FACULTY
OF ELECTRICAL
ENGINEERING
CTU IN PRAGUE

# HABILITATION THESIS

Jana Nosková

# Applied Statistics in Computer Vision and Cosmic Ray Physics

**Prague, November 2022**

Applied Statistics in Computer Vision and Cosmic Ray Physics
Habilitation Thesis
Prague, November 2022

Jana Nosková

Department of Mathematics
Faculty of Civil Engineering
Czech Technical University
Thákurova 7
166 29 Prague 6
Czech Republic

jana.noskova@cvut.cz
https://mat.fsv.cvut.cz/noskova/

# Contents

# 1 Introduction

This habilitation thesis is a collection of five articles [A01, A04, A06, A07, A08] I have contributed to, supplemented by a brief introduction. Specifically, together with my colleagues, I have been dealing with a robust model fitting of geometrical models in computer vision problems and with the visual object tracking [A01, A04, A06]. I have been also involved in the analysis of astroparticle data with the aim to identify cosmic ray sources [A07, A08]. These two fields of science are remarkably different from the point of view of an adoption of newly proposed uncertainty models. In computer vision, where we have at our disposal extremely large amount of cheap data and one can check models, we can choose the most successful one and so this model is quickly widely cited. In astroparticle physics, where the right answer is unknown and the data is rare and expensive, it takes quite a long time to adopt a new approach.

Early in my professional life I have realized that in spite of the fact that we have at our disposal a huge number of statistical procedures there arise problems for which no procedure suits well. If one has a good theoretical background in mathematical statistics, working on such problems could lead to novel or significantly modified statistical methods. Solving any such particular problem will be widely useful if this problem arises in applications in other fields of science.

About ten years ago, I have been invited to work at Visual Recognition Group of Department of Cybernetics CTU where I have become acquainted with current problems in computer vision. During this cooperation we have published several papers. Among them, our papers [A01, A04, A06], which form part of this work, address important issues for the computer vision community and my co-authors were mostly interested in. It turned out that mathematical statistics is a very suitable tool for solving such problems. To this end, I have proposed several mathematical methods.

First, I have introduced a novel loss function and its local optimization procedure to be used in RANSAC type procedures [A01]. I have also improved the mean-shift tracker in order to be scale adaptive [A04] and, finally, I have worked on the HMMTxD-tracker which uses the hidden Markov model for a fusion of multiple trackers [A06]. Main ideas of the mathematical methods and suggested algorithms, which are my own activities in these projects, are described in Sections 2, 3 and 4. These methods were utilized to create computer vision algorithms with which many different experiments with large data sets were successfully performed. Nonetheless the computer code implementation and subsequent experiments were fully provided by my co-authors.

It is worth noting that the predecessor of our article [A04] presented in a conference proceedings [A05] was awarded the best paper prize at the Scandinavian Conference on Image Analysis 2013; both these studies are widely cited. Similarly, the predecessors of Ref. [A01], i.e. conference contributions [A02] and [A03] that were published at the Conference on Computer Vision and Pattern Recognition

2019 and 2020, respectively, are also widely cited and, moreover, they are frequently used outside the computer vision community e.g. in geoscience and remote sensing, robotics, data mining and knowledge discovery or neurocomputing.

More than fifteen years ago I have started my cooperation with the astrophysics group at the Faculty of Mathematics and Physics. Thanks to the unprecedented development of the physics of cosmic rays in the last two decades, many interesting data on their arrival directions have been collected, which should indicate the hitherto unknown sources of this extremely energetic radiation. However, the detected signals are very weak and the sources that produce them cannot be conclusively identified.

In several internal technical reports, which are not publicly available, we dealt with the way how to confirm a weak sources within the framework of the standard statistical method, known in the astroparticle physics as an on-off problem. It turned out that the optimal method of analysis is the Bayesian approach, which does not need to rely on the asymptotic behavior of the statistics used and can advantageously utilize previously acquired knowledge about the observed phenomena. I have proposed a method based on Bayesian reasoning that allowed us to estimate the significance of the tested source of detected events [A07]. We showed in details its connection to commonly used procedures. In a subsequent study [A08], other relevant quantities were introduced to demonstrate the usefulness of the Bayesian approach in the field of cosmic ray physics. In particular, relying upon previous observation, we dealt with the waiting time for the next observation in a counting experiment within the Bayesian settings. We also discussed the problem how to compare significances of sources when obtained in different observations carried out in different experiments.

