The Likelihood-Ratio Test and Efficient Robust Estimation

Andrea Cohen[1], Christopher Zach[2]
[1]Department of Computer Science, ETH Zurich, Switzerland
[2]Toshiba Research, Cambridge, UK
acohen@inf.ethz.ch
chzach@crl.toshiba.uk

Abstract

Robust estimation of model parameters in the presence of outliers is a key problem in computer vision. RANSAC inspired techniques are widely used in this context, although their application might be limited due to the need of a priori knowledge on the inlier noise level. We propose a new approach for jointly optimizing over model parameters and the inlier noise level based on the likelihood ratio test. This allows control over the type I error incurred. We also propose an early bailout strategy for efficiency. Tests on both synthetic and real data show that our method outperforms the state-of-the-art in a fraction of the time.

1. Introduction

Robust estimation of model parameters from data points contaminated with outlier noise is a fundamental problem in geometric computer vision. In many applications the random sampling paradigm (RANSAC [7]) is employed due to its efficiency and its robustness to large amounts of outliers. In statistical terms, RANSAC (and its many modern variants), aims to find a maximum likelihood estimate (MLE) for the desired model parameters and inlier fraction by using a mixture model of inlier and outlier distributions. Since the objective for MLE has many local minima, RANSAC uses many restarts by sampling model parameters from subsets of data to find an approximate MLE. The solely optimized parameters of the mixture are the model parameters and the mixture coefficient. Other parameters such as the inlier noise level are assumed to be provided in advance.

In several applications the inlier noise level is not known in advance, and even a sensible estimate may be unavailable, e.g. if two image-based 3D models with unknown scale need to be merged, or if nothing is known on how the data points were obtained. More frequently, the model parameters have usually too few degrees of freedom—the models are underspecified—for efficiency reasons, e.g. lens distortion is typically not modeled in robust estimation techniques for multi-view geometry. Underspecified models therefore may invalidate assumptions on the inlier noise level. Overall, adding the inlier noise level as additional unknown to be determined from given data is beneficial in many applications of robust model estimation.

Jointly optimizing over model parameters and the inlier noise level requires much more emphasis on reasoning about statistical errors that can occur in robust estimation. In standard RANSAC with a fixed inlier noise level, the only relevant statistical error is a type II error (false negative rate) of missing a well-supported model due to the bounded number of sampling iterations. If the noise level is unknown, reasoning about the type I error (false positive rate) of hallucinating a model with many inliers due to a large chosen value for the noise level is critical. The need of controlling the type I error in such setting has been identified in [19, 13], but to our knowledge using the generalized likelihood ratio test (LRT) for robust model estimation with unknown inlier noise level has not been considered in the literature so far. In this work our contributions are: (i) we propose to use the LRT test statistic as the objective function for robust model estimation with unknown inlier noise level. It allows us to control the type I error of fitting an insignificant model into random, unstructured data in a statistically sound manner. (ii) We show how the number of sampling iterations given by the RANSAC formula leads to reduction of the search range for the inlier noise level, accelerating model verification. (iii) We propose a bailout test to further speed up model verification, which makes our approach (which also determines the inlier noise level) comparable in runtime to modern RANSAC implementations (which assume a given noise level).

2. Related Work

When the inlier noise level is unknown, standard robust estimation with a fixed criterion (i.e. inlier threshold) to determine inliers and outliers is not applicable. Numerous methods have been proposed to address this task (e.g. least median of squares [17], MINPRAN [19], AMLESAC [8], AC-RANSAC [13, 14], RECON [16], kernel density esti-

mation [22, 23] and the pbM-estimator [12, 18, 20].

Similar to MINPRAN [19] and AC-RANSAC [13, 14] we cast the problem of joint estimation of a good model and the noise level as a statistical test between two hypotheses: is the set of given data points generated by an interesting and informative distribution, or are the data points drawn from an uninformative background distribution. Among other (technical) differences between our and these methods, the most important difference is that both MINPRAN and AC-RANSAC link the details of how models are generated with the test statistic of the hypothesis test: due to an underlying independence assumption, MINPRAN uses a very conservative estimate for the type I error, hence rejecting any estimated structure in the data points in the limit (i.e. when sampling many model parameters). In AC-RANSAC the specifics of how models are sampled enters as a coefficient in the test statistic, and it relies on models being generated from random sets sampled from the data points. We fundamentally believe that the criterion for testing the hypothesis whether a set of data points exhibits structure against its alternative should be independent of how models are generated. The fact that in the majority of cases we cannot generate all models exhaustively should, e.g., only affect the type-II error (since we may miss a good model), but any criterion should be independent of this limitation. We share with MINPRAN and AC-RANSAC the utilization of an uninformative background distribution to model the null hypothesis.

RECON [16] uses a completely different approach: the method assumes that good models generated by random sampling combined with the correct noise level share most of their inliers and have similar residual distributions. One obvious but degenerate solution by setting the selected noise level to a large enough value, such that all data points are reported as inliers, is prohibited by a user-provided upper bound for the noise level. The method does not provide any statistical reasoning (such as on the type I error of hallucinating a structure in random data), and its run-time complexity is quadratic in the number of sampled models.

AMLESAC [8] (and a related approach [5]) is an extension of MLESAC [21] by including the inlier noise level into the set of parameters for maximum-likelihood estimation, in addition to the mixture coefficient and model parameters. Neither MLESAC nor AMLESAC include hypothesis testing, and both approaches rely on frequent non-convex non-linear optimization to find the MLE.

The pbM-estimator [12, 18, 20] estimates the inlier noise level by using kernel density estimation. However, this method is intrinsically linked to linear models.

