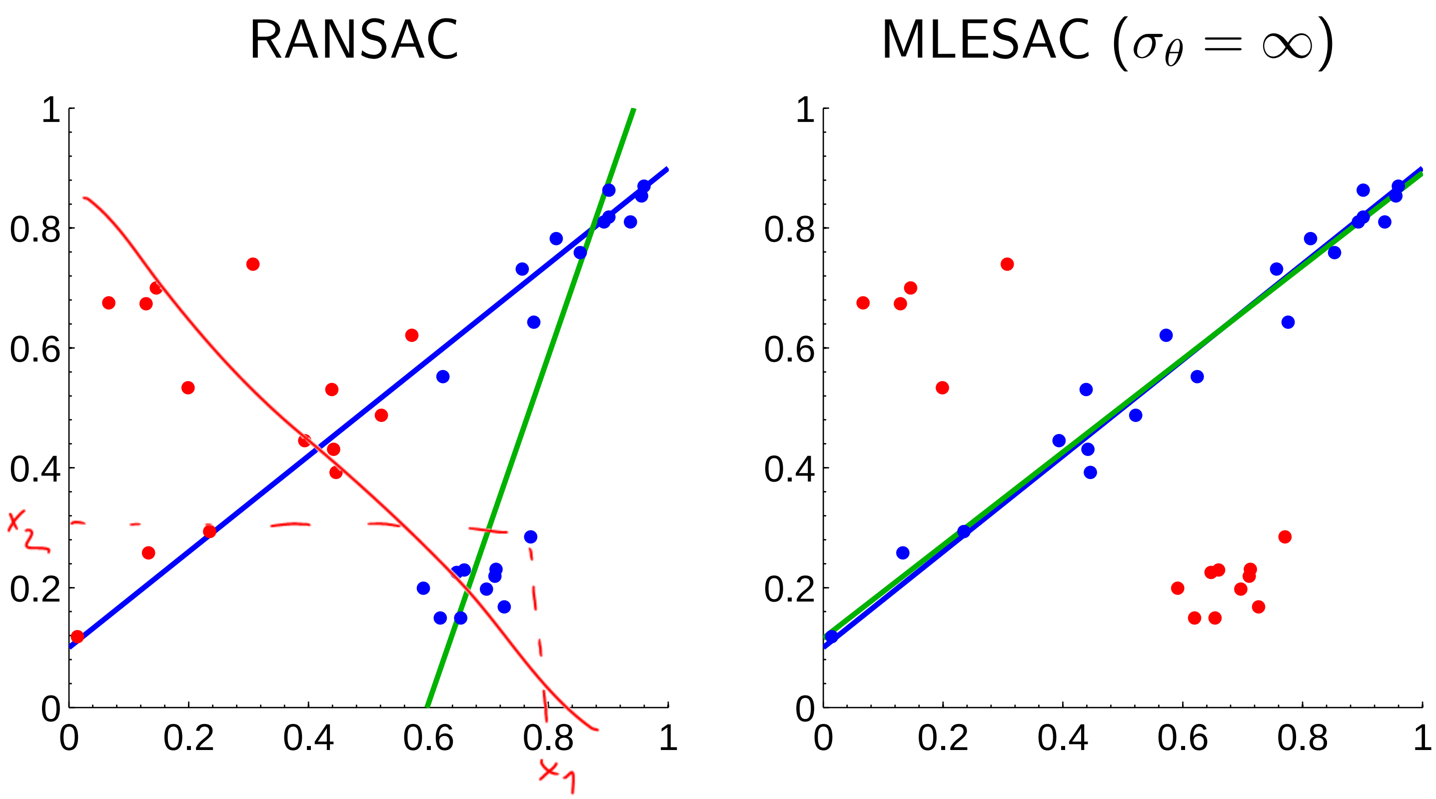
- dokáže lépe rozlišit nerozhodnutelné případy pokud je σ velké
- lze přidat apriorní model θ_0 dovoluje i menší počet bodů ve výběru než $s = |\theta|$

$$E = \sum_i \rho(e_i) + \frac{\|\theta - \theta_0\|^2}{\sigma_\theta^2}$$

- dostaneme randomizovaný algoritmus nelineární optimalizace E MAPSAC

Stabilita MLESACu

- randomizovaný algoritmus je stabilní když při opakovaném běhu vrací podobný výsledek



celkem 20 kontaminujících bodů (tj 50% všech bodů), použito $\sigma = 0.12$

- existují tři možné modely s 20 a více body, RANSAC je vrací všechny tři
- MLESAC vrací spravný model mnohem častěji
- při zařazení apriorního modelu θ_0 se stabilita ještě zlepší

Nejmenší medián kvadratické chyby, LMS

- podobná randomizovaná metoda — někdy označována LMedS
- nevyžaduje σ, ale vyžaduje to, aby správnému modelu vyhovovalo alespoň 50% bodů

