随机抽样一致算法的讲解以及应用

(random sample consensus)

colorful

讲解内容

1、最小二乘法

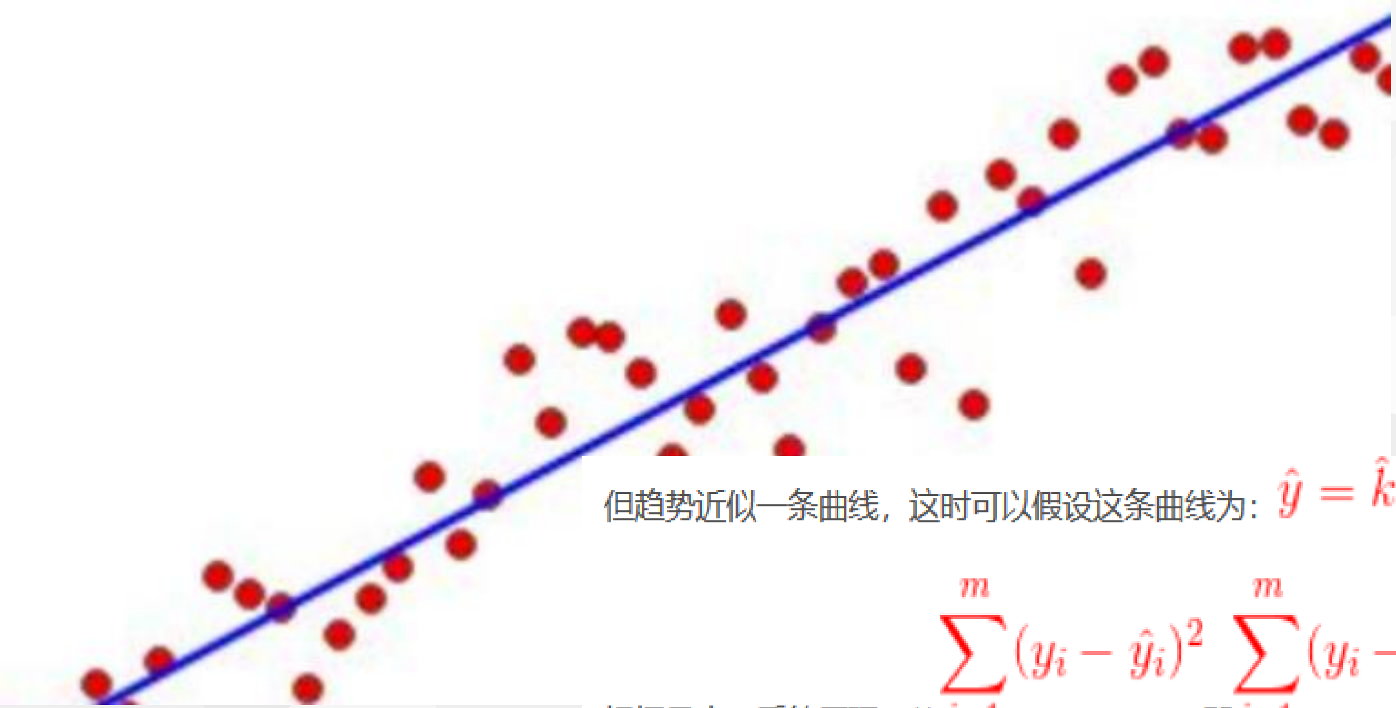
2、原论文内容

3、算法讲解

4、算法应用

• 最小二乘法的复习

最小二乘法：有一种直线拟合的方式。它是一种数学优化技术，原理是通过**最小化误差的平方和**寻找数据的最佳模型。



比如研究x和y之间的关系，假设我们拥有的数据是：

$(x_1,y_1),(x_2,y_2),(x_3,y_3),(x_4,y_4),...,(x_{m-1},y_{m-1}),(x_m,y_m)$

将这些数据描绘在x-y直角坐标系中，发现这些点并没有能够连接成一条直线。

但趋势近似一条曲线，这时可以假设这条曲线为：$\hat{y}=\hat{k}x+b$。

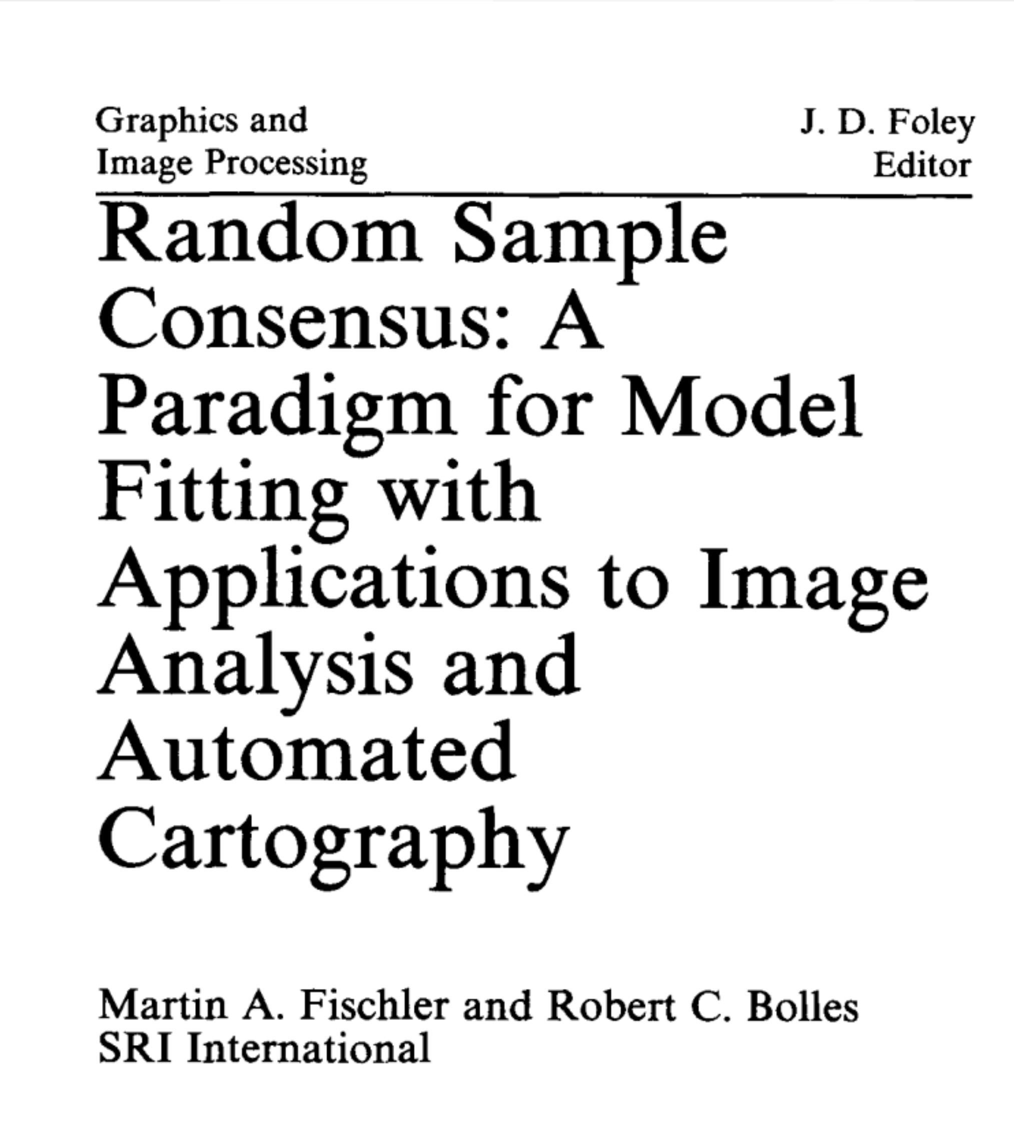
根据最小二乘的原理，使 $\sum_{i=1}^{m}(y_i-\hat{y}_i)^2$ 即 $\sum_{i=1}^{m}(y_i-\hat{k}x_i-b)^2$ 最小化,可以得到$\hat{k}$值，再根据直线过点$(\bar{x},\bar{y})$得出b的值。$\bar{x}$为横坐标的平均值，$\bar{y}$为纵坐标的平均值。

其中，

$$\hat{k}=\frac{\sum_{i=1}^{m}xy-\frac{1}{m}\sum_{i=1}^{m}x*\sum_{i=1}^{m}y}{\sum_{i=1}^{m}x^2-\frac{1}{m}(\sum_{i=1}^{m}x)^2}$$