If the inlier noise level is unknown, the model evaluation stage is computationally more expensive, since models have to be evaluated with respect to a family of candidate noise levels. Both MINPRAN [19] and AC-RANSAC [13] sort the data points with respect to their residuals to evaluate the model for increasing noise levels. In order to handle large datasets we propose to utilize an early bailout approach in the spirit of [3, 2, 10, 4], but we extend it to the case of evaluating multiple noise levels simultaneously.

3. The likelihood-ratio test for robust estimation

In this section we describe the basic method for robust estimation with unknown inlier noise level, and we state the underlying assumptions of our approach. Performance enhancements to improve the runtime efficiency are deferred to Section 3.5. The problem of interest is to answer the following questions given data points $\mathbf{X} = (X_1, \dots X_N)$:

1. Given a model class with parameters denoted by θ (which need to be estimated from data), does $\mathbf{X}$ exhibit a sufficient non-random structure?
2. If the answer is yes, then determine the model θ^* that explains the given observations $\mathbf{X}$ the best.

We use the likelihood-ratio test as described in the following to answer both questions: if the likelihood ratio test statistic is above a critical value we can assess (with user-specified type I error) the non-randomness of the dataset. Further, we use a RANSAC-type argument on the number of required sampling iterations in order to guarantee (for a given confidence value) that the best model so far cannot be improved. This is in contrast to a-contrario approaches [13, 14] or RECON [16], which stop the iterations once a non-random model has been found (which is subsequently refined).

3.1. Notations

The task at hand is to find model parameters $\theta \in \Theta$ that robustly explain n given data points $\mathbf{X} = (X_1, \dots, X_n)$, where w.l.o.g. each data point is an element from a bounded D-dimensional domain $[0,1]^D$. Θ is the space of model parameters, which have d degrees of freedom (i.e. Θ is a d-dimensional manifold). Robust fitting means that a model is required to explain only *inlier* data points, that have their respective residual $e(X;\theta) \leq \sigma$ for an inlier noise level σ. Here $e(\cdot;\cdot)$ measures the error of a data point X with respect to the model parameters θ. σ is unknown as well and is an element from a finite set $\Sigma = \{\sigma_{\min}, \dots, \sigma_{\max}\}$, i.e. we use a quantized set of candidate noise levels. For model parameters θ and a noise level σ we define the inlier region $I(\theta,\sigma) \stackrel{\text{def}}{=} \{X \in [0,1]^D : e(X;\theta) \leq \sigma\} \subseteq [0,1]^D$.

3.2. Non-random structures

We use the same basic idea of using a non-informative background distribution as in [19, 13] to assess the likelihood of finding a model in the data supported by an observed number of inliers, but we differ in our choice of

test statistic for the hypothesis test. We utilize the following generative model for the observed data: data points are drawn from a mixture of an inlier distribution (with weight $\rho \in [0,1]$) and a non-informative outlier distribution (with weight $1-\rho \in [0,1]$). For given model parameters θ and inlier threshold σ, the inlier distribution is assumed to be uniform in the set $I(\theta,\sigma)$. We denote the area of $I(\theta,\sigma)$ by p_σ (and for simplicity we omit the usually weak dependence of p_σ on the model parameters θ). Thus, p_σ is the probability of a point drawn from the uniform background distribution to be an inlier. Overall, we have the probability density for a data point X_i given by

$$P(X_i;\theta,\sigma,\rho) = (1-\rho)1_{[0,1]^D}(X_i) + \frac{\rho}{p_\sigma}1_{I(\theta,\sigma)}(X_i). \tag{1}$$

If we have n data points and k observed inliers, then the joint density of $\mathbf{X} = (X_1,\ldots,X_n)$ is

$$P(\mathbf{X};\theta,\sigma,\rho) = \left(1-\rho+\rho/p_\sigma\right)^k (1-\rho)^{n-k}. \tag{2}$$

For the probability the exact instance $\mathbf{X}$ and the model parameter θ is not of importance, but it depends only on the number of observed inliers k. In the following we will write Eq. 2 as $P(k;\sigma,\rho)$. k is a random variable defined via

$$k \stackrel{\text{def}}{=} |\{i : X_i \in I(\theta,\sigma)\}| = |\{i : e(X_i;\theta) \le \sigma\}|. \tag{3}$$

In order to have a compact notation we make the dependence of k on the realization $\mathbf{X}$, the current model parameters θ, and the inlier threshold σ implicit.

We define the (simple) null hypothesis H_0 by fixing $\rho = 0$, i.e. all data points are generated by a uniform background distribution with density $1_{[0,1]^D}(\cdot)$ (and hence the values of θ and σ do not matter). The alternative hypothesis H_1 is a composite one with parameters θ, σ and ρ to be determined from the data. Note that the hypotheses are nested, i.e. H_0 is a special case of H_1. The (generalized) likelihood ratio test is therefore applicable, and the test statistic is given by

$$\Lambda(k) = \frac{\sup_{\theta,\sigma,\rho} P(\mathbf{X};\theta,\sigma,\rho)}{\prod_i 1_{[0,1]^D}(X_i)} = \sup_{\theta,\sigma,\rho} P(k;\sigma,\rho), \tag{4}$$

since the denominator is 1. Wilks' theorem tells us that under H_0 the quantity $2\log\Lambda(k)$ is asymptotically χ^2 distributed with $d+2$ degrees of freedom.[1] This allows us to choose critical values c for the hypothesis test given a user-defined choice α of the type I error. Note that specifying α allows to use a uniform value with a clear interpretation across all applications.