Dáno:

$$n \geq \frac{\log(1-p)}{\log(1-0.5^s)}$$

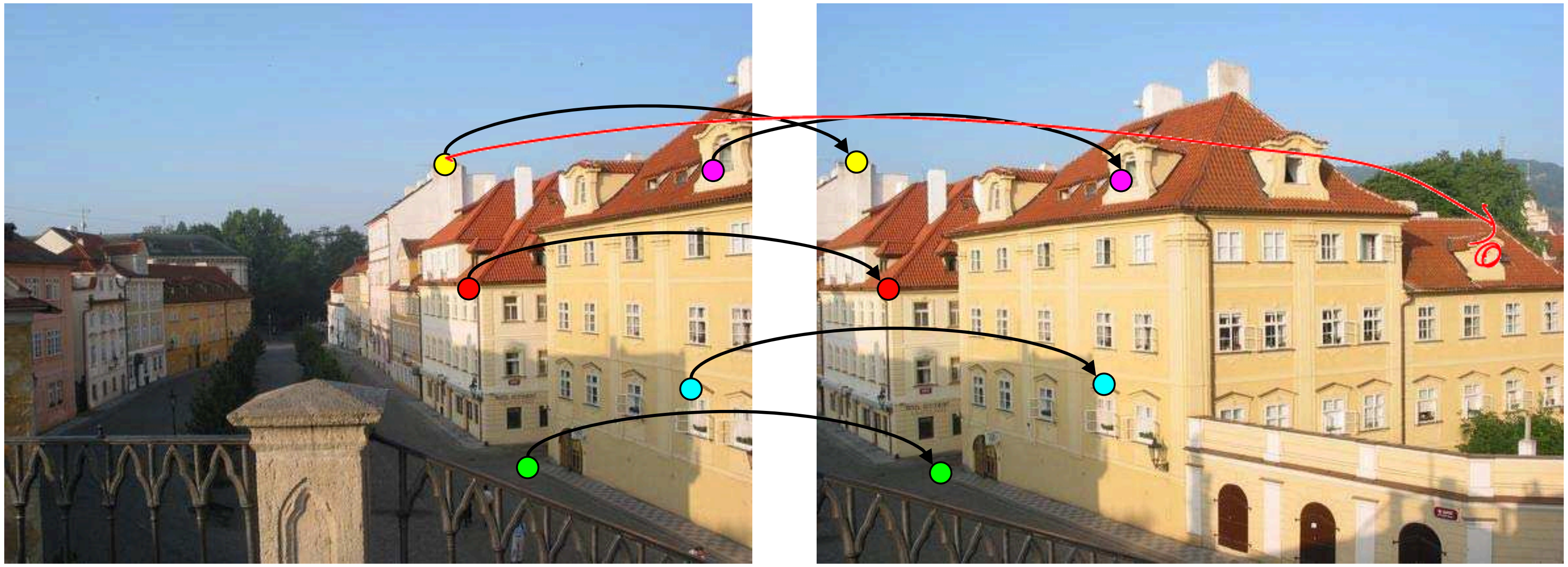
1. množina bodů $P = \{\mathbf{x}_i \mid i = 1, 2, \dots, k\}$
2. počet pokusů n

Procedura:

1. Inicializuj $M := \infty$.
2. Opakuj pro $j = 1, 2, \dots, n$:
   a. Náhodně vyber dvojici bodů $\mathbf{x}_1$, $\mathbf{x}_2$ z P
   b. Z $\mathbf{x}_1$, $\mathbf{x}_2$ vypočti parametry přímky $\theta_j := (\mathbf{a}_j, b_j)$
   c. Vypočti vzdálenost $e_1, e_2, \dots e_k$ všech bodů $\mathbf{x}_i$ vzhledem k $(\mathbf{a}_j, b_j)$
   d. Jestliže $\operatorname*{med}_i e_i^2 < M$ potom
      i. $\theta := \theta_j$
      ii. $M := \operatorname*{med}_i e_i^2$

Složitější problémy s latentními parametry

Příklad: nalezení korespondencí mezi obrazy



- binární indikátor pro každý bod: regulérní/kontaminující bod
- θ_n – nelatentní parametry neznámé parametry kamer: poloha, ohnisková vzdálenost,. . .
- θ_l – latentní parametry který regulérní bod v 2. obrazu patří ke každému bodu v 1. obrazu

Všechny korespondující body $\mathbf{x}_i \leftrightarrow \mathbf{y}_i$ musí vyhovovat rovnici

$$\tilde{\mathbf{y}}_i^\top \mathbf{F}\, \tilde{\mathbf{x}}_i = 0,$$ podmínka neurčuje jednoznačně pozici $\mathbf{y}_i$

kde $\mathbf{F}$ je neznámá 3×3 matice hodnosti 2 $\theta_n = \mathbf{F}$

- k odhadu $\mathbf{F}$ postačuje 8 korespondencí $\mathbf{x}_i \leftrightarrow \mathbf{y}_i \Rightarrow s = 8$