，$b=\hat{y}-\hat{k}x$。

• 1、原论文内容

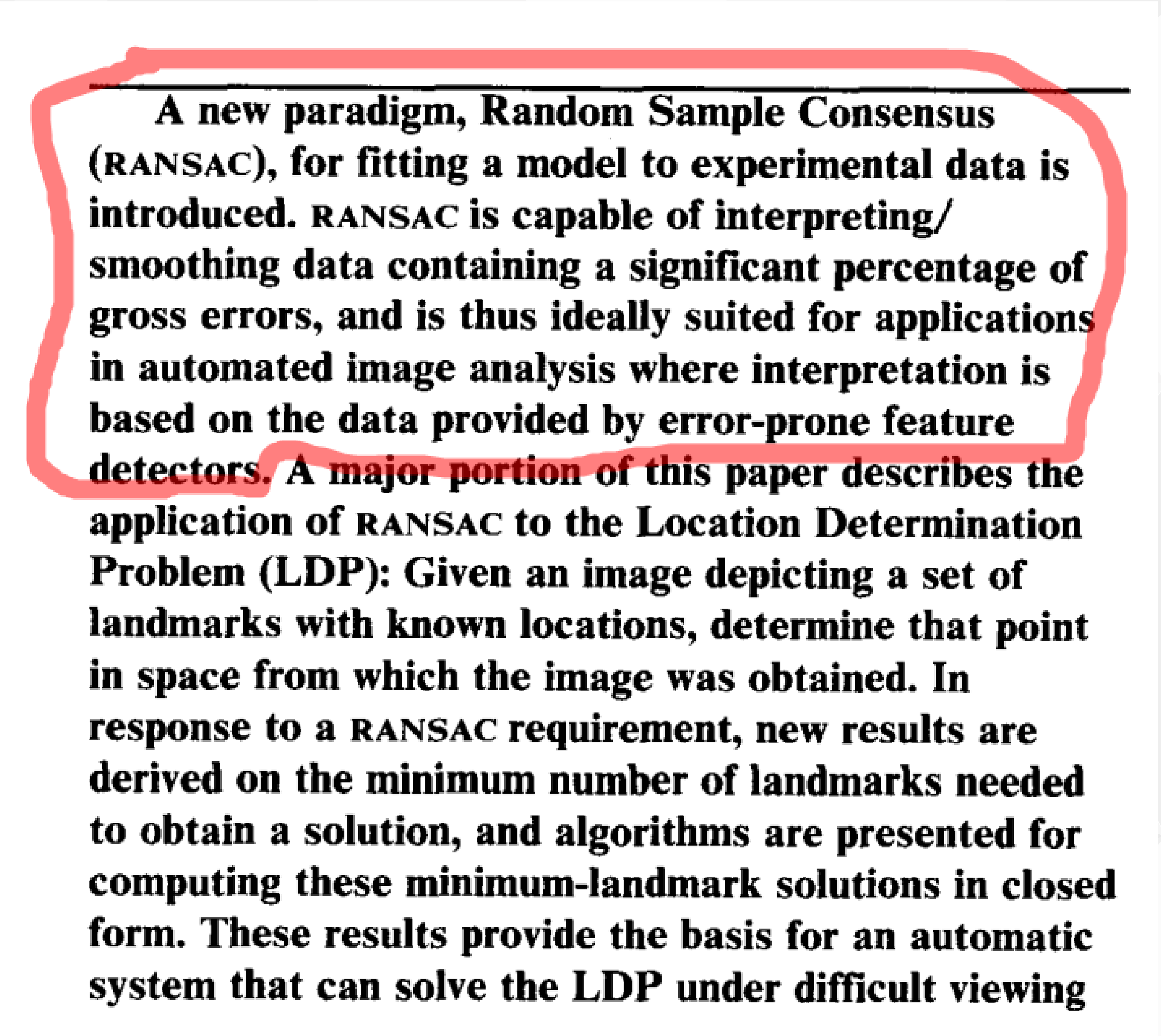


Graphics and
Image Processing

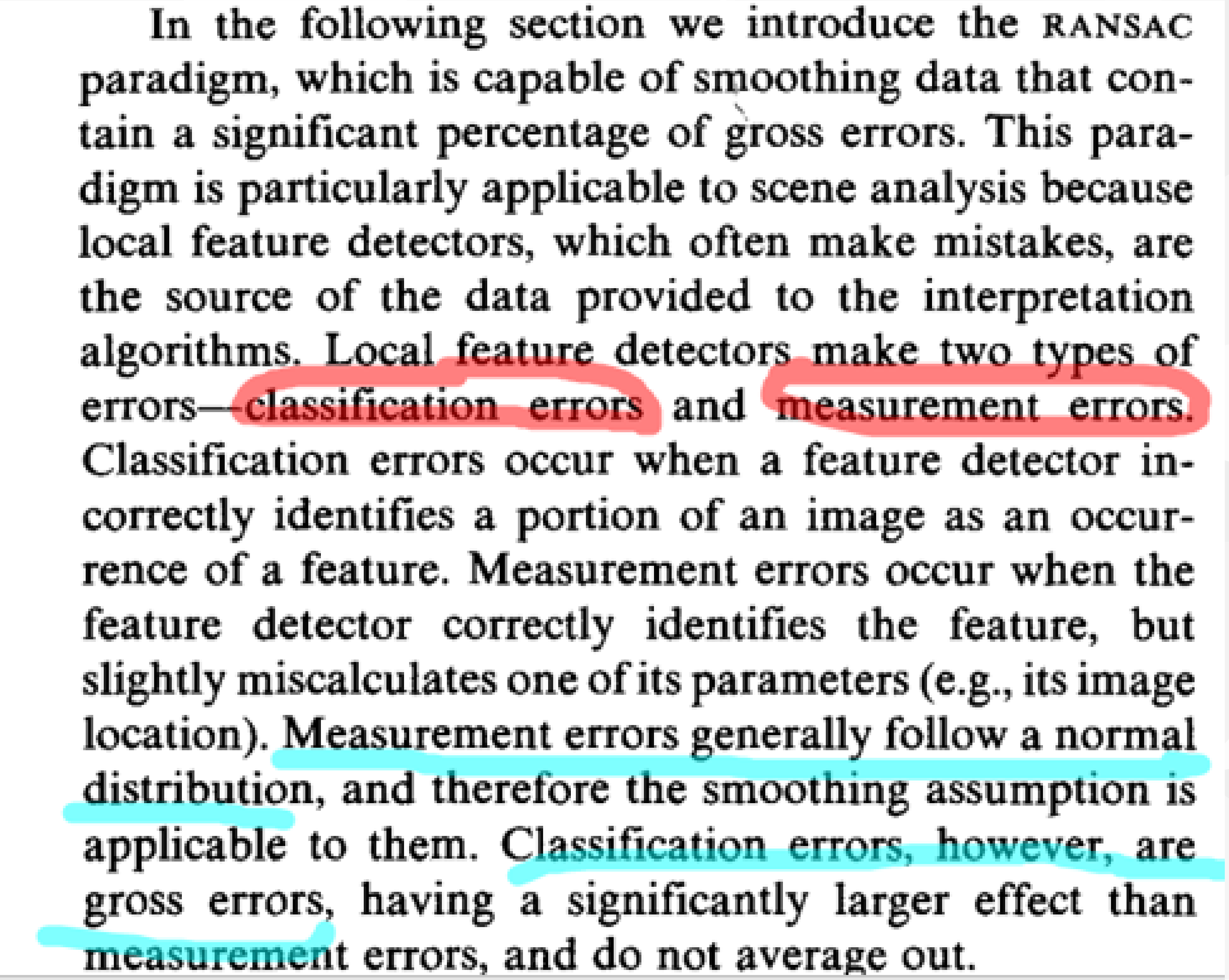
J. D. Foley
Editor

Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography

Martin A. Fischler and Robert C. Bolles
SRI International



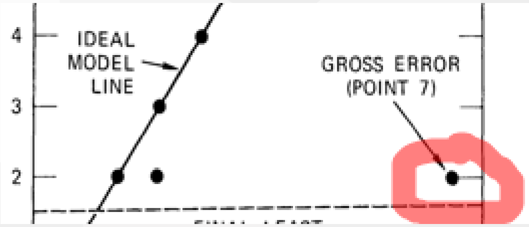
A new paradigm, Random Sample Consensus (RANSAC), for fitting a model to experimental data is introduced. RANSAC is capable of interpreting/smoothing data containing a significant percentage of gross errors, and is thus ideally suited for applications in automated image analysis where interpretation is based on the data provided by error-prone feature detectors. A major portion of this paper describes the application of RANSAC to the Location Determination Problem (LDP): Given an image depicting a set of landmarks with known locations, determine that point in space from which the image was obtained. In response to a RANSAC requirement, new results are derived on the minimum number of landmarks needed to obtain a solution, and algorithms are presented for computing these minimum-landmark solutions in closed form. These results provide the basis for an automatic system that can solve the LDP under difficult viewing



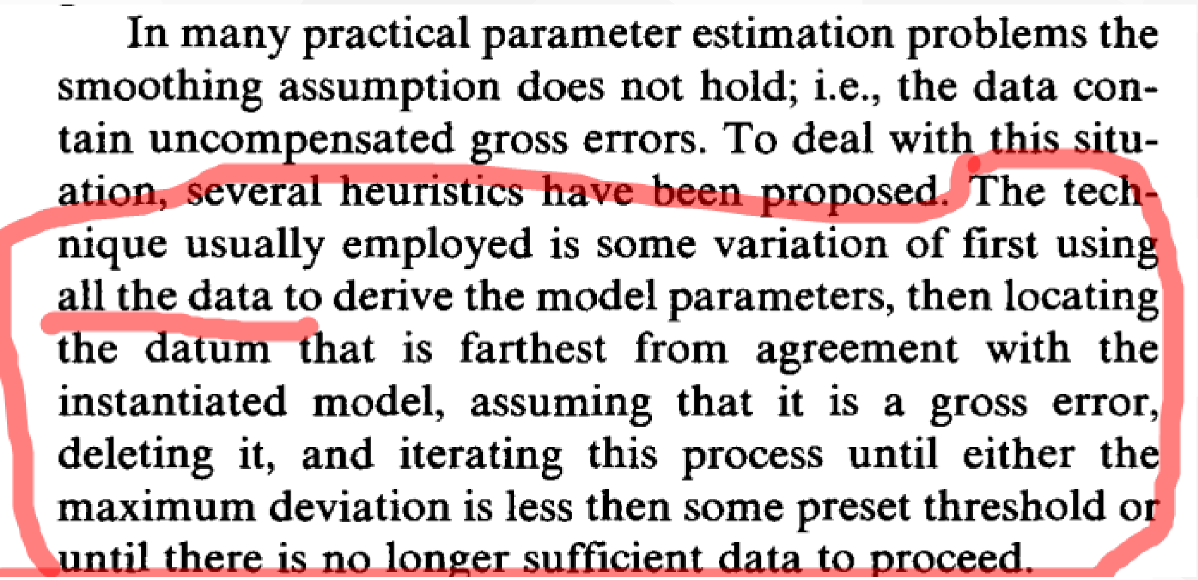
In the following section we introduce the RANSAC paradigm, which is capable of smoothing data that contain a significant percentage of gross errors. This paradigm is particularly applicable to scene analysis because local feature detectors, which often make mistakes, are the source of the data provided to the interpretation algorithms. Local feature detectors make two types of errors—classification errors and measurement errors. Classification errors occur when a feature detector incorrectly identifies a portion of an image as an occurrence of a feature. Measurement errors occur when the feature detector correctly identifies the feature, but slightly miscalculates one of its parameters (e.g., its image location). Measurement errors generally follow a normal distribution, and therefore the smoothing assumption is applicable to them. Classification errors, however, are gross errors, having a significantly larger effect than measurement errors, and do not average out.