For the likelihood ratio test one needs to maximize the likelihood w.r.t. the unknowns θ, σ, and ρ. Optimizing with respect to θ is performed by random sampling, and maximization with respect to σ is performed by exhaustive evaluation for values $\sigma \in \{\sigma_{\min},\ldots,\sigma_{\max}\}$. One can easily show that the maximum likelihood estimate for $\hat{\rho}$ given the number of *apparent* inliers k is given by

$$\hat{\rho} = \max\left\{0, \frac{k/n - p_\sigma}{1-p_\sigma}\right\} = \max\left\{0, \frac{\varepsilon - p_\sigma}{1-p_\sigma}\right\}, \tag{5}$$

where we introduced the *apparent* inlier ratio $\varepsilon = k/n$. This is an intuitive relation for the following reasons:

1. If $\varepsilon < p_\sigma$, then we observe fewer apparent inliers than even expected under H_0. It also leads to a negative mixture coefficient $\hat{\rho}$, which is infeasible. Therefore an upper bound for σ is induced by $p_\sigma \le \varepsilon$.
2. If $\varepsilon \ge p_\sigma$, then the above relation can be rewritten as
$$\varepsilon = \hat{\rho} + (1-\hat{\rho})p_\sigma, \tag{6}$$
i.e. the observed inlier ratio is a mixture of the "true" inlier fraction (or its MLE) and the chance of "random" inliers from the background distribution.

In the following we add the constraint $\varepsilon \ge p_\sigma$ and use the relation $\hat{\rho} = (\varepsilon - p_\sigma)/(1-p_\sigma)$. Plugging the expression for $\hat{\rho}$ into the test statistic $L(\varepsilon,\sigma) \stackrel{\text{def}}{=} 2\log P(k;\sigma,\hat{\rho})$ yields (after rearrangements)

$$L(\varepsilon,\sigma) = 2n\left(\varepsilon\log\left(\frac{\varepsilon}{p_\sigma}\right) + (1-\varepsilon)\log\left(\frac{1-\varepsilon}{1-p_\sigma}\right)\right). \tag{7}$$

if $\varepsilon < p_\sigma$, then $L_\theta(\varepsilon,\sigma) = 0$.

Remark 1. Compared to MLESAC [21] or AMLESAC [8], using uniform inlier and outlier distributions allows a closed form MLE for the mixture coefficient ρ and a non-linear optimization step is not required. Further, we believe that the uniform inlier distribution can cope better with under-specified models (e.g. not including lens distortion parameters into the model for geometric vision problems).

Remark 2. In contrast to MINPRAN [19] or a-contrario estimation [13] the likelihood ratio test statistic is agnostic on how the score $L(\varepsilon,\sigma)$ is maximized. The fact that we cannot test all model parameters $\theta \in \Theta$ exhaustively only implies, that e.g. if $L_\theta(\varepsilon,\sigma)$ is maximized by random sampling, we may miss a model θ with the highest score (i.e. we increase the type II error of incorrectly "accepting" H_0[2]).

Remark 3. For fixed σ (and therefore p_σ), $L(\varepsilon,\sigma)$ is a convex and monotonically increasing function with respect to ε, and strictly convex in $\varepsilon \in [p_\sigma, 1]$. This will be important in Section 3.3, since for a fixed σ we need to compute a critical inlier ratio $\underline{\varepsilon}$ such that $L(\underline{\varepsilon},\sigma) = c$ for a given c.

[1] d d.o.f. from the model parameters θ, one additional d.o.f. from σ and ρ, respectively.

[2] Since the type II error is not controlled, the proper statistical terminology is "not rejecting" H_0.

Straightforward implementation of maximizing the likelihood ratio test statistic by random sampling of models and subsequent evaluation of the model for all values of σ is computationally expensive and requires $O(n|\Sigma|)$ time per model. By sorting the residuals one can virtually set $\Sigma = \mathbb{R}$ and reduce the runtime complexity to $O(n \log n)$ [19, 13], but this is still slow in practice. In the following section we propose an approach that bails out of the evaluation step early if a model is unlikely to outperform the best hypothesis so far. It shares its motivation with randomized RANSAC [3, 10, 4], but uses a different approach, since models are evaluated for many values of σ *in parallel.*

3.3. Limited search range for σ

For a user-defined critical value c and every inlier noise level σ there is a minimal value $\underline{\varepsilon}$ such that $L(\varepsilon, \sigma) \geq c$ for all $\varepsilon > \underline{\varepsilon}$ (which follows from the monotonicity of $L(\cdot, \sigma)$). $\underline{\varepsilon} \in [0, 1]$ may not exist for combinations of c and σ, i.e. $L(\varepsilon, \sigma) < c$ for all choices of $\varepsilon \in [0, 1]$. This implies that given the critical value c (obtained by specifying the type I error) there is a global upper bound for the range σ to consider, and any inlier noise level larger than this threshold can never lead to a test statistic exceeding the critical value. Consequently, we maintain an array $\underline{\varepsilon}(\sigma)$, $\sigma \in \Sigma = \{\sigma_{\min}, \ldots, \sigma_{\max}\}$ with minimally required inlier ratios $\underline{\varepsilon}(\sigma)$ to exceed the critical value c. Since $L(\varepsilon, \sigma)$ is monotonically increasing as a function of ε, numerical computation of $\underline{\varepsilon}(\sigma)$ can be efficiently performed via the bisection method.

The same reasoning can be applied for the minimum inlier ratio required for each σ in order to exceed the best scoring model so far. Hence, whenever $L(\hat{\varepsilon}, \hat{\sigma})$ for a current model θ, currently observed inlier ratio $\hat{\varepsilon}$, and corresponding value $\hat{\sigma}$ is a new maximum Λ^* of the test statistic, the required inlier ratios $\underline{\varepsilon}(\sigma)$ are updated (increased) for all σ.