In all these activities, I mainly proposed solution methods and oversaw the correct statistical interpretation of the obtained results. Work with the data, numerical calculations as well as the final physical interpretation were performed by my coauthors. A brief introduction to the on-off problem is given in Section 5.

# 2 Robust model fitting in computer vision

Our study presented in Ref. [A01] is introduced in the following. Section 2.1 starts with a general description of robust methods. After a brief summary of robust estimators, M-estimators, in Section 2.1.1, we introduce the iteratively reweighted least squares (IRLS) method for their solving in Section 2.1.2. Robust estimators in computer vision are discussed in Section 2.2. The relationship of M-estimators to computer vision RANSAC-type estimators is mentioned in Section 2.2.1. Main results of our study [A01] are highlighted in Section 2.2.2.

## 2.1 Robust statistics

Advanced statistical methods rely on probability models containing some theoretical assumptions. The most widely used methods, e.g. the least squares method (LSM), assume that the observed data have a normal (Gaussian) distribution. These methods, called classical statistical methods, were used in statistics at least for two centuries. It was understood that the statistical model is just an approximation of reality. Then, one expects that outputs of such models are also approximately correct. This is unfortunately not true. Even a single atypical observation, called an outlier, can spoil the output provided by a classical statistical method.

Robust statistics derives methods which produce reliable outputs also when its theoretical assumptions hold only approximately. If the data contain no outlier, robust methods give practically the same results as the classical methods. In case of presence of outliers they fit the bulk of the data and can be used for outlier detection.

Attempts to make statistical methods more robust were made at least at the nineteenth century. The fundaments of robust statistics were set by John Tukey, Peter Huber and Frank Hampel only in the 60s and early 70s of the last century. It has to be mentioned that many Czech statisticians contributed not only to further development of robust statistics but also have formed its fundaments[1)] and helped improve its reputation.[2)]

### 2.1.1 M-estimators

Currently popular robust estimators, M-estimators where “M”stands for “maximum likelihood-type”estimators, were proposed by Huber in 1964 [1]. They are a gener-

[1] From this point of view, let us mention the important visit of prof. Jaroslav Hájek, the former head of Department of Probability and Statistics at Charles University in Prague, to University of California at Berkeley in 1961-62, where at the same time Peter Huber was at his post-doc stay. Hájek’s doctoral student, current professor emeritus at Charles University prof. Jana Jurečková, belongs to co-founders of robust statistics.

[2] For example, 10th International Conference on Robust Statistics in 2010 was held in Prague due to international eminence and influence of prof. Jana Jurečková. At this conference, beside other, the victory of all robust statisticians was mentioned because robust statistics has become a part of mainstream statistics and robust procedures were incorporated in many commercial softwares.

alization of the maximum likelihood estimators (MLE).

The same notation as in Refs. [A01, A02, A03] is adopted. Let us denote

$\mathcal{P} = \{p | p \in \mathbb{R}^\nu, \nu \in \mathbb{N}\}$ the set of observed data,

$\Theta = \{\theta | \theta \in \mathbb{R}^d, d \in \mathbb{N}\}$ the set of possible parameters of our model and

$R : \Theta \times \mathcal{P} \longrightarrow \mathbb{R}$ the residual function of the model with a parameter $\theta$ and an observation $p$.

An M-estimator is defined as follows [2]:

**Definition 1**

A $\rho$-function will denote a function $\rho$ such that

1. $\rho(x)$ is a non-decreasing function of $|x|$,
2. $\rho(0) = 0$,
3. $\rho(x)$ is increasing for all $x > 0$ for which $\rho(x) < \rho(\infty)$,
4. if $\rho$ is bounded $\rho(\infty) = 1$,

then the M-estimator of parameter $\theta$ is

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \sum_{p \in \mathcal{P}} \rho(R(\theta, p)). \tag{1}$$

MLE is a special type of M-estimator with $\rho$-function equivalent to the minus log-likelihood function. For example, MLE for a mean of normal distribution is an M-estimator with quadratic $\rho$-function, i.e. L2 estimator. Looking for robust estimators under presence of large amount of outliers, M-estimators with bounded $\rho$-function are of a special interest. Such estimators cannot be formulated as MLE for any density function.