测量错误，局部特征检测器将特征点的数据轻微地计算错误。这种误差通常较小且符合正态分布，

分类错误，局部特征检测器，将本不是特征点当作是了特征点，这个就是重大误差点，如图



• 最小二乘法拟合模型



In many practical parameter estimation problems the smoothing assumption does not hold; i.e., the data contain uncompensated gross errors. To deal with this situation, several heuristics have been proposed. The technique usually employed is some variation of first using all the data to derive the model parameters, then locating the datum that is farthest from agreement with the instantiated model, assuming that it is a gross error, deleting it, and iterating this process until either the maximum deviation is less then some preset threshold or until there is no longer sufficient data to proceed.

最小二乘法：**使用尽可能多得数据去拟合出一个模型，使得尽可能多的点在模型当中，期望通过"平均"消除偏差**，然而在很多实际情况下（数据中带有重大误差点），不能够得到一个好的结果。

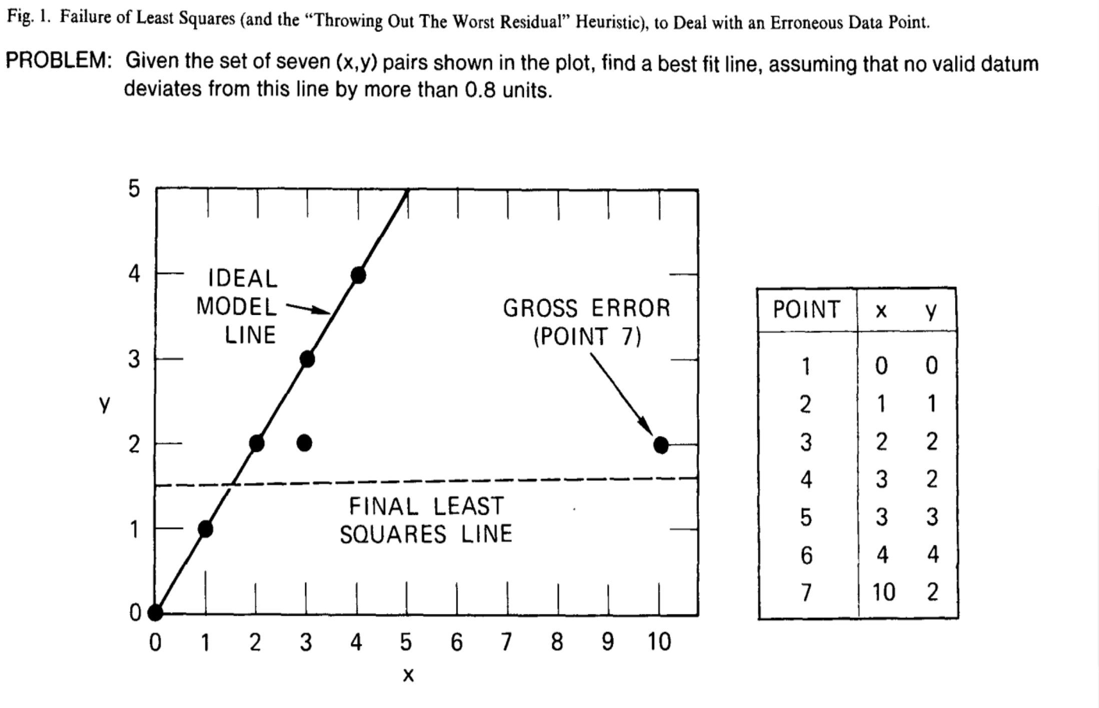
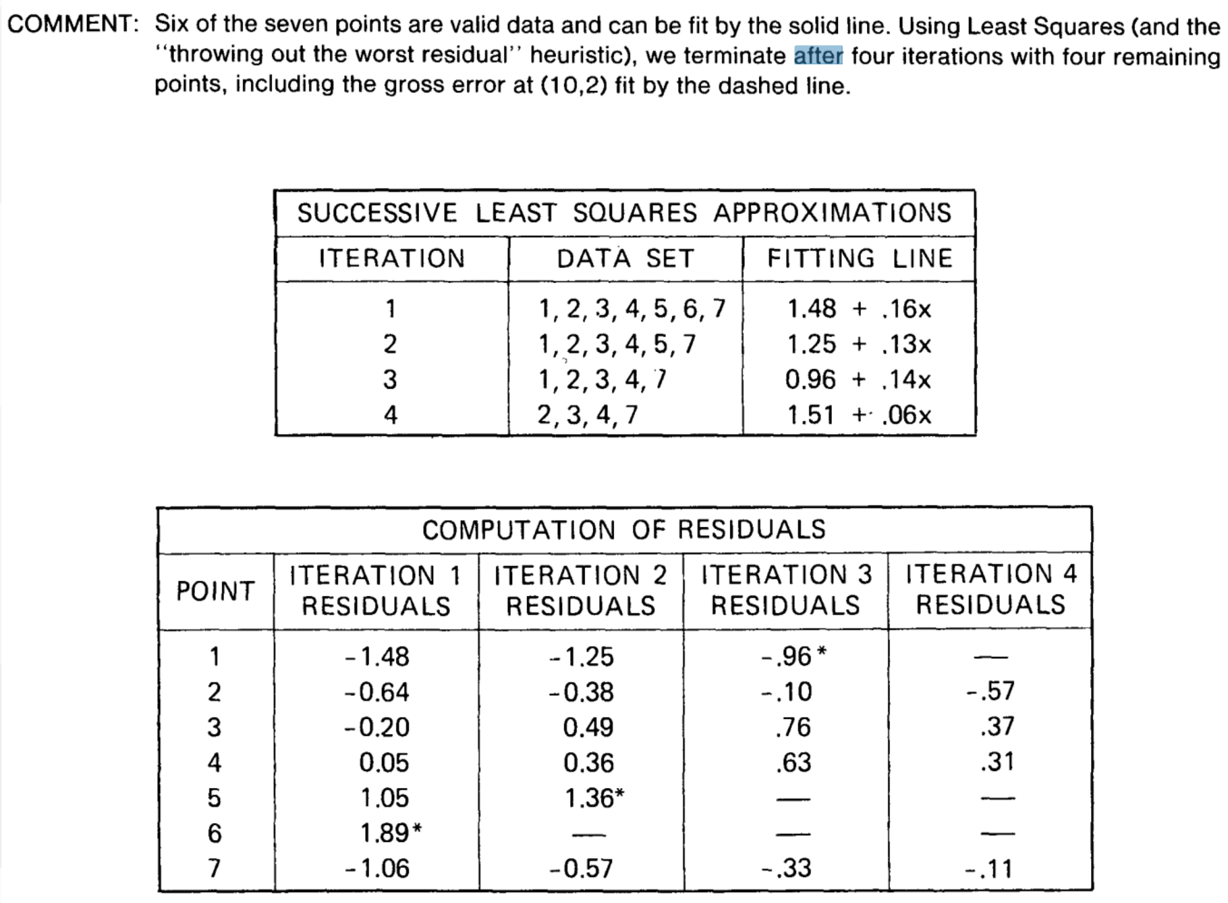


Fig. 1. Failure of Least Squares (and the "Throwing Out The Worst Residual" Heuristic), to Deal with an Erroneous Data Point.

PROBLEM: Given the set of seven (x,y) pairs shown in the plot, find a best fit line, assuming that no valid datum deviates from this line by more than 0.8 units.

POINT	x	y
1	0	0
2	1	1
3	2	2
4	3	2
5	3	3
6	4	4
7	10	2



COMMENT: Six of the seven points are valid data and can be fit by the solid line. Using Least Squares (and the "throwing out the worst residual" heuristic), we terminate after four iterations with four remaining points, including the gross error at (10,2) fit by the dashed line.