3.4. Number of sampling iterations

One important assumption linking the test statistics $L(\varepsilon, \sigma)$ with sampling arguments is the following: meaningful structures (meeting the critical value or outperforming the best model found so far) are only reported by non-contaminated samples, i.e. $\max_\theta L(\varepsilon, \sigma) > \max\{c, \Lambda^*\}$[3] only if the model θ was estimated from an uncontaminated sample. The converse is not necessarily true, since even the correct model might not be statistically significant. This assumption implies that we can rule out values of σ, if the number of sampling iterations was large enough to generate at least one good model with high confidence. If for a value of σ the test statistic $L(\varepsilon, \sigma) < \max\{c, \Lambda^*\}$ after

$$M(\sigma) \stackrel{\text{def}}{=} \frac{\log(\gamma)}{\log(1 - \underline{\varepsilon}(\sigma)^s)} \tag{8}$$

sampling iterations, where s is the size of the minimal sample set, we can (with high confidence $1 - \gamma$) conclude that the data points show no sufficient structure at noise level σ. This allows to successively eliminate noise levels σ to test over the iterations. Since $\underline{\varepsilon}(\sigma)$ is monotonically increasing by construction, it means that $\sigma_{\max}$ can be reduced successively. It also allows us to adaptively refine the required number of sampling iterations.

Since by design the evaluation of model parameters with respect to a fixed $\sigma \in \Sigma$ is stopped after $M(\sigma)$ sampling iterations, such that the chance of missing an uncontaminated model is γ, we have to compute the respective probability of missing an uncontaminated model for any value of σ. Application of the law of total probability,

$$\begin{aligned} P(A) &= \sum_j P(A, B_\sigma) = \sum_\sigma P(A \mid B_\sigma) P(B_\sigma) \\ &= (1 - \gamma) \sum_\sigma P(B_\sigma) = 1 - \gamma \end{aligned}$$

(and defining A as the event of drawing at least one good model regardless of the value of σ, and B_σ is the event of σ attaining its respective value), implies that $P(\text{at least one good model drawn for } \textit{any } \sigma) \geq 1 - \gamma$. This result implies that if the number of actually performed sampling iterations T is larger than $M(\sigma)$ for every σ (which due to monotonicity is equivalent of testing whether $T \geq M(\sigma_{\min})$), then we can terminate the iterations and return the best scoring model parameters θ with the respective noise level σ. It also implies that the type II error of incorrectly accepting H_0 increases by γ.

3.5. Early bailout

A statistically justified approach for early stopping of model parameter evaluation proposed in [10, 4] is based on Wald's sequential probability ratio test (SPRT). Unfortunately, SPRT is not applicable in our setting, since we have a composite alternative hypothesis (with unknown mixture coefficient ρ and inlier threshold σ). SPRT has been generalized to composite hypotheses, and asymptotic results for the type I and type II error are available [9]. Nevertheless, we employ a more standard argument using concentration bounds to control the probability of incorrect early stopping.

In the evaluation phase for a current model θ the task is to determine quickly whether $\hat{\varepsilon}(\sigma, \theta) \geq \underline{\varepsilon}(\sigma)$ for any σ can be achieved, or if $\hat{\varepsilon}(\sigma, \theta) < \underline{\varepsilon}(\sigma)$ for all σ with high probability. Let $\hat{\varepsilon}_m(\sigma, \theta)$ be the observed inlier ratio after evaluating $m \leq n$ data points. We will use a one-sided version of Hoeffding's inequality,

$$P\left(\bar{Z} \leq \mathbf{E}(\bar{Z}) - t\right) \leq e^{-2t^2 m}, \tag{9}$$

for independent random variables $Z_i \in [0, 1]$ and $\bar{Z} \stackrel{\text{def}}{=} \frac{1}{m} \sum_{i=1}^m Z_i$. In our setting $Z_i = 1$ if the i-th data point

[3] Recall that ε has an implicit dependence on θ and $\mathbf{X}$.

X_i is an inlier with respect to the current model θ and noise level σ, and 0 otherwise. $\bar{Z}$ therefore corresponds to the observed inlier ratio after m data points, $\hat{\varepsilon}_m(\sigma, \theta)$. Under the hypothesis that $\mathbf{E}(\hat{\varepsilon}_m(\theta, \sigma)) \geq \underline{\varepsilon}(\sigma)$, this bound implies

$$P\big(\hat{\varepsilon}_m(\sigma, \theta) \leq \underline{\varepsilon}(\sigma) - \tau_m(\sigma)\big) \leq e^{-2\tau_m(\sigma)^2 m}, \qquad (10)$$

For a deviation $\tau_m(\sigma)$. This means that the probability of discarding a model with inlier ratio not worse than $\underline{\varepsilon}(\sigma)$ decreases exponentially fast to 0 w.r.t. m. For computational efficiency we want m (the number of evaluated data points) and the r.h.s. of the inequality (the type II error denoted by β') to be the same for all values of σ (and therefore $\tau_m(\sigma) = \tau_m$ is independent of σ). Thus, in the evaluation phase we check the criterion $\hat{\varepsilon}_m(\sigma, \theta) \leq \underline{\varepsilon}(\sigma) - \tau_m$ after m data points for all (still considered) values of σ. Recall that the type I error of early committing to $\hat{\varepsilon}(\theta, \sigma) \geq \underline{\varepsilon}(\sigma)$ is 0, and for a given user-specified type II error β' we obtain

$$\beta' = e^{-2\tau_m^2 m} \qquad \text{or} \qquad \tau_m = \sqrt{-\frac{\log \beta'}{2m}}.$$

In contrast to the $T(d, d)$ test [3], the choice of m does not fix the type II error,[4] hence we have freedom to choose m.