**M-estimators of location**

In this case the data $p$ and the model parameter $\theta$ are scalars and the residual function is [2]

$$R(\theta, p) = (p - \theta)/\sigma, \tag{2}$$

where the parameter $\sigma$ is a scale of data $\mathcal{P}$, $\sigma > 0$. Unfortunately, M-estimators of location with bounded $\rho$-function depend on the choice of the parameter $\sigma$ which is usually unknown. Thus, some proper robust estimate of $\sigma$ has to be used instead, e.g. the normalized median of absolute deviation about median. There are no explicit expressions for the distribution of M-estimators of location in finite sample sizes. However, under quite general conditions, these M-estimators are asymptotically normal in the same sense as MLEs and M-estimators with bounded $\rho$-function can cope merely with 50% of outliers in a data sample.

**M-estimators in linear model**

Here, we assume that the data $p$ and the model parameter $\theta$ to be vectors of real numbers [2]

$$p = (y(p), x(p)), x(p) = (x_1(p), x_2(p), \dots, x_{\nu-1}(p)), \quad p \in \mathbb{R}^\nu, \tag{3}$$

$$\theta = (\theta_1, \theta_2, \dots, \theta_{\nu-1}), \quad \theta \in \mathbb{R}^{\nu-1}, \tag{4}$$

and the residual function

$$R(\theta, p) = (y(p) - x(p)\theta^T)/\sigma, \tag{5}$$

where the parameter $\sigma$ is a scale of data $\mathcal{P}$, $\sigma > 0$ and here $T$ means transposition.

Similarly as for location in previous example, also M-estimators in linear model with bounded $\rho$-function depend on the choice of the parameter $\sigma$ and, therefore, some proper robust estimate of $\sigma$ has to be used. Again, there are no explicit expressions for the distribution of M-estimators in linear model in finite sample sizes. Nonetheless, under quite general conditions, they are asymptotically normal in the same sense as MLEs. In case of fixed or bounded vectors $x(p)$, M-estimators with bounded $\rho$-function can cope merely with 50% of outliers in a data sample.

### 2.1.2 Iteratively reweighted least squares

For computing M-estimates several function minimization methods can be used. Iteratively reweighted least squares (IRLS) method is the recommended one [2].

**Definition 2**

If there exist the first derivative of a $\rho$-function, $\rho'$, and the second derivative at 0, $\rho''(0)$, a weight $w$-function is a function $w$ such that $w(0) = \rho''(0)$ and $w(x) = \rho'(x)/x$ otherwise.

**IRLS algorithm**

1. Compute an initial estimate $\theta_0$ and some robust estimate of $\sigma$.
2. For $k = 0, 1, 2, \dots$ let $\theta_{k+1} = \underset{\theta \in \Theta}{\arg\min} \sum_{p \in \mathcal{P}} w(R(\theta_k, p)) R^2(\theta, p)$.

3. Stop when $\max_{p\in\mathcal{P}} |(R(\theta_k, p) - (R(\theta_{k+1}, p)|$ falls below some predefined threshold.

When the weight function $w(x)$ is non-increasing in $|x|$, function

$$L(\theta) = \sum_{p\in\mathcal{P}} \rho(R(\theta, p)), \tag{6}$$

non-increases in each iteration of IRLS algorithm. For convex $L(\theta)$ function, IRLS attains its minimum. However, for bounded and thus non-convex $\rho$-functions, $L(\theta)$ is not convex. It was shown [2] that in linear model IRLS converges to its local minimum. In any case the proper choice of the initial estimator is important. The recommended initial estimator for $\theta_0$ is L1 estimator. For the parameter $\sigma$ it is the normalized median of $|R(\theta_0, p)|, p \in \mathcal{P}$.