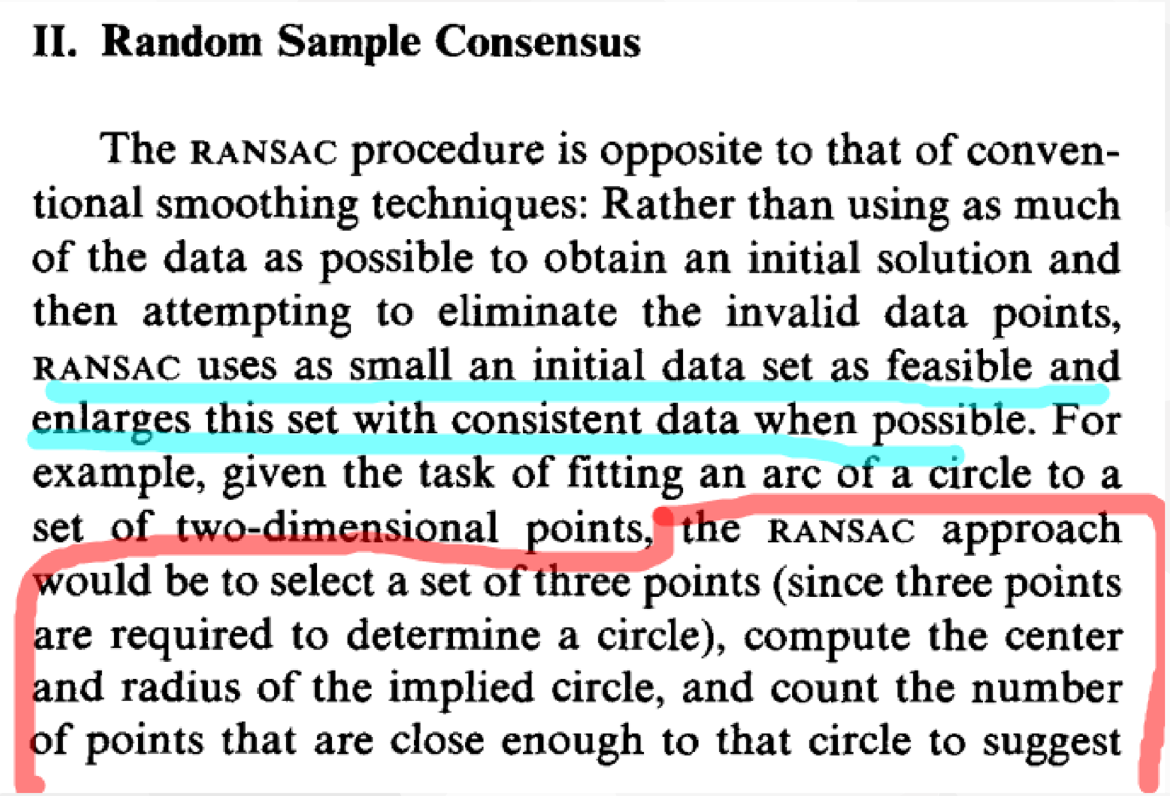
SUCCESSIVE LEAST SQUARES APPROXIMATIONS

ITERATION	DATA SET	FITTING LINE
1	1, 2, 3, 4, 5, 6, 7	1.48 + .16x
2	1, 2, 3, 4, 5, 7	1.25 + .13x
3	1, 2, 3, 4, 7	0.96 + .14x
4	2, 3, 4, 7	1.51 + .06x

COMPUTATION OF RESIDUALS

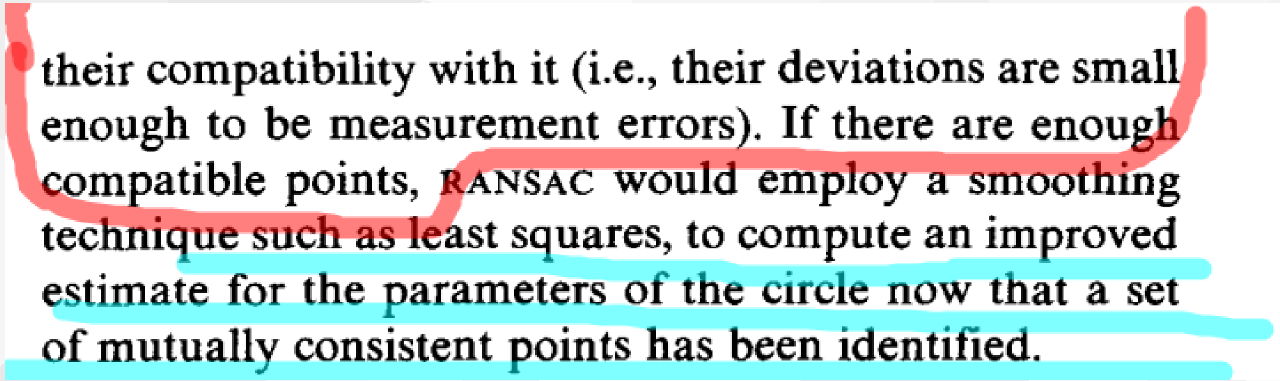
POINT	ITERATION 1 RESIDUALS	ITERATION 2 RESIDUALS	ITERATION 3 RESIDUALS	ITERATION 4 RESIDUALS
1	-1.48	-1.25	-.96*	—
2	-0.64	-0.38	-.10	-.57
3	-0.20	0.49	.76	.37
4	0.05	0.36	.63	.31
5	1.05	1.36*	—	—
6	1.89*	—	—	—
7	-1.06	-0.57	-.33	-.11

• RANSAC拟合模型过程简单介绍

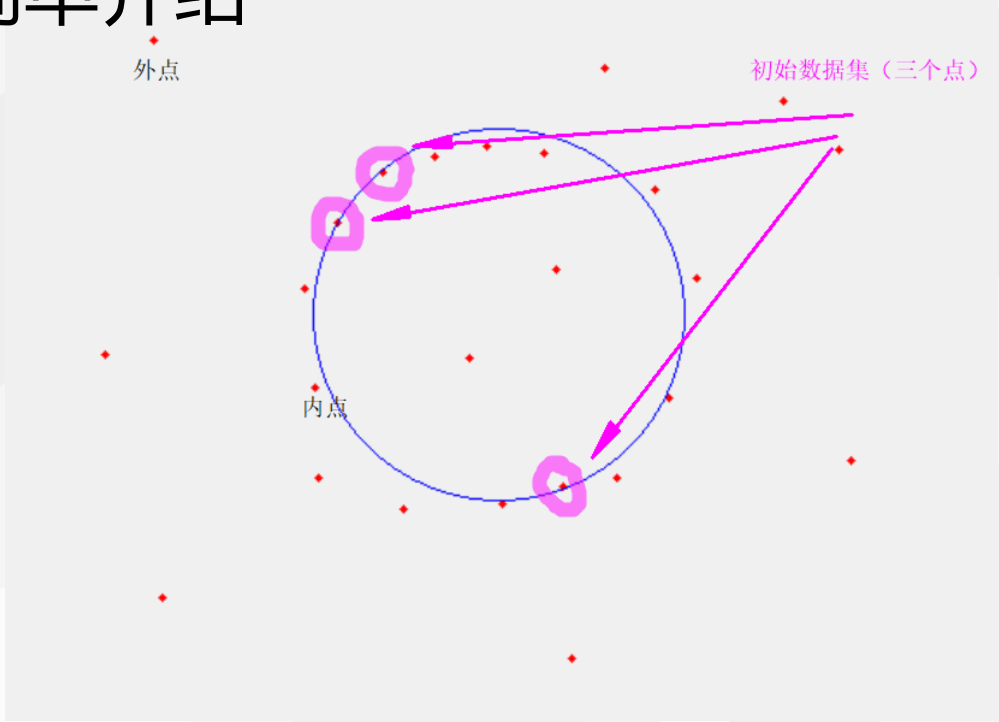


II. Random Sample Consensus

The RANSAC procedure is opposite to that of conventional smoothing techniques: Rather than using as much of the data as possible to obtain an initial solution and then attempting to eliminate the invalid data points, RANSAC uses as small an initial data set as feasible and enlarges this set with consistent data when possible. For example, given the task of fitting an arc of a circle to a set of two-dimensional points, the RANSAC approach would be to select a set of three points (since three points are required to determine a circle), compute the center and radius of the implied circle, and count the number of of points that are close enough to that circle to suggest



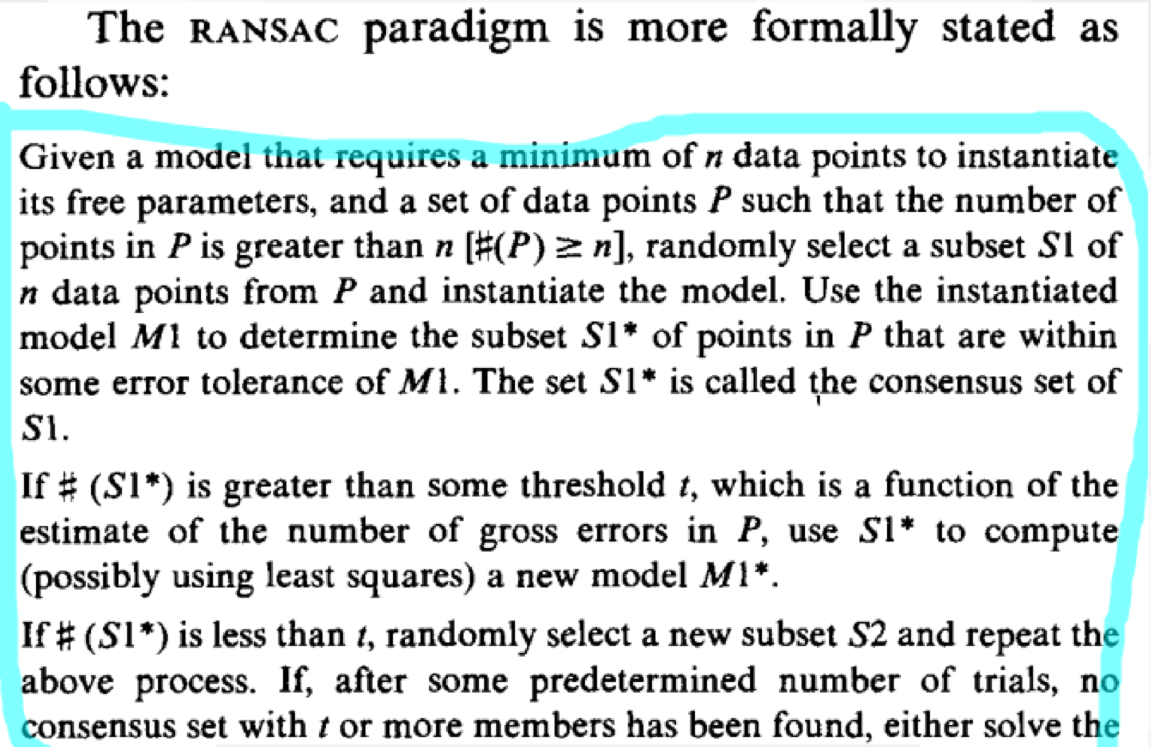
their compatibility with it (i.e., their deviations are small enough to be measurement errors). If there are enough compatible points, RANSAC would employ a smoothing technique such as least squares, to compute an improved estimate for the parameters of the circle now that a set of mutually consistent points has been identified.