The presentation above applies if the bailout test is applied exactly once after m data points. A natural extension is to repeat the bailout test after every B data points, where B is the size of a batch. Since checking the bailout criterion comes at a certain computation cost (of complexity $O(|\Sigma|)$), it is –in contrast to SPRT– not efficient to apply the test after *every* data point, but to use larger batches of size B. We determine B empirically such that the runtime of evaluating B data points is similar to the time required for the bailout test (hence neither the evaluation of data points nor the bailout test dominate the runtime). In such setting the bailout test is applied $Q = \lceil n/B \rceil$ times. Via the union bound the total type II error β of bailing out incorrectly at any of the Q tests is bounded by $Q\beta'$, hence we choose $\beta' = \beta/Q$ for a user-specified type II error β.[5] Plugging this into the expression for τ_m yields

$$\tau_m = \sqrt{-\frac{\log(\beta/Q)}{2m}} = \sqrt{-\frac{\log \beta - \log Q}{2m}}. \qquad (11)$$

Note that the number Q of bailout tests only mildly influences the deviation τ_m: for $\beta = 0.05$ and $B = 100$ we have $\tau_B \approx 0.16276$ for $n = 1000$, $\tau_B \approx 0.19495$, and $\tau_B \approx 0.22253$ for $n = 100000$. From the expression for τ_m we see that the largest difference in their values is for the smallest tested value of m, i.e. $m = B$, as τ_m monotonically converges to 0 with increasing m.

[4] In the $T(d, d)$ test τ_m is essentially fixed.

[5] This means we essentially apply the Bonferroni correction for multiple hypotheses tests. Thus, we are not underestimating the type II error (unlike the bailout method proposed in [2] as pointed out in [4]).

Remark 4. The increase of the type II error by early bailout in the evaluation step has the following consequence: the expected number of generated models needed to see (and fully evaluate) the first good model is increased from $1/\varepsilon^s$ to $1/[\varepsilon^s/(1-\beta)]$, where β is the overall type II error of the early bailout step. Thus, Eq. 8 needs to be modified to

$$M(\sigma) \stackrel{\text{def}}{=} \frac{\log(\gamma)}{\log\big(1 - \underline{\varepsilon}(\sigma)^s(1-\beta)\big)}. \qquad (12)$$

The total type I error α is therefore unaffected.

Algorithm 1 Robust estimation using LRT

Require: Model class Θ, data points $\mathbf{X} = (X_1, \ldots, X_n)$
Require: Type I error α, type II error β, confidence $1 - \gamma$
$\forall\sigma$: compute $\underline{\varepsilon}(\sigma) : L\big(\underline{\varepsilon}(\sigma), \sigma\big) = c$ via bisection
$T \leftarrow M(\sigma_{\min})$, $\Lambda^* \leftarrow 0$
repeat
  Hypothesize θ from random sample
  for all $X \in \mathbf{X}_i = \{X_0, \ldots X_N\}$ **do**
    Compute $e(X_i; \theta)$ and update $\hat{\varepsilon}(\sigma)$
    if $i \bmod B = 0$ **and** $\forall\sigma, \hat{\varepsilon}(\sigma) < \underline{\varepsilon}(\sigma) - \tau_{i+1}$ **then**
      Stop, discard θ and go back to step 4
    end if
  end for
  $\hat{\Lambda} \leftarrow \max_\sigma \{L(\hat{\varepsilon}(\sigma), \sigma)\}$
  if $\hat{\Lambda} > \Lambda^*$ **then**
    $\Lambda^* \leftarrow \hat{\Lambda}$, $\theta^* \leftarrow \theta$
    $\forall\sigma$: update $\underline{\varepsilon}(\sigma) : L\big(\underline{\varepsilon}(\sigma), \sigma\big) = \Lambda^*$
    $\forall\sigma$: update $M(\sigma)$ (Eq. 12), $T \leftarrow M(\sigma_{\min})$
    Decrease $\sigma_{\max}$ **if** $M(\sigma_{\max}) < T$
  end if
until T iterations are reached
return best model θ^* if $\Lambda^* \geq c$, null otherwise

3.6. Summary of the method

Algorithm 1 incorporates the ideas described in the previous sections. In addition to the model class and the data points it has three further inputs: the type I error α, that controls the false positive rate of hallucinating a model in random data; a type II error β that affects the bailout test in the model verification stage; and the confidence level $1 - \gamma$ that steers the number of sampling iterations. The algorithm returns the best model parameters if the null hypothesis is rejected, or null otherwise.

Our implementation uses a Fenwick tree [6] to accumulate $\hat{\varepsilon}(\sigma)$, which reduces the time to update the array $\hat{\varepsilon}(\sigma))_{\sigma\in\Sigma}$ for each data point from $O(|\Sigma|)$ to $O(\log(|\Sigma|))$.

4. Results

We evaluated our method on a number of estimation problems and we compared our performance with the cur-

		Plane		H		F	
Params	Method	T	err	T	err	T	err
$\varepsilon = 0.6$ $\sigma = 3.0$	RANSAC$_T$	31.5	1.41	66.4	2.97	271.1	3.1
	RECON	38.2	1.35	47	2.9	133.1	3.04
	AC-RANSAC	1001	0.76	1007	2.79	1001	0.85
	OURS	12.8	1.43	29.2	3.78	139.2	2.3
$\varepsilon = 0.4$ $\sigma = 2.0$	RANSAC$_T$	99.7	0.88	304.7	2.05	2431.5	2.2
	RECON	72.5	0.89	166.9	2.04	1879.3	2.34
	AC-RANSAC	1006	0.51	1037	1.83	1001	10.45
	OURS	47.8	0.96	156.0	2.63	2270.65	1.83
$\varepsilon = 0.3$ $\sigma = 2.0$	RANSAC$_T$	212	0.82	891.2	2.55	30153.5	2.42
	RECON	149.5	0.76	490	2.56	13751.4	2.44
	AC-RANSAC	1015	0.52	1122	1.83	1001	107.82
	OURS	115.4	0.98	521.5	2.54	17907.9	1.77
$\varepsilon = 0.3$ $\sigma = 4.0$	RANSAC$_T$	291.8	1.67	1373.4	4.61	35983.9	4.06
	RECON	172.4	1.72	612.3	4.47	13894.4	3.98
	AC-RANSAC	1017	0.98	1147	3.69	1001	118.95
	OURS	118.4	1.93	554.1	4.36	16734.8	3.01

Table 1: Results on synthetic data.

