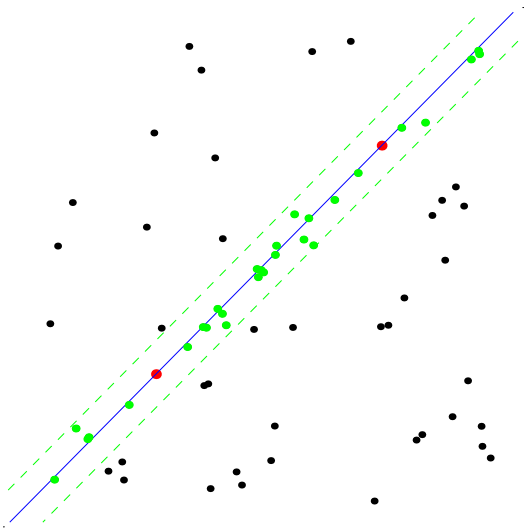
Probability of selecting uncontaminated sample in K trials



Uncontaminated sample

- N - number of data points
- w - fraction of inliers
- s - size of the sample

Prob. of selecting a sample with all inliers[3]: $\approx w^s$

Prob. of not selecting a sample with all inliers: $1 - w^s$

Prob. of not selecting a good sample K times: $(1 - w^s)^K$

The sought probability of selecting uncontaminated sample in K trials at least once: $P = 1 - (1 - w^s)^K$

[3] Approximation valid for $s \ll N$, see the lecture notes

Derivation of the probability

Drawing samples from a pool of data *without replacement* is not an independent process, it is not like tossing a coin. Assume N data points, I of them inliers. Probability of selecting one inlier in a random trial is clearly I/N. Probability of selecting another inlier from the remaining data set is $(I-1)/(N-1)$. The probability of selecting s uncontaminated (= all inliers) samples is then

$$P_s(I) = \prod_{i=0}^{s-1} \frac{I-i}{N-i}$$

which for small sample size and large number of points $s \ll N$ may be well approximated as w^s.

End

References