RANSAC算法：使用尽可能少的初始数据去拟合出一个模型，通过迭代次数，逐渐扩大内点数量，找到一个更好的模型

比如说三个点就足够定义一个圆，一开始我们可以从数据集中随机选取三个点，并假设连接它们的构成的圆就是正确的模型。具体算法看后面。

• RANSAC算法具体内容



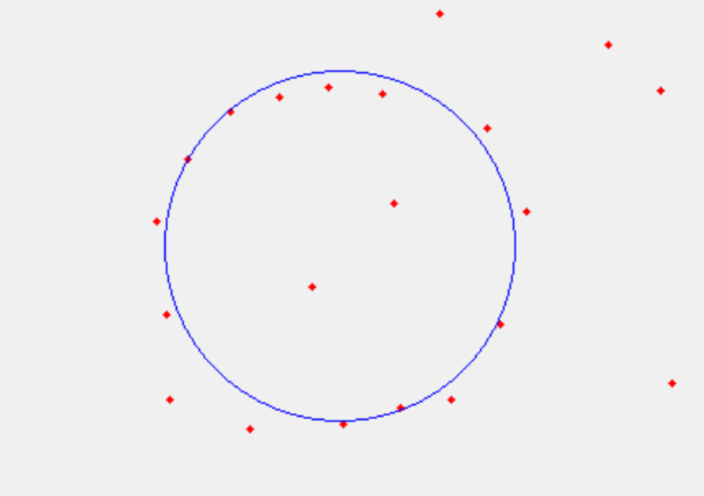
The RANSAC paradigm is more formally stated as follows:

Given a model that requires a minimum of n data points to instantiate its free parameters, and a set of data points P such that the number of points in P is greater than n [$\#(P) \geq n$], randomly select a subset $S1$ of n data points from P and instantiate the model. Use the instantiated model $M1$ to determine the subset $S1^*$ of points in P that are within some error tolerance of $M1$. The set $S1^*$ is called the consensus set of $S1$.

If $\#(S1^*)$ is greater than some threshold t, which is a function of the estimate of the number of gross errors in P, use $S1^*$ to compute (possibly using least squares) a new model $M1^*$.

If $\#(S1^*)$ is less than t, randomly select a new subset $S2$ and repeat the above process. If, after some predetermined number of trials, no consensus set with t or more members has been found, either solve the model with t

in failure.



The
paramete
whether
the num
which is
that the
discussed
rameters

unspecified
to determine
a model, (2)
threshold t,
used to imply
Methods are
for these pa-

P：为“局部特征检测器”检测获得的所有数据，包括重大误差点；

n：为构建一个模型所需的最少数据，例如一个圆需要3个点才能确定下来，那么n=3；

注意：RANSAC算法第一次拟合模型-圆时，用的数据只有n=3个，但是第二次、第三次.....拟合所用的数据会更多n>3。

S1：内点的集合；

M1：每次迭代后，拟合出的那个模型；

the error tolerance：容限误差，论文中没有提到用什么符号表示，我假设用“ε”(伊普西隆)表示。当一个点与模型的残差小于ε，那么我就判定该点为内点，否则为外点；

t：阈值，当内点的数量大于t时，判定拟合出的那个模型合理，否则不合理；

算法步骤:

1、假设我们要将P个数据点 {x1, x2,… , xn}拟合一个由至少n个点决定的模型(P≥n ,对于圆来说n=3)。

2、设迭代计数k=1。

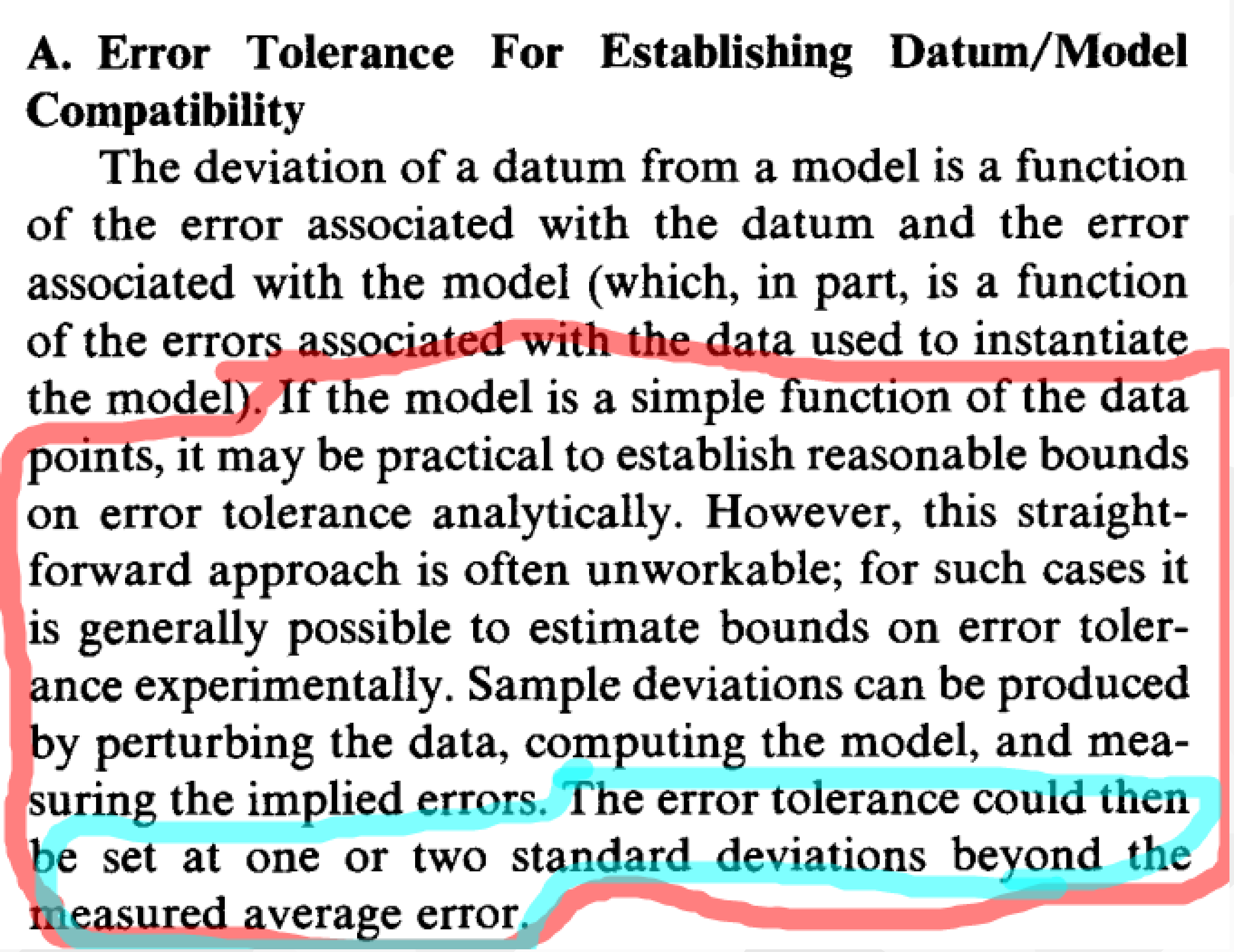
3、从P中随机选取n个点拟合一个模型，记为M1。n在一开始为3，之后会越来越大。

4、给定容限误差 ε,计算数据点{x1, x2,… , xn}中相对于模型的残差在偏差ε内的元素个数，如果内点个数大于阈值t，算法终止。之后我们可以根据内点集合重新拟合模型(可以利用最小二乘或其变种)，。

5、设k=k+1,如果k小于预先设定的K,跳至第3步，新的内点集合和模型分别记为S1*和M1*。

否则采用具有当前内点最多的点集的模型，或者算法失败。

• 容限误差的确定



A. Error Tolerance For Establishing Datum/Model Compatibility

The deviation of a datum from a model is a function of the error associated with the datum and the error associated with the model (which, in part, is a function of the errors associated with the data used to instantiate the model). If the model is a simple function of the data points, it may be practical to establish reasonable bounds on error tolerance analytically. However, this straightforward approach is often unworkable; for such cases it is generally possible to estimate bounds on error tolerance experimentally. Sample deviations can be produced by perturbing the data, computing the model, and measuring the implied errors. The error tolerance could then be set at one or two standard deviations beyond the measured average error.

如果模型是一个简单函数，那么我们可以通过计算解析解的方式求得容限误差的值。
但是实际情况下，这种方式不可行。
在复杂的情况下，我可以通过经验确定容限误差的值。样本偏差可以通过以下几个方面确定：噪声、计算出的模型、测量隐藏误差。