Let us note that it is not necessary for $\theta_{k+1}$ to minimize $\sum_{p\in\mathcal{P}} w(R(\theta_k, p))R^2(\theta, p)$ with respect to $\theta$. It is sufficient to decrease the sum, i.e.

$$\sum_{p\in\mathcal{P}} w(R(\theta_k, p))R^2(\theta_{k+1}, p) < \sum_{p\in\mathcal{P}} w(R(\theta_k, p))R^2(\theta_k, p). \tag{7}$$

**Commonly used $\rho$-functions**

The commonly used $\rho$-functions with their $w$-functions are depicted in Fig.1. These functions are defined as follows

1. L2 estimator: $\rho(x) = x^2/2$ and $w(x) = 1$.

2. L1 estimator: $\rho(x) = |x|$ and $w(x) = 1/|x|$.

3. Huber estimator:
$\rho(x) = x^2/2$ for $|x| < 1$, $\rho(x) = |x| - 1/2$ otherwise and
$w(x) = 1$ for $|x| < 1$, $w(x) = 1/|x|$ otherwise.

4. Welsch estimator: $\rho(x) = 1 - \exp(-x^2)$ and $w(x) = 2\exp(-x^2)$.

5. Hampel type estimator:
$\rho(x) = \frac{9}{4}x^2$ for $|x| < 1/3$, $\rho(x) = \frac{3}{2}|x| - \frac{1}{4}$ for $1/3 \le |x| \le 2/3$,
$\rho(x) = -\frac{9}{4}x^2 + \frac{9}{2}|x| - \frac{5}{4}$ for $2/3 \le |x| \le 1$ and $\rho(x) = 1$ otherwise.
The corresponding weight function is $w(x) = \frac{9}{2}$, $w(x) = \frac{3}{2|x|}$, $w(x) = \frac{9}{2}(\frac{1}{|x|} - 1)$ and $w(x) = 0$.

6. Tukey estimator:
$\rho(x) = 1 - (1 - x^2)^3$ for $|x| < 1$, $\rho(x) = 1$ otherwise and
$w(x) = 6(1 - x^2)^2$ for $|x| < 1$, $w(x) = 0$ otherwise.

