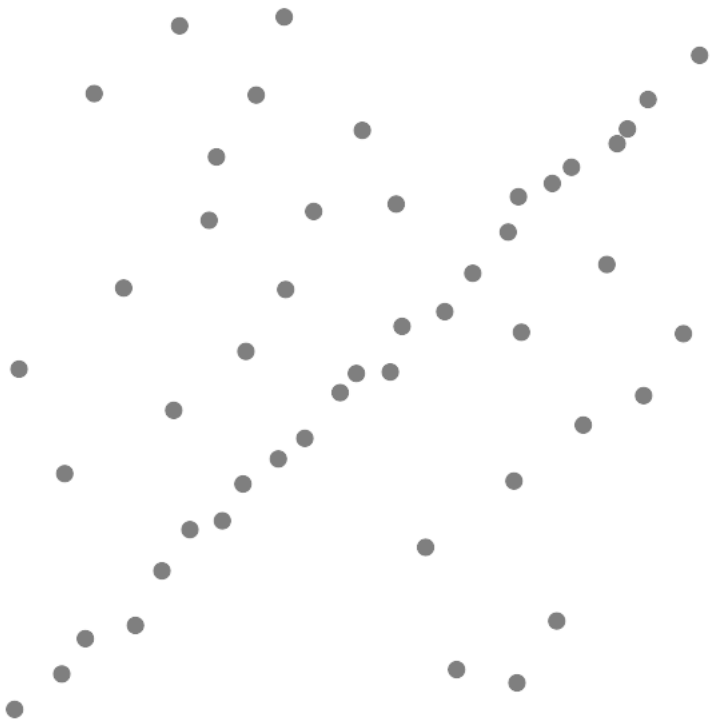
ROBUSZTUS BECSLÉSI MÓDSZEREK

1. Bevezetés

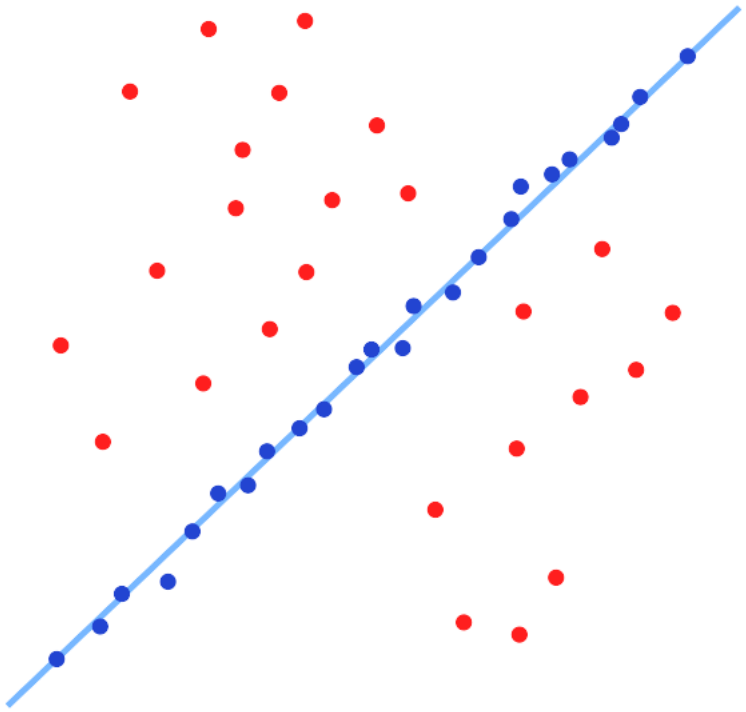
Becslésekkel már korábban is foglalkoztunk, tudjuk, hogy számtalan matematikailag megalapozott becslési módszer létezik. Ezek közül is kiemelkedik a legkisebb négyzetes módszer, mely a legegyszerűbb becslési eljárás, ráadásul lineáris esetben optimális becslést tudunk produkálni.

A gondok akkor kezdődnek, amikor hibás adatok kerülnek a ponthalmazainkba: hibás pontokból a hagyományos eljárások csak nagyon rossz becslést képesek produkálni.