容限误差=平均偏差+1~2个标准差

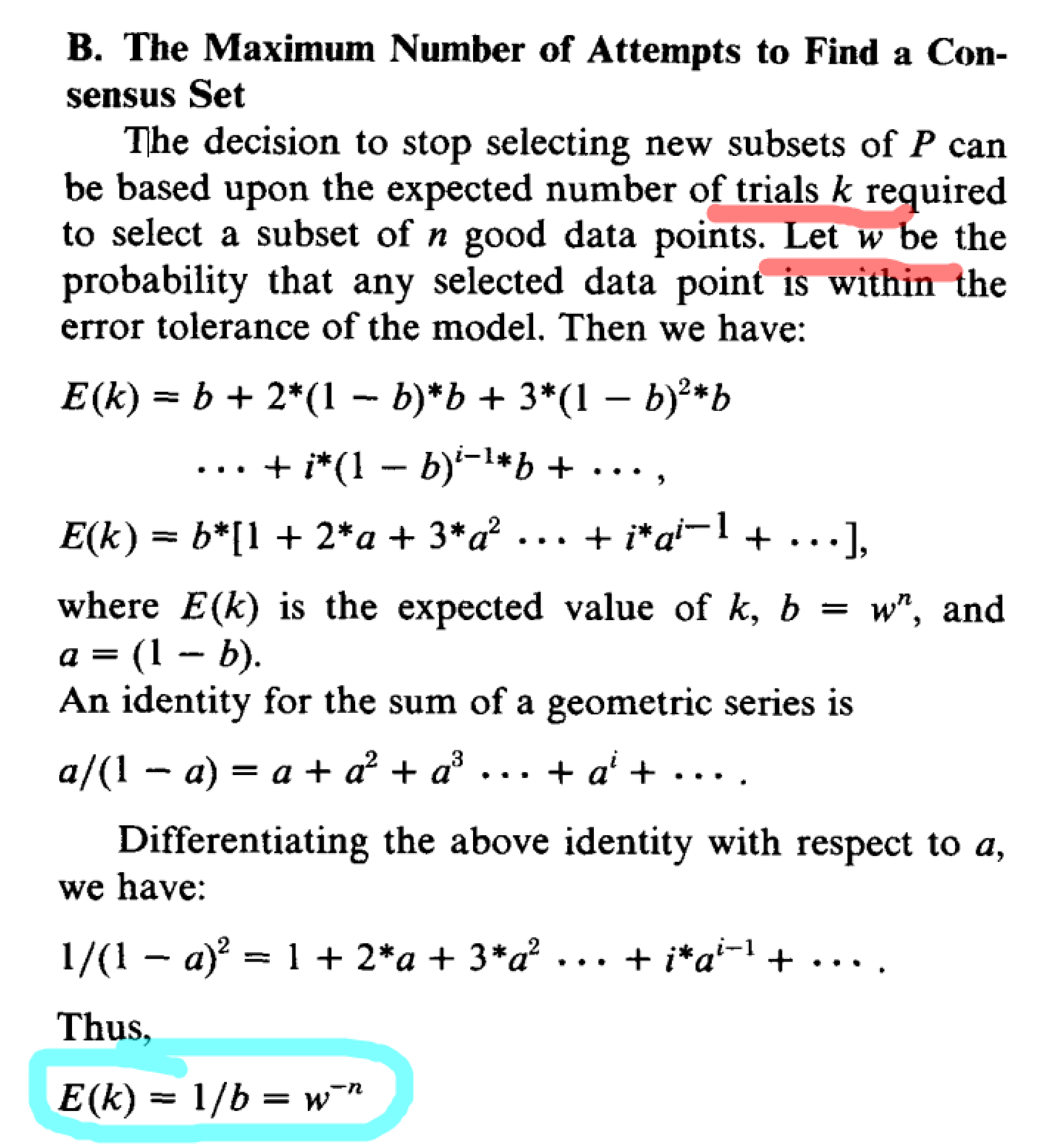
x^2=5

solution: x=sqrt(5) -- analytical solution（解析解，闭合解）

x=2.236 -- numerical solution（数值解）

SIFT算法中的子像元插值（SIFT关键点定位）就属于数值解

• 迭代次数k的确定



B. The Maximum Number of Attempts to Find a Consensus Set

The decision to stop selecting new subsets of *P* can be based upon the expected number of trials *k* required to select a subset of *n* good data points. Let *w* be the probability that any selected data point is within the error tolerance of the model. Then we have:

$$E(k) = b + 2{*}(1 - b){*}b + 3{*}(1 - b)^2{*}b$$
$$\cdots + i{*}(1 - b)^{i-1}{*}b + \cdots,$$

$$E(k) = b{*}[1 + 2{*}a + 3{*}a^2 \cdots + i{*}a^{i-1} + \cdots],$$

where $E(k)$ is the expected value of k, $b = w^n$, and $a = (1 - b)$.

An identity for the sum of a geometric series is

$$a/(1 - a) = a + a^2 + a^3 \cdots + a^i + \cdots.$$

Differentiating the above identity with respect to a, we have:

$$1/(1 - a)^2 = 1 + 2{*}a + 3{*}a^2 \cdots + i{*}a^{i-1} + \cdots.$$

Thus,

$$E(k) = 1/b = w^{-n}$$

k：寻找合适的拟合模型时需要算法的迭代次数；

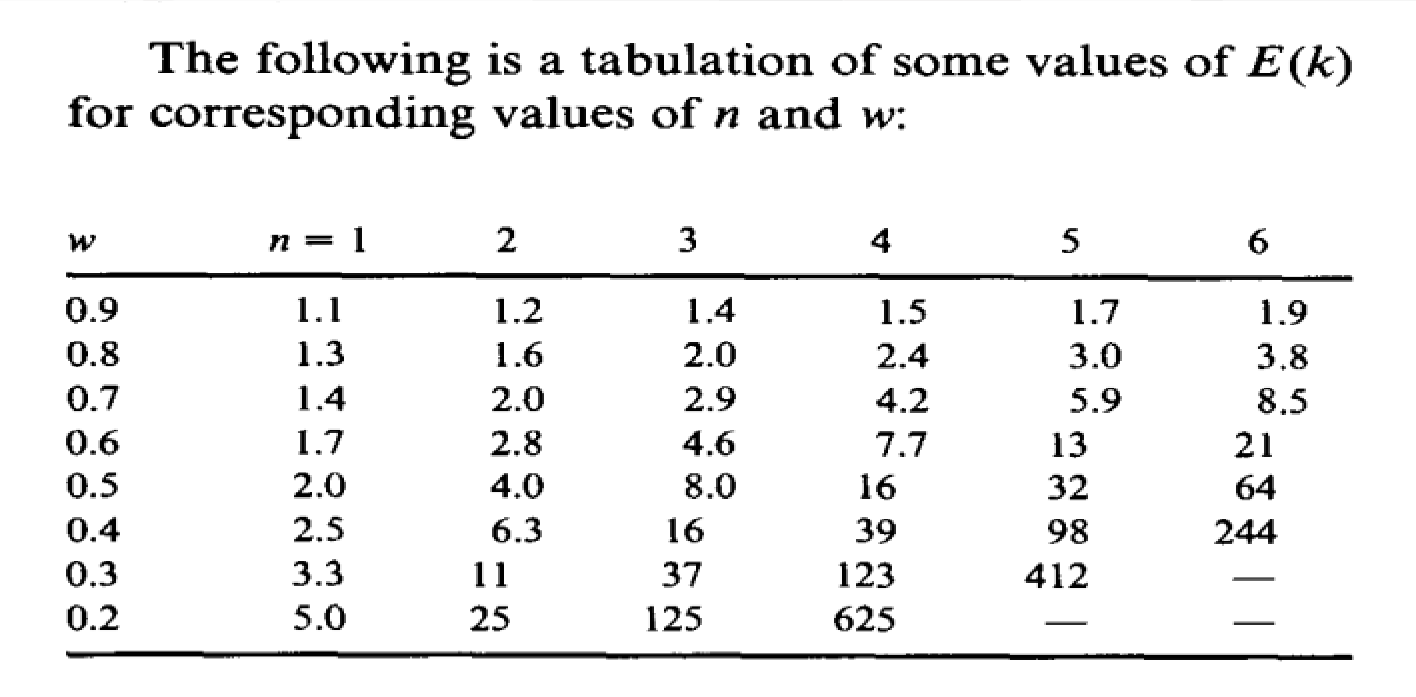
E(k)：k的期望值；

ω：从P个点中选出来的一点是符合容限误差的点的概率（ω 是一个先验概率，可以实现确定的）；

b：算法迭代中，某一次拟合模型所用的点，所有点(n个点)都是符合容限误差的概率；

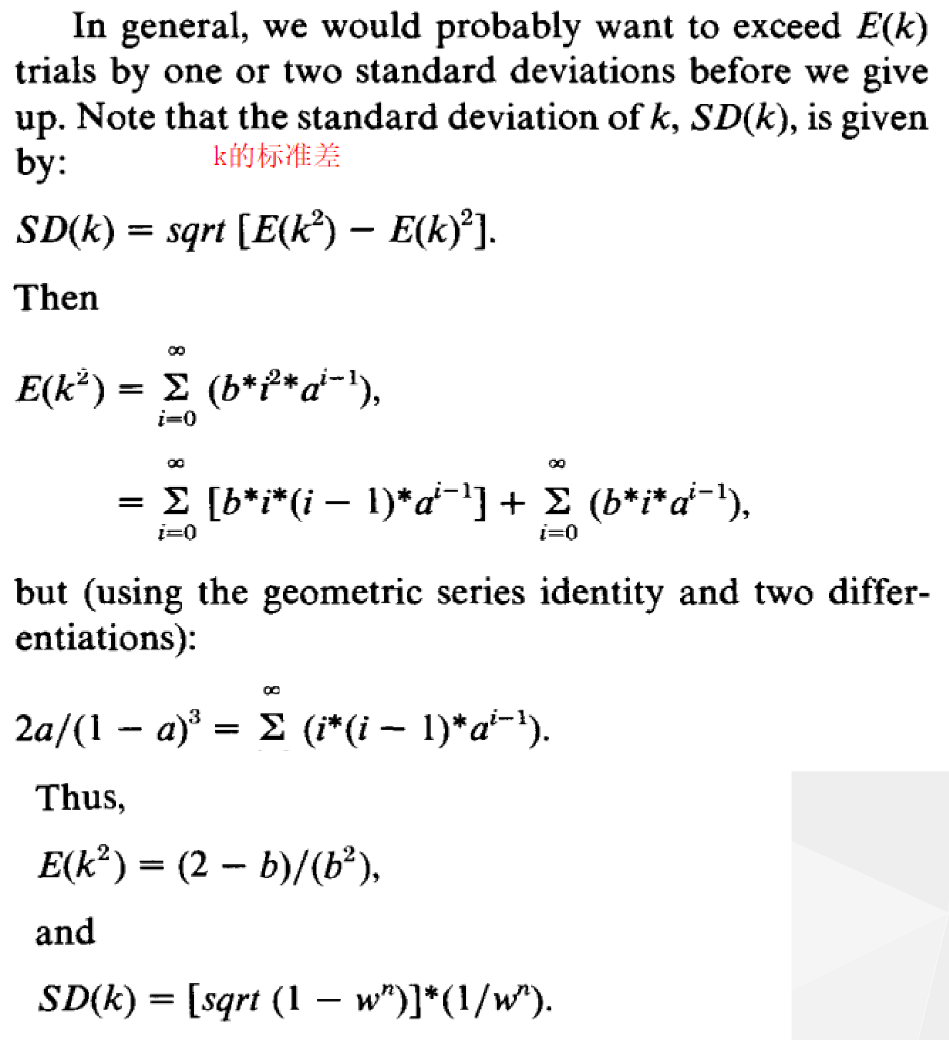
a=1-b 表示所有点(n个点)中至少有一个不符合容限误差的概率；

z：算法迭代中，至少有一次拟合模型所用的点都是符合容县误差的概率；



The following is a tabulation of some values of $E(k)$ for corresponding values of n and w:

w	$n = 1$	2	3	4	5	6
0.9	1.1	1.2	1.4	1.5	1.7	1.9
0.8	1.3	1.6	2.0	2.4	3.0	3.8
0.7	1.4	2.0	2.9	4.2	5.9	8.5
0.6	1.7	2.8	4.6	7.7	13	21
0.5	2.0	4.0	8.0	16	32	64
0.4	2.5	6.3	16	39	98	244
0.3	3.3	11	37	123	412	—
0.2	5.0	25	125	625	—	—



In general, we would probably want to exceed $E(k)$ trials by one or two standard deviations before we give up. Note that the standard deviation of k, $SD(k)$, is given by:

k的标准差

$$SD(k) = sqrt\ [E(k^2) - E(k)^2].$$

Then

$$E(k^2) = \sum_{i=0}^{\infty} (b*i^2*a^{i-1}),$$

$$= \sum_{i=0}^{\infty} [b*i*(i-1)*a^{i-1}] + \sum_{i=0}^{\infty} (b*i*a^{i-1}),$$

but (using the geometric series identity and two differentiations):

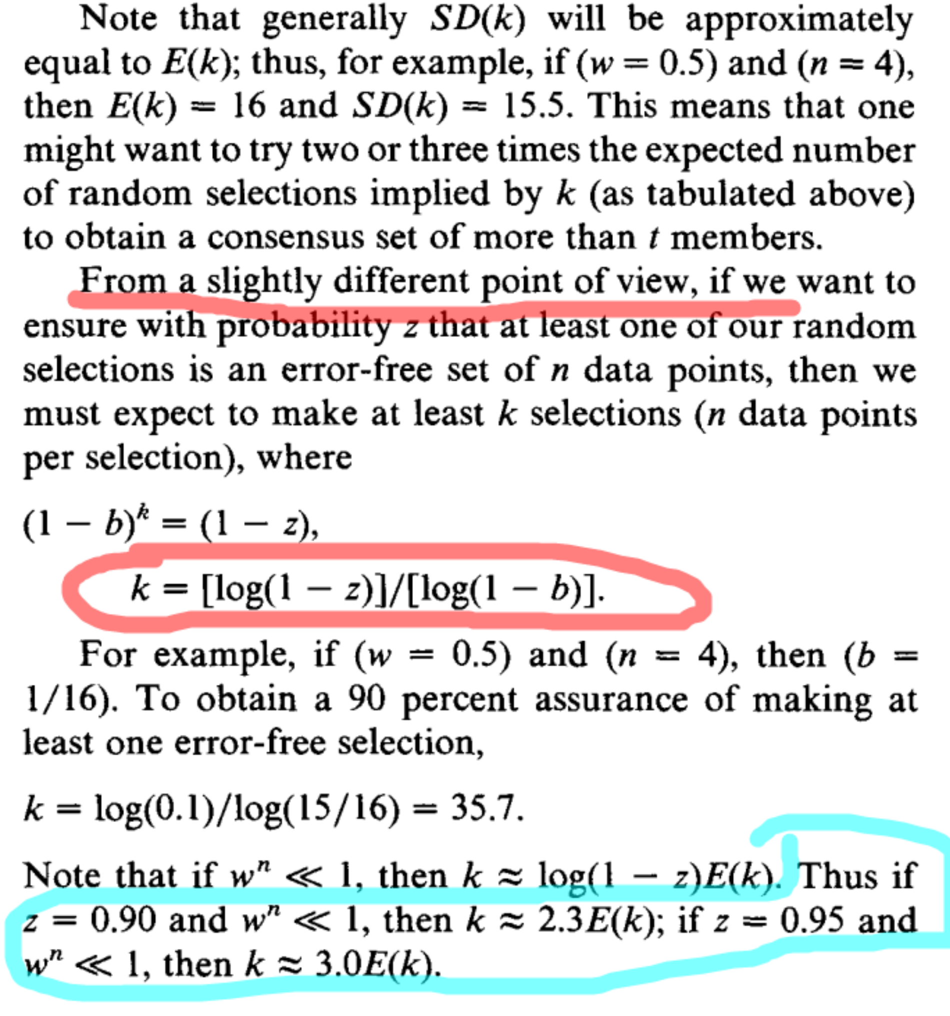
$$2a/(1-a)^3 = \sum^{\infty} (i*(i-1)*a^{i-1}).$$

Thus,

$$E(k^2) = (2-b)/(b^2),$$

and

$$SD(k) = [sqrt\ (1-w^n)]*(1/w^n).$$



Note that generally $SD(k)$ will be approximately equal to $E(k)$; thus, for example, if $(w = 0.5)$ and $(n = 4)$, then $E(k) = 16$ and $SD(k) = 15.5$. This means that one might want to try two or three times the expected number of random selections implied by k (as tabulated above) to obtain a consensus set of more than t members.

From a slightly different point of view, if we want to ensure with probability z that at least one of our random selections is an error-free set of n data points, then we must expect to make at least k selections (n data points per selection), where

$$(1-b)^k = (1-z),$$

$$k = [\log(1-z)]/[\log(1-b)].$$

For example, if $(w = 0.5)$ and $(n = 4)$, then $(b = 1/16)$. To obtain a 90 percent assurance of making at least one error-free selection,

$$k = \log(0.1)/\log(15/16) = 35.7.$$

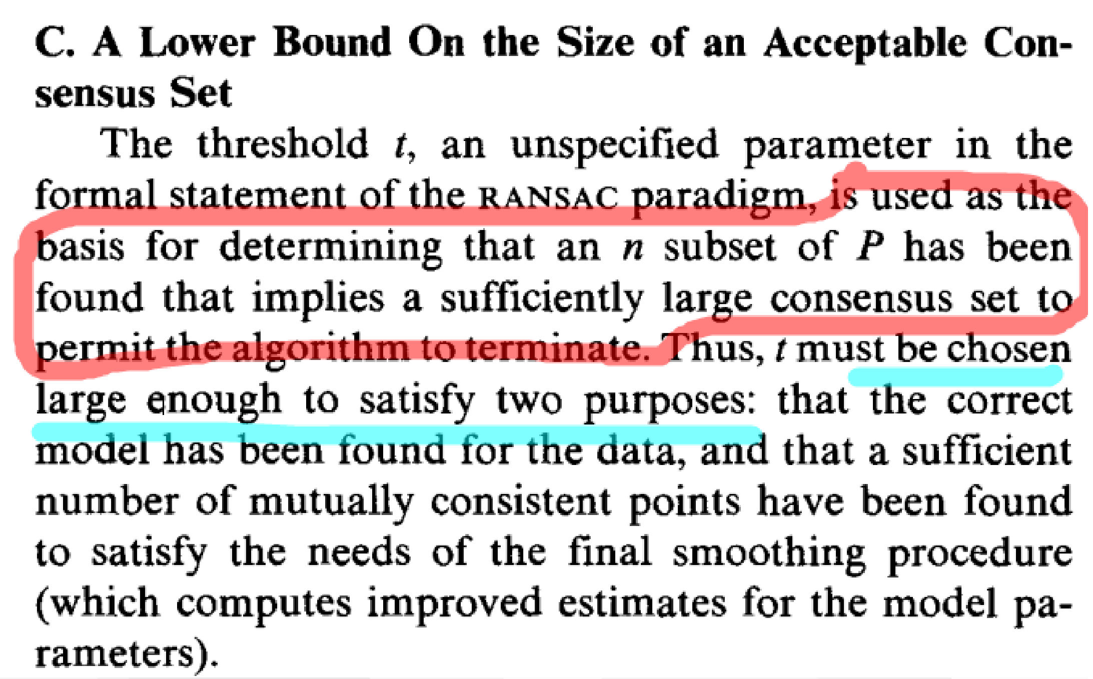
Note that if $w^n \ll 1$, then $k \approx \log(1-z)E(k)$. Thus if $z = 0.90$ and $w^n \ll 1$, then $k \approx 2.3E(k)$; if $z = 0.95$ and $w^n \ll 1$, then $k \approx 3.0E(k)$.