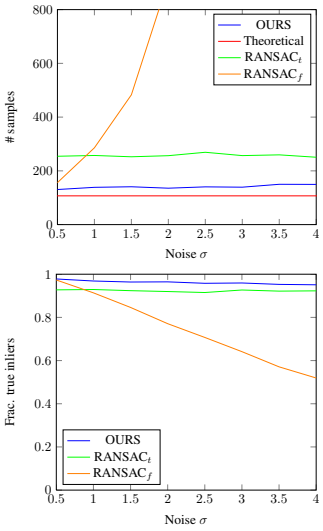


Figure 1: Top: iterations vs σ. Bottom: true inliers vs σ.

rent state-of-the-art as well as with standard RANSAC. For all experiments we used $\alpha = 1\%$, $\gamma = 1\%$ and $\beta = 5\%$. $\sigma_{\min}$ was set to 0.25 pixels, while $\sigma_{\max}$ was computed as described in Section 3.3. p_σ is computed as described in [14]. We tested our method on both synthetic and real data, achieving state-of-the-art results. The raw results obtained by our method are presented with no further local optimization. Our results are comparable to RECON and superior to both MINPRAN and AC-RANSAC while being more than an order of magnitude faster, especially for large datasets.

4.1. Synthetic data

We tested the performance of our method for three different problems: 3D plane estimation, homography estimation and fundamental matrix estimation. The results are presented in Table 1. We show the number of samples drawn T and mean error of the true inliers (err). We used the same settings as in [16]: we generated data with an inlier ratio ε varying from 0.25 to 1.0 in steps of 0.05. The inliers were altered with a zero-mean Gaussian noise with σ ranging from 0.5 to 4.0. The total number of points was chosen randomly between 500 and 2000 for each of the 500 trials. We compare our results with RANSAC using the true noise level as threshold (RANSAC$_T$), as well as with RECON and AC-RANSAC. The results of MINPRAN are not reported since it failed to achieve acceptable results for inlier ratios as high as 60% with an average running time of $13s$. Table 1 shows a subset of these experiments. The results obtained by RECON and our method are very similar and comparable to those achieved by RANSAC$_T$. However, RECON requires some knowledge on maximal noise level. This can be limiting for some applications where the scale of the data might be unknown. Note that AC-RANSAC failed to find a good solution for inlier ratios smaller than 40%. RECON's reported running times vary from $22ms$ to $1.7s$, while AC-RANSAC takes between $0.12s$ and $0.31s$. Our method achieved similar results in $0.9ms$ to $6ms$ for the lower inlier ratios. In addition, the accuracy of the solution, measured in the percentage of true inliers recovered, as well as running time, do not vary with respect to the noise levels, as opposed to standard RANSAC with a fixed threshold (RANSAC$_F$). This behaviour is illustrated in Figure 1.

4.2. Real Data

We evaluated our method on real data taken from [15] for three different applications: fundamental matrix, essential matrix and homography estimation. Results were averaged over 500 runs. We compare our performance with standard RANSAC, USAC1.0 [15] (both with a fixed inlier threshold of 1) and AC-RANSAC. Tables 2, 3 and 4 report number of inliers found (k), inlier error (error), number of samples (T), number of verifications per model (vpm) as well as total running time in milliseconds for each of the different estimation problems. For USAC1.0, T reports the number of samples/models rejected by the sample/model check steps. For most datasets, all methods obtain very similar results, with USAC1.0 showing a smaller variation in the number of inliers detected over all runs, due to the use of local optimization. Even though a fixed threshold of 1 works well for most cases, we show that for some datasets the noise level is actually higher, *e.g.* dataset A for homography estimation. Both AC-RANSAC and our method found a better solution with σ close to 4 pixels. AC-RANSAC fails to find a good solution for homography estimation on dataset E, probably due to the fixed number of iterations used in the method,

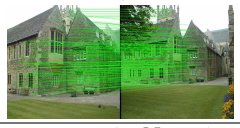
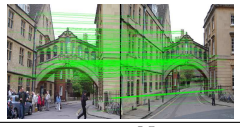
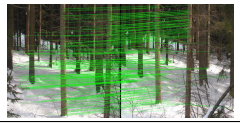
		RANSAC$_F$	USAC 1.0	AC-RANSAC	OURS
A: $\varepsilon = 0.48$, $N = 3154$ 	k	1412 ± 50	1495 ± 8	1414 ± 40	1455 ± 50
	error	1.28 ± 0.58	0.63 ± 0.24	0.36 ± 0.08	0.60 ± 0.29
	T	1420	$2/0$	1001	1142
	vpm	3154.0	940.9	3154.0	47.4
	time	255.90	11.84	1065.9	1.88
B: $\varepsilon = 0.57$, $N = 575$ 	k	315 ± 12	328 ± 3	312 ± 7	323 ± 7
	error	0.71 ± 0.45	0.06 ± 0.24	0.34 ± 0.15	0.70 ± 0.61
	T	385	$3/0$	1001	279
	vpm	575.0	423.8	575.0	38.5
	time	14.98	3.61	169.7	5.92
C: $\varepsilon = 0.38$, $N = 1088$ 	k	381 ± 13	406 ± 4	343 ± 44	397 ± 19
	error	1.79 ± 1.27	0.58 ± 0.28	0.36 ± 0.27	0.57 ± 0.24
	T	7935	$2/0$	1001	6585
	vpm	1088.0	472.2	1088.0	51.74
	time	546.53	13.74	327.3	9.51
D: $\varepsilon = 0.22$, $N = 1516$ 	k	324 ± 10	334 ± 3	596 ± 495	335 ± 34
	error	0.93 ± 0.47	0.49 ± 0.22	24.4 ± 39.9	1.03 ± 1.51
	T	267465	$4/0$	1003	252657
	vpm	1516.0	268.6	1516.0	53.12
	time	23892.92	3.32	462.19	937.31
E: $\varepsilon = 0.92$, $N = 786$ 	k	685 ± 37	722 ± 0	620 ± 53	686 ± 36
	error	2.77 ± 6.43	0.29 ± 0.00	0.15 ± 0.03	0.47 ± 0.24
	T	14	$1/0$	1001	12
	vpm	786.0	675.7	786.0	327.7
	time	0.85	15.61	219.2	0.16