A probléma súlyosságát egy konkrét példával szemléltetjük: adott n darab kétdimenziós pont, amelyre szeretnénk egyeneseket illeszteni. A pontok így néznek ki:



A sok véletlen pont között felsejlik középen egy egyenes képe, amely dominál, de rengeteg külső pont is található. A külső pontokat outlier-eknek, a modellhez tartozó pontokat inliereknek szokás hívni. Érezhető, hogy ezt az eredményt szeretnénk kapni:



2. Monte-Carlo elvű robusztus módszerek

Ezek a robusztus módszereket azért nevezték el a szerencsejátékok fővárosáról, Monte-Carlóról, mert véletlenszerű kiválasztáson alapulnak: néhány mintából próbálják meg a szükséges modelleket elképzelni. Ezt sokszor végzi el, és végül az így kapott modellek közül azt választja ki, amelyik a legjobban megfelel bizonyos kritériumoknak. A kaszinók rulettjét kell elképzelni, ahol számokat sorsolunk ki, és nem fontos, hogy minden egyes pörgetéskor nyerjünk, elég, ha jókor megütjük a főnyereményt.

A robusztus algoritmusokat az alábbi két lépésre lehet bontani:

(1) Modellek alkotása Monte-Carlo elvű módszerek esetén véletlen pontok ismétlődő kiválasztásával.
(2) Legjobb modell kiválasztása
(3) Kiválasztott (legjobb) modellnek megfelelő pontok meghatározása.
(4) A modell újraszámítása a kiválasztott pontok segítségével.

Az egyes robusztus módszerek különféleképpen valósítják meg a két lépést, de általában minden egyes módszer ezt a két fő lépést tartalmazza. Itt két módszert vizsgálunk meg: az ún. RANSAC és az LMedS/LTS eljárásokat.

A vizsgálat előtt azonban még meg kell állapítanunk, hogy hányszor kell a véletlen ponkiválasztást megismételni. Tegyük fel, hogy p darab pontból szeretnénk a modellt kiszámolni, az outlierek aránya pedig legyen κ. Az inlierek aránya ekkor értelemszerűen $1-\kappa$. Annak a valószínűsége, hogy mind a p darab pont inlier legyen $(1-\kappa)^p$. A véletlen pontkiválasztást ismételjük meg m-szer. Annak a valószínűsége,

hogy egyik modellalkotás sem jár sikerrel, azaz a p darab pontba minden esetben legalább egy outlier keveredik, felírható az alábbi összefüggés segítségével:

$$1 - \Gamma = (1 - (1 - \kappa)^p)^m$$

Értelemszerűen Γ jelöli annak a valószínűségét, hogy legalább egy darab jó mintánk születik, azaz $\Gamma = 1 - (1 - (1 - \kappa)^p)^m$.

A véletlen kiválasztás számát pedig m kifejtésével kaphatjuk meg:

$$m = \frac{ln(1 - \Gamma)}{ln[1 - (1 - \kappa)^p]}$$

Ennél a képletnél beszédesebb, ha kiszámoljuk néhány konkrét értékre a szükséges mintaszámot. Legyen 95% annak a valószínűsége, hogy jó modellt kapunk legalább egyszer. Az outlierek arányát és a szükséges kiválasztási számat az alábbi táblázattal szemléltethetjük:

- Ha egyenest szeretnénk illeszteni, két pont szükséges, tehát $p = 2$

$Outlier\%$	5	10	20	30	40	50	60	70	80
m	2	2	3	5	7	11	18	32	74

- $Hap = 3$ (pontregisztregisztrációs probléma)

$Outlier\%$	5	10	20	30	40	50	60	70	80
m	2	3	5	8	13	23	46	110	373

- Ha $p = 4$ (gyengén perspektív rekonstrukció)

$Outlier\%$	5	10	20	30	40	50	60	70	80
m	2	3	6	11	22	47	116	369	1871

- Ha $p = 7$ (7 potntos sztereó)

$Outlier\%$	5	10	20	30	40	50	60	70	80
m	3	5	13	35	106	382	1827	13692	233963

Jól látható, hogy az outlierek növekedésével nagyon romlik a szükséges műveletszám, éz ezáltal drasztikusan nő a futási idő. Ezért sebesség szempontjából nagyon fontos, hogy minél kevesebb outlier keveredjen a mintáinkba.