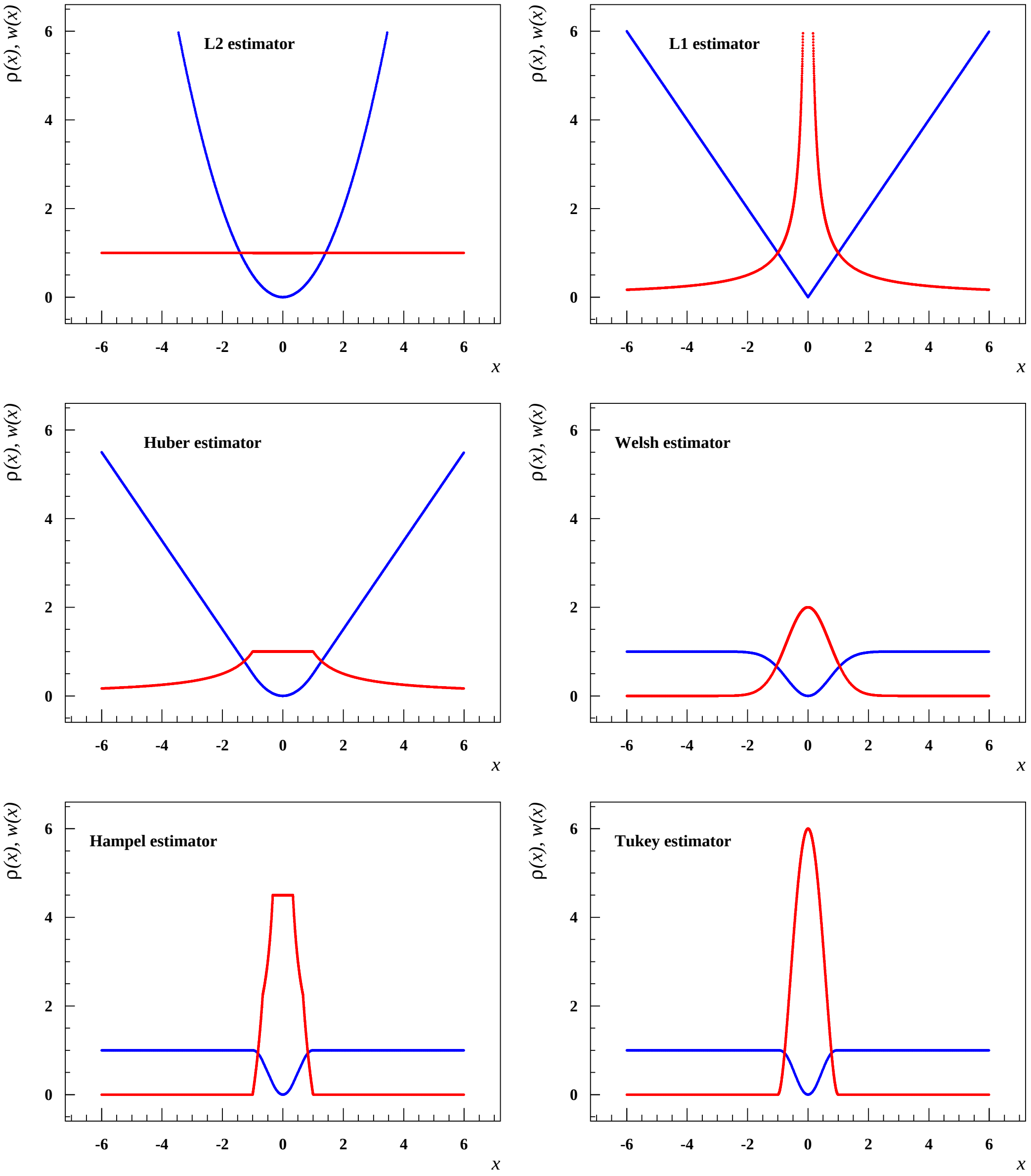


Figure 1: **Commonly used M-estimators.** $\rho$-functions (blue) and $w$-functions (red) of L2, L1, Huber, Welsh, Hampel and Tukey estimators are depicted (from top to bottom).

It is documented in Fig.1 that L2, L1 and Huber M-estimators have unbounded $\rho$-functions. Welsch, Hampel and Tukey M-estimators have bounded $\rho$-functions. Only $w$-functions of Hampel and Tukey estimators are equal to zero for $|x| > 1$. Vanishing $w$-function outside some bounded interval is important for IRLS procedure, since usually only a small part of data is involved in each iteration.

## 2.2 Robust estimation in computer vision

Many problems that require robust estimation arise in computer vision e.g. fundamental matrix, homography and essential matrix fitting. Unfortunately, in these tasks M-estimators or any other results of robust statistics cannot be applied directly. There are two main reasons why.

The first one is that the ratio of the number of outliers in the data is very often larger than 50%. In statistics, our model is assumed to describe most of the data, thus, it cannot cope with more than 50% of outliers. In case of a bounded $\rho$-function one expects that M-estimators could work under a vague assumption that outliers "do not cooperate"with each other. Then, there are no reliable initial estimators for $\theta_0$ and $\sigma$, however.

The second reason is the form of the residual function $R(\theta, p)$. Different computer vision problems use different residual functions but generally they are not linear functions of parameters $\theta$ and cannot be simply linearized. It implies that one has to give up attempts to obtain asymptotic distribution of the M-estimator $\hat{\theta}$.

### 2.2.1 RANSAC-type estimators

In 1981 Fischler and Bolles proposed a new paradigm for model fitting the RANdom SAmple Cosensus (RANSAC) [3], which has become a widely used robust estimator not only in computer vision. The basic idea is to repeatedly randomly select a minimal sample of $m$ points from $\mathcal{P}$. These points are required to compute a candidate for parameter $\theta$. Then one calculates residua $R(\theta, p)$ of all points $p \in \mathcal{P}$ and the model determined by this candidate and counts all points $p$ which are closer to this model than some predefined threshold $\sigma$. These close points are called inliers. Finally, after completing all these loops, the candidate with the largest number of inliers determines the estimate of the parameter $\theta$.