结论：

当z=0.9， b= ω^n<<1时 k≈2.3E(k) E(k)为期望值

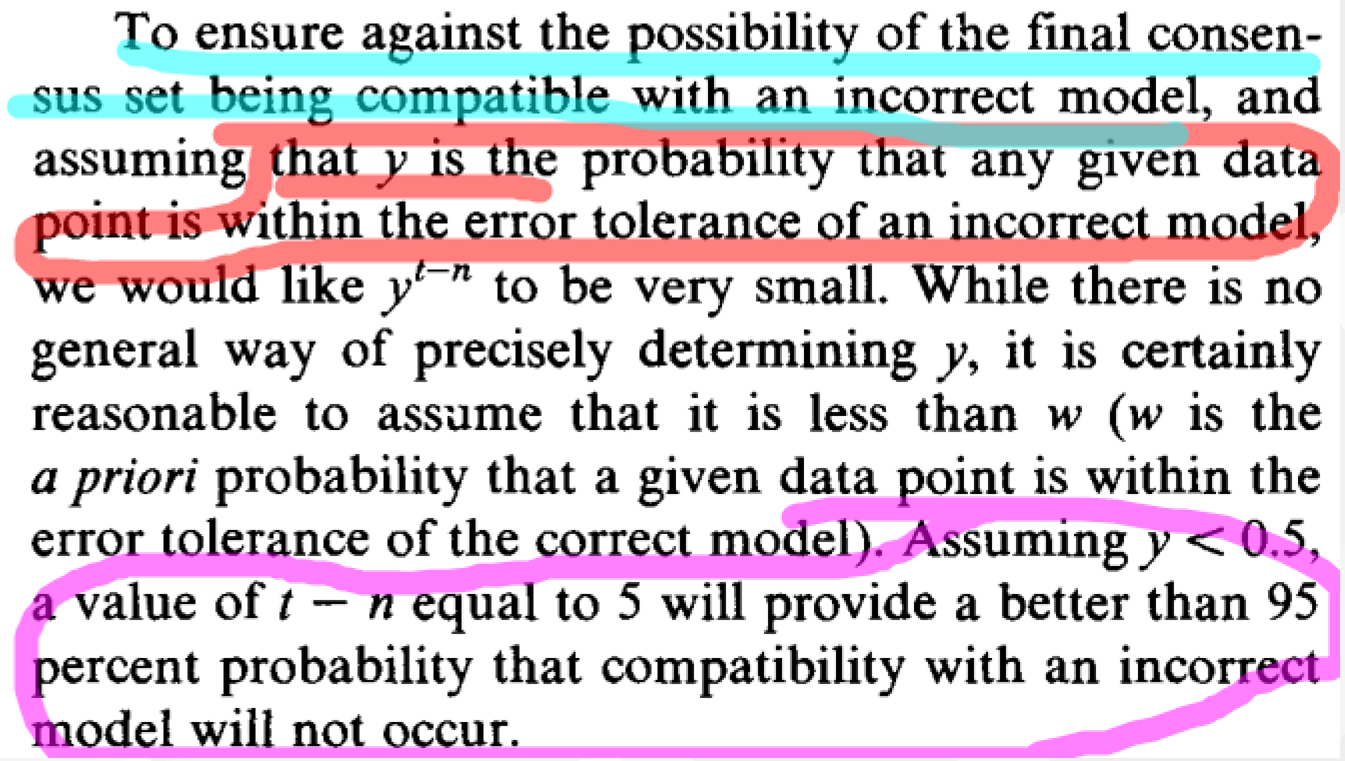
当z=0.95， b= ω^n<<1时 k≈3E(k) E(k)为期望值

• 阈值t的确定



C. A Lower Bound On the Size of an Acceptable Consensus Set

The threshold t, an unspecified parameter in the formal statement of the RANSAC paradigm, is used as the basis for determining that an n subset of P has been found that implies a sufficiently large consensus set to permit the algorithm to terminate. Thus, t must be chosen large enough to satisfy two purposes: that the correct model has been found for the data, and that a sufficient number of mutually consistent points have been found to satisfy the needs of the final smoothing procedure (which computes improved estimates for the model parameters).



To ensure against the possibility of the final consensus set being compatible with an incorrect model, and assuming that y is the probability that any given data point is within the error tolerance of an incorrect model, we would like y^{t-n} to be very small. While there is no general way of precisely determining y, it is certainly reasonable to assume that it is less than w (w is the *a priori* probability that a given data point is within the error tolerance of the correct model). Assuming $y < 0.5$, a value of $t - n$ equal to 5 will provide a better than 95 percent probability that compatibility with an incorrect model will not occur.

t：阈值，当内点的数量大于t时，判定拟合出的那个模型合理，否则不合理；

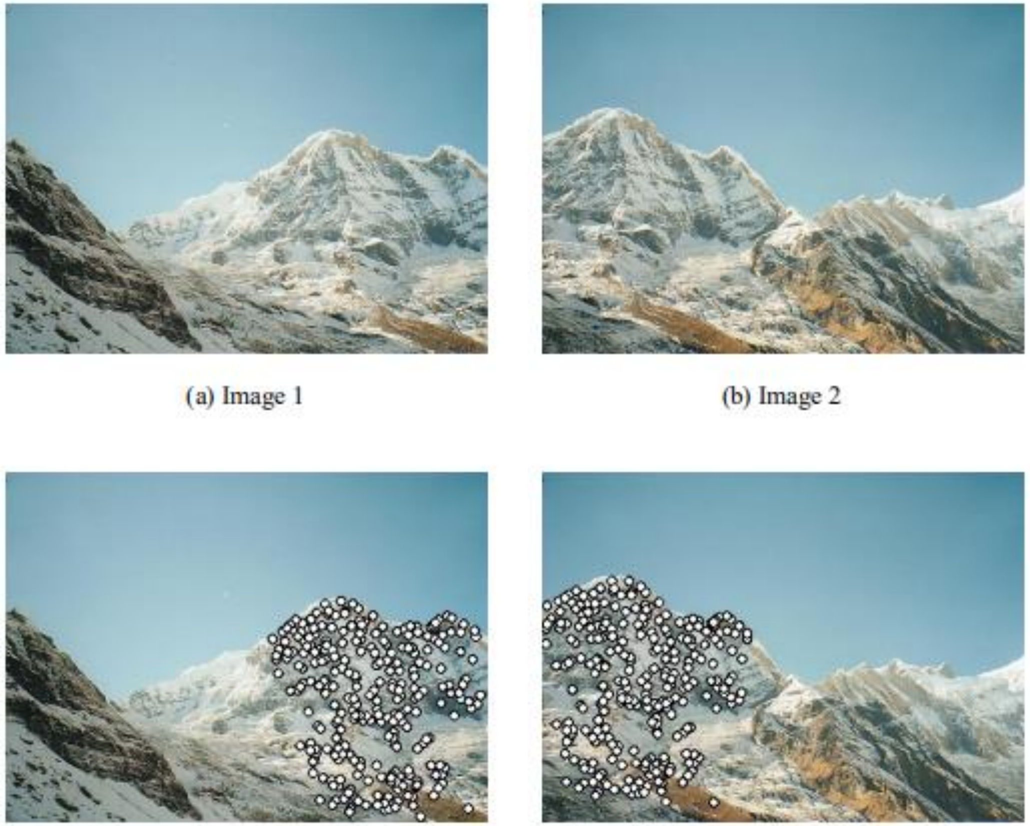
t：应该选择的足够大，能够达到两个目的：能够使得算法选出足够的内点（符合容限误差），能够拟合出正确的模型；

为了避免最终的点都符合容限误差，但是拟合不出一个正确的模型的情况。

y：表示符合容限误差的点拟合不出正确模型的概率；

假设y<0.5，当t-n>5时，有超过95的概率确定出一个正确的模型；

• 应用：SIFT+RANSAC图像拼接

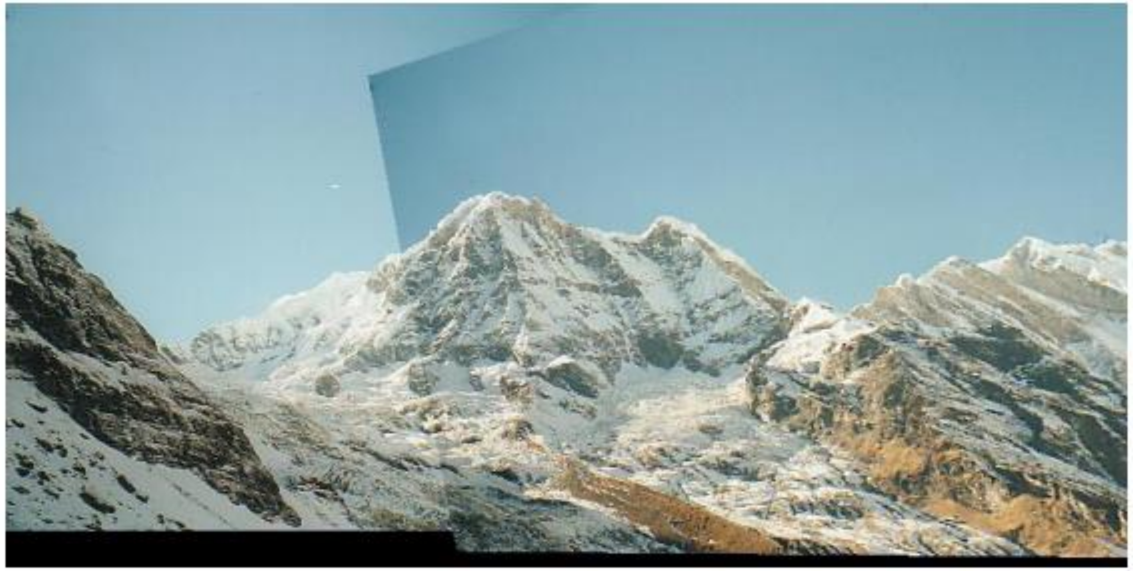


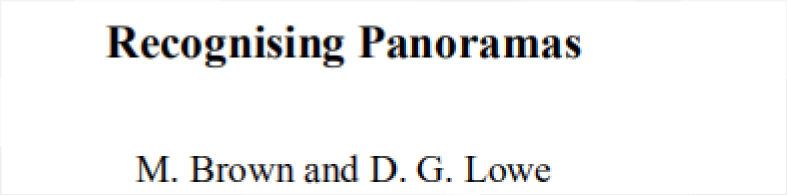
(a) Image 1

(b) Image 2

(c) SIFT matches 1

(d) SIFT matches 2





Recognising Panoramas

M. Brown and D. G. Lowe

THANK YOU!