Table 2: Results for fundamental matrix estimation

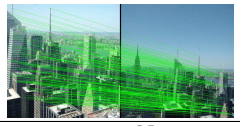
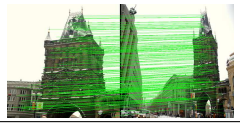
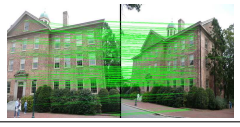
		RANSAC$_F$	USAC 1.0	AC-RANSAC	OURS
A: $\varepsilon = 0.35$, $N = 1207$ 	k	395 ± 16	418 ± 8	596 ± 10	619 ± 22
	error	2.21 ± 1.48	0.98 ± 0.54	0.91 ± 0.10	1.63 ± 0.96
	T	1321	$19/0$	1001	131
	vpm	1207.0	72.2	1207.0	62.9
	time	522.77	6.12	661.49	9.47
B: $\varepsilon = 0.65$, $N = 1110$ 	k	646 ± 18	713 ± 4	795 ± 11	828 ± 14
	error	2.94 ± 2.20	0.31 ± 0.24	0.63 ± 0.04	1.32 ± 0.76
	T	73	$12/0$	1001	18
	vpm	1110.0	94.5	1110.0	174.51
	time	24.1	5.21	450.90	2.61
C: $\varepsilon = 0.26$, $N = 2273$ 	k	537 ± 23	586 ± 5	541 ± 16	602 ± 60
	error	1.82 ± 0.57	1.08 ± 0.28	1.07 ± 0.15	2.46 ± 1.6
	T	6826	$35/0$	1002	4138
	vpm	2273.0	25.6	2273.0	56.18
	time	4100.09	17.81	1276.53	22.30

Table 3: Results for essential matrix estimation

since this dataset has only a 10% inlier ratio. Our method is the fastest for most datasets with running times comparable to USAC1.0 while automatically recovering σ.

Note that problems may arise with smaller sized datasets; as explained in Section 3, the relation between the critical value c and the type I error only holds asymptotically. Therefore, for a lower number of points, the critical value is more difficult to reach, potentially leading to an increase in type II error (rejection of valid models). However, in practice, datasets with as little as 500 data points did not present any problem for our method.

We also ran fundamental matrix estimation on three challenging optical flow datasets (taken from the robust vision challenge data [11]), where $656 \times 541 = 354896$ putative correspondences were determined by a fast PatchMatch [1] inspired optical flow method. Results obtained by RANSAC$_F$ and our method are presented in Table 5 (the statistics are the same as in the previous tables, time is measured in seconds). The images on the table illustrate (from left to right): the optical flow image, the inliers recovered

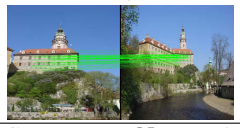
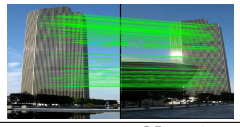
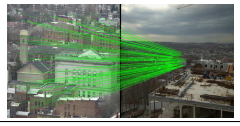
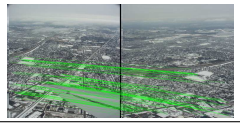
		RANSAC$_F$	USAC 1.0	AC-RANSAC	OURS
A: $\varepsilon = 0.46$, $N = 2540$ 	k	994 ± 68	1148 ± 2	1574 ± 13	1612 ± 39
	error	1.71 ± 0.21	1.04 ± 0.00	1.50 ± 0.15	2.32 ± 1.40
	T	220	2/0	1001	26
	vpm	2540.0	940.9	2540.0	284.1
	time	40.62	11.84	504.80	1.90
B: $\varepsilon = 0.15$, $N = 514$ 	k	70 ± 4	74 ± 3	71 ± 2	76 ± 12
	error	1.88 ± 0.68	1.19 ± 0.33	0.46 ± 0.25	2.56 ± 2.31
	T	16766	9/1	1001	13081
	vpm	514.0	110.4	514.0	22.0
	time	940.73	1.81	86.94	208.1
C: $\varepsilon = 0.23$, $N = 1317$ 	k	286 ± 17	302 ± 6	294 ± 16	318 ± 40
	error	1.63 ± 0.44	0.89 ± 0.13	0.44 ± 0.05	2.20 ± 2.23
	T	2433	9/1	1001	1688
	vpm	1317.0	374.1	1317.0	60.7
	time	254.62	1.26	219.77	3.14
D: $\varepsilon = 0.34$, $N = 495$ 	k	151 ± 11	168 ± 0	170 ± 2	173 ± 3
	error	2.22 ± 0.45	1.43 ± 0.00	0.23 ± 0.02	0.56 ± 0.61
	T	663	8/2	1001	327
	vpm	495.0	124.0	495.0	50.0
	time	36.00	6.13	83.39	6.54
E: $\varepsilon = 0.10$, $N = 994$ 	k	93 ± 6	99 ± 0	20 ± 13	80 ± 11
	error	3.43 ± 1.42	2.59 ± 0.27	0.13 ± 0.68	4.43 ± 1.60
	T	75950	7266/6511	1023	173192
	vpm	994.0	38.0	994.0	151.8
	time	6532.22	25.74	20.54	3616.11