#### The sampling algorithm

Since choosing all minimal samples of $m$ points from $\mathcal{P}$ is usually intractable we need some stopping criterion. Suppose our data contain a portion $\epsilon \in (0, 1)$ of outliers, the probability of all-inlier sample is $(1-\epsilon)^m$ and the probability of at least one all-inlier sample in $k$ samples is

$$1 - (1 - (1 - \epsilon)^m)^k. \tag{8}$$

If one wants this probability to be larger than some predefined confidence probability $\mu \in (0, 1)$, the number of chosen samples $k$ has to be

$$k \geq \frac{\ln(1 - \mu)}{\ln(1 - (1 - \epsilon)^m)}. \tag{9}$$

The sampling algorithm finds the initial estimate $\theta_0$. But the problem of a reliable estimate of $\sigma$ remains.

Let us finally note that the same sampling algorithm with the same stopping criterion was proposed by Leroy and Rousseeuw in their technical report in 1984 [4].

**$\rho$-functions of RANSAC-type estimators**

RANSAC can be formulated as an M-estimator with $\rho$-function

$$\rho(x) = 0 \quad \text{for} \quad |x| < 1 \quad \text{and} \quad \rho(x) = 1 \quad \text{otherwise.} \tag{10}$$

RANSAC is extremely sensitive to a choice of a parameter $\sigma$. Since its weight function

$$w(x) = 0 \quad \text{for} \quad x \in \mathbb{R}, \quad |x| \neq 1, \tag{11}$$

IRLS cannot be used for a local optimization to get an estimate of $\theta$.

Torr and Zisserman proposed an approach [5] where other bounded $\rho$-functions are used within a sampling algorithm. They preferred a $\rho$-function [5]

$$\rho(x) = x^2 \quad \text{for} \quad |x| < 1 \quad \text{and} \quad \rho(x) = 1 \quad \text{otherwise,} \tag{12}$$

An estimator with this $\rho$-function is known as MSAC. Its weight function

$$w(x) = 2 \quad \text{for} \quad |x| < 1 \quad \text{and} \quad w(x) = 0 \quad \text{otherwise,} \tag{13}$$

therefore, IRLS method could be used.

Indeed, although it is not directly claimed, local optimization of MSAC in sense of IRLS is in fact used in Refs. [6] and [7], where a new estimator called LO-RANSAC is proposed. LO-RANSAC is less sensitive to a choice of a threshold $\sigma$ than pure RANSAC or MSAC. Nevertheless, a proper choice of $\sigma$ is still a hard and crucial problem.

For the sake of completeness, RANSAC and MSAC $\rho$-functions and $w$-functions are depicted in Fig.2.

### 2.2.2 Marginalizing sample consensus

In our study [A01], we proposed a novel class of $\rho$-functions, see also Ref. [A03]. This class of $\rho$-functions is parameterized by the dimension of the data $\nu$, with a gradually decreasing $w$-functions vanishing outside some bounded interval. M-estimators generated by these $\rho$-functions are called MAGSAC++. In this case IRLS method can be used for a local optimization to obtain an M-estimator of the parameter $\theta$. MAGSAC++ $\rho$-functions and $w$-functions for $\nu = 2, 3, 4, 5$ are depicted in Fig.3.

Many minimal solvers for most computer vision tasks were proposed, among other reasons, due to the 40-year popularity of RANSAC. They are necessary for computation of a candidate for parameter $\theta$ using just several points from $\mathcal{P}$, at least some minimal sample of $m$ points from $\mathcal{P}$, when looking for a solution in the