2.1. **RANSAC (RANdom SAmpling Consensus).** A RANSAC algoritmus talán a legnépszerűbb robusztus módszer.

Először is szükségünk van egy modellre, amely a rendelkezésre álló pontokból előállítható. Például egyenes illesztése esetén az egyenest leíró modell két valós paraméter, a és b, hiszen az egyenes $y = ax + b$ alakban adható meg. Két pont meghatároz egy egyenest.

A RANSAC módszer lényege, hogy a lehető legkevesebb pontból meghatározza a modellt, és utána megnézi, hogy mely pontok illeszkednek a modellre. Az egyenes illesztés esetén tehát meghatározzuk a és b paramétereket, majd az i-edik pont koordinátáit behelyettesítve megkapjuk a hibát a pontra:

$$\epsilon_i = y_i - b - ax_i$$

Ezek után számoljuk össze azokat a pontokat, amelyek megadott küszöbön (küszöböt jelöljük $\epsilon_{thr} - rel$. A thr rövidítés utal a küszöb angol nevére, a threshold-ra) belül

vannak, akkor megkapjuk, hogy hány pont támogatja a modellt (latin szóval konszenzust alkot a modellel). Ezek után újabb két pontot választunk, és újra egy egyenest húzunk, ahol a távolságokat és a küszöbön belül levő pontokat újfent meghatározhatjuk.

A sok lehetséges modell közül azt választjuk ki, amelyikhez a legtöbb konszenzusos pont tartozik.

A RANSAC modell nagy előnye, hogy egyszerű, gyorsan implementálható. Hátránya, hogy egy küszöböt meg kell adni, és ez nem is olyan egyszerű feladat: ha a küszöb szigorú, sok jó pontot kidobunk, ha nagyon laza a küszöb, akkor outlierek is bekerülhetnek az adatainkba (és ezért rontják a végeredmény minőségét).

2.2. **LMedS (Least MEDian of Squares).** A Least MEDian of Squares segítségével a küszöb eltüntethető. Statisztikusok arra a megállapításra jutottak, hogy ha az egyes ϵ_i-k Gauss-eloszlást követnek, akkor az inlierek pozícióinak szórására robusztus becslést adhatunk. Ehhez be kell vezetni egy súlyozó számot:

$$s^0 = 1,4826\frac{5}{n-p}medin\{\epsilon_i\}$$

ahol n az összes pont száma, amiből p daraboit kell választani a modellépítéshez. Sokszor ismételjük a modellalkotást, véletlenszerűen kiválasztunk p darab pontot. Minden esetben kiszámoljuk s^0-t. Végül azt választjuk ki, amelyik s^0a legkisebb. Ezzel a helyes modellt megbecsültük. Az lett a helyes modell, amelyiknél a medián (azaz a sorba rakott $\epsilon_i - k$ közül a középső) a legkisebb.

Ha megvan a minimális s^0, akkor meg lehet becsülni a valódi szórást az alábbi összefüggés segítségével:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(w_i\epsilon_i^2)}{\sum_{i=1}^{N} w_i - 4}}$$

ahol N jelöli a minták számát, w_i pedig egy bináris változó: ha $e_i < 2,5s^0$, akkor $w_i = 1$, egyébként nulla.

Végezetül meg kell határozni a becsült helyes modellhez tartozó pontokat. Azok a pontokat jelöljük inliereknek (modellhez tartozónak), amelyekhez tartozó ϵ_i-k 2.5σ-nál kisebbek.

2.2.1. *LTS (Least Trimmed Squares).* Az LMedS módszernek az a hátránya, hogy nem működik, ha az outlierek aránya ötven százalék fölé megy, hiszen a medián akkor outlierből származó hibára fog mutatni helyes modell esetén is. Ezért egy apró változtatást végeztek a kutatók: a σ^0-t meghatározó összefüggésben lecserélték a mediánt az első d darab legjobb hiba összegére. Az a szerencsés választás, ha d a várható outlier aránynál nem sokkal kisebb (de mindenképpen kisebb!).