Table 4: Results for homography estimation

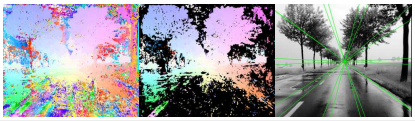
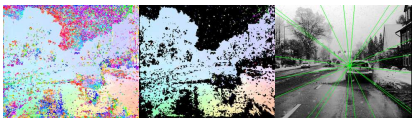
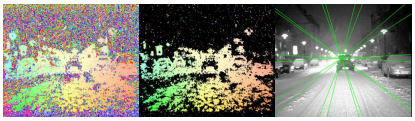
		RANSAC$_F$	OURS
A: $\varepsilon = 0.55$, $N = 354896$ 	k	182910 ± 4523	200997 ± 3995
	error	0.30 ± 0.03	0.48 ± 0.25
	T	514	384
	vpm	354896	3722.26
	time (s)	6.72	2.77
B: $\varepsilon = 0.54$, $N = 354896$ 	k	160666 ± 4856	192404 ± 5216
	error	0.22 ± 0.03	0.17 ± 0.04
	T	230	301
	vpm	354896	5031.4
	time (s)	3.16	2.40
C: $\varepsilon = 0.40$, $N = 354896$ 	k	115141 ± 2685	143607 ± 35489
	error	0.28 ± 0.04	1.72 ± 2.56
	T	1270	9733
	vpm	354896	205.2
	time (s)	130.89	54.39

Table 5: Results for fundamental matrix estimation using optical flow field

by our method, and the estimated epipolar geometry. Once again, automatic noise estimation allows for the recovery of a bigger inlier ratio. Our method achieves very good results in reasonable time while AC-RANSAC has an average running time of 200s on these datasets.

5. Conclusions

We introduced a new method for robust model estimation based on the likelihood-ratio test. Our approach is justified by a sound theoretical analysis and it doesn't require a-priori knowledge on the inliers noise level. We also proposed an early-bailout technique with statistical guarantees on the bounds for the error incurred. Tests on both synthetic and real data for a number of different estimation problems show that our method achieves state-of-the-art results while being faster than previously proposed methods. We plan to extend this work in order to account for multiple structures.

References

[1] C. Barnes, E. Shechtman, A. Finkelstein, and D. Goldman. Patchmatch: a randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics-TOG*, 28(3):24, 2009.

[2] D. P. Capel. An effective bail-out test for RANSAC consensus scoring. In *Proc. BMVC*, 2005.

[3] O. Chum and J. Matas. Randomized RANSAC with T(d, d) test. *Proc. BMVC*, 2:448–457, 2002.

[4] O. Chum and J. Matas. Optimal randomized RANSAC. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(8):1472–1482, 2008.

[5] C. Feng and Y. Hung. A robust method for estimating the fundamental matrix. In *DICTA*, pages 633–642. Citeseer, 2003.

[6] P. M. Fenwick. A new data structure for cumulative frequency tables. *Softw., Pract. Exper.*, 24(3):327–336, 1994. corrections: SPE 24(7): 677 (July 1994).

[7] M. Fischler and R. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[8] A. Konouchine, V. Gaganov, and V. Veznevets. AMLESAC: A new maximum likelihood robust estimator. In *Proc. Graphicon*, volume 5, pages 93–100, 2005.

[9] X. Li, J. Liu, and Z. Ying. Generalized sequential probability ratio test for separate families of hypotheses. *Sequential Analysis*, 33(4):539–563, 2014.

[10] J. Matas and O. Chum. Randomized RANSAC with sequential probability ratio test. In *Proc. ICCV*, volume 2, pages 1727–1732, 2005.

[11] S. Meister, B. Jähne, and D. Kondermann. Outdoor stereo camera system for the generation of real-world benchmark data sets. *Optical Engineering*, 51(02):021107, 2012.

[12] S. Mittal, S. Anand, and P. Meer. Generalized projection-based m-estimator. *TPAMI*, 34(12):2351–2364, 2012.

[13] L. Moisan and B. Stival. A probabilistic criterion to detect rigid point matches between two images and estimate the fundamental matrix. *IJCV*, 57(3):201–218, 2004.

[14] P. Moulon, P. Monasse, and R. Marlet. Adaptive structure from motion with a contrario model estimation. In *Proc. ACCV*, pages 257–270, 2013.

[15] R. Raguram, O. Chum, M. Pollefeys, J. Matas, and J. Frahm. USAC: a universal framework for random sample consensus. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):2022–2038, 2013.

[16] R. Raguram and J.-M. Frahm. RECON: Scale-adaptive robust estimation via residual consensus. In *Proc. ICCV*, 2011.

[17] P. J. Rousseeuw. Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880, 1984.

[18] S. Rozenfeld and I. Shimshoni. The modified pbm-estimator method and a runtime analysis technique for the ransac family. In *CVPR*, CVPR '05, pages 1113–1120, Washington, DC, USA, 2005. IEEE Computer Society.

[19] C. V. Stewart. MINPRAN: A new robust estimator for computer vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(10):925–938, 1995.

[20] R. Subbarao and P. Meer. Beyond ransac: User independent robust regression. In *Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop*, CVPRW '06, pages 101–, Washington, DC, USA, 2006. IEEE Computer Society.

[21] P. H. Torr and A. Zisserman. MLESAC: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, 78(1):138–156, 2000.

[22] H. Wang, D. Mirota, and G. D. Hager. A generalized kernel consensus-based robust estimator. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(1):178–184, 2010.

[23] H. Wang and D. Suter. Robust adaptive-scale parametric model estimation for computer vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(11):1459–1474, 2004.