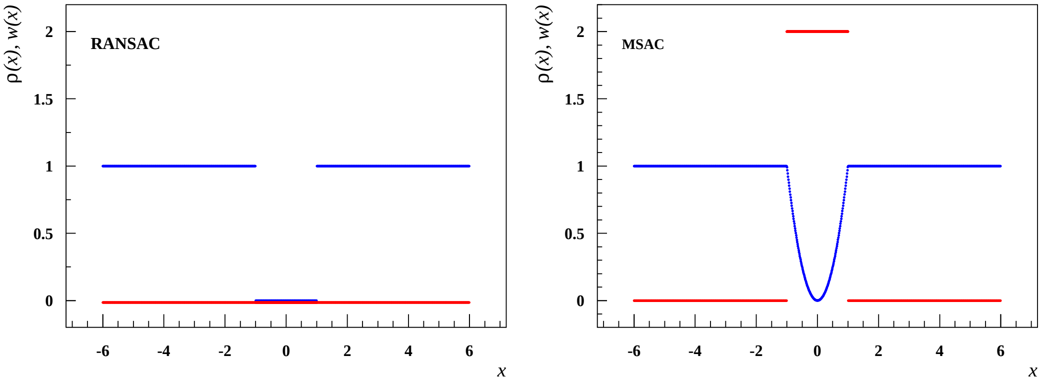


Figure 2: **M-estimators used in RANSAC.** $\rho$-functions (blue) and $w$-functions (red) of RANSAC and MSAC estimators are visualized.

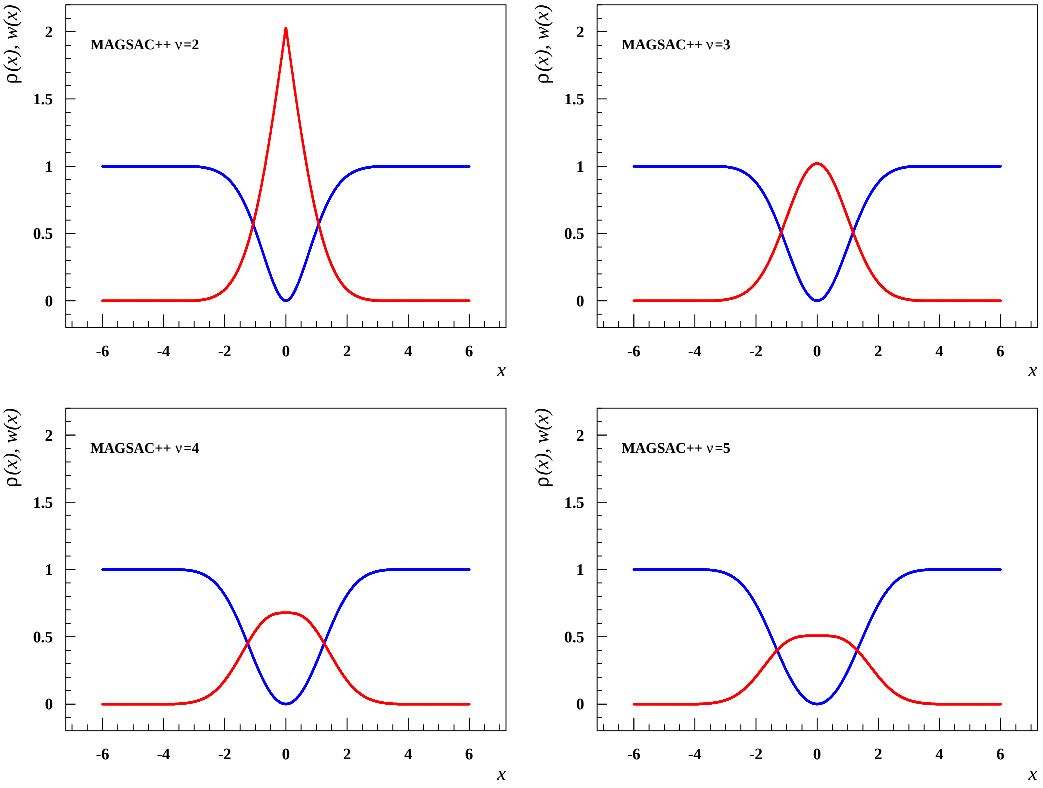


Figure 3: **MAGSAC++ M-estimators.** $\rho$-functions (blue) and $w$-functions (red) of MAGSAC++ for $\nu = 2, 3, 4$ and 5 versions are depicted.

inner loop of RANSAC, see the introductory text to Section 2.2.1. Each computer vision task utilizes its own residual function $R(\theta, p)$. Its application to a minimal sample leads to a system of polynomial equations in $\theta$ with some constrains which has to be solved. For different computer vision tasks, different systems of equations are derived, see the overview of minimal solvers in Ref. [8]. These minimal solvers can be directly used for the IRLS in the MAGSAC++ loops, see the introductory text in Section 2.2.1.

For many computer vision tasks, the parameter $\sigma$ in MAGSAC++ can be chosen from a wide range of values while having practically no impact on the resulting estimate [A01, A03]. In this case, the IRLS method usually converges in a few iterations.

# 3 Mean-shift tracker

Visual object tracking is a computer vision task of automatically identifying objects in video sequences. The mean-shift tracker was proposed for real-time tracking of non-rigid objects by Comaniciu, Ramesh and Meer at Conference on Computer Vision and Pattern Recognition 2000 [9]. In 2010, this paper was awarded the Longuet-Higgins Prize for Computer Vision and Pattern Recognition papers from ten years ago that have made a significant impact on computer vision research.

The mean-shift tracker is a short-term tracker of a single target and the only information about the target provided is its bounding box in the first frame. It tracks by minimizing the distance between a target RGB color histogram and RGB color histograms of target candidates. The mean-shift algorithm is used for that minimization. The procedure is recapitulated in Section 3.1 followed by the description of the mean-shift algorithm and its relationship to iteratively reweighted least squares (IRLS) in Section 3.2. The mean-shift tracker with a variable bandwidth proposed in our study [A04] is commented in Section 3.3.

## 3.1 Mean-shift procedure

The mean-shift is used in an iterative procedure for seeking modes of kernel estimates of a density function. It was proposed by Fukunaga and Hostetler in 1975 [10]. This study was spread at computer vision community by Comaniciu and Meer in 2002 [11].

**Kernel estimator of density function**

Let $K : \mathbb{R} \to \mathbb{R}$ be a kernel function, i.e. a non-negative symmetric function with

$$\int_{-\infty}^{+\infty} K(u)\ \mathrm{d}u = 1, \tag{14}$$

then

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{15}$$

is a kernel density estimator of a sample $\{x_i \in \mathbb{R}, 1 \leq i \leq n\}$ at $x \in \mathbb{R}$ with a bandwidth $h > 0$.

One possible way how to define a multivariate kernel density estimator in $d$-dimensional space $\mathbb{R}^d$ using one-dimensional kernel function is an introduction of the so called profile $k$ of a kernel $K$ [9]:

$$k(x^2) = K(x). \tag{16}$$

Let $\mathbf{x}_i \in \mathbb{R}^d$ for $1 \leq i \leq n, \| \ \mathbf{x}\|^2 = \ \mathbf{x}^T\mathbf{x}$. Then

$$\hat{f}(\mathbf{x}) = \frac{1}{cnh^d} \sum_{i=1}^{n} k\left( \| \frac{\mathbf{x} - \mathbf{x}_i}{h} \|^2 \right) \tag{17}$$

is a kernel density estimator of a sample $\{\mathbf{x}_i \in \mathbb{R}^d, 1 \leq i \leq n\}$ at $\mathbf{x} \in \mathbb{R}^d$ with a bandwidth $h > 0$ and

$$c = \int\limits_{-\infty}^{+\infty} k(\|\mathbf{v}\|^2) \ \mathrm{d}\mathbf{v}, \tag{18}$$

for $\mathbf{v} \in \mathbb{R}^d$.

**Mean-shift**

Assuming that the first derivative $k'$ of a kernel profile $k$ exists and $g(x) = -k'(x)$,

$$\mathbf{m}(\mathbf{x}) = \frac{\sum\limits_{i=1}^{n} \mathbf{x}_i g\left( \| \frac{\mathbf{x} - \mathbf{x}_i}{h} \|^2 \right)}{\sum\limits_{i=1}^{n} g\left( \| \frac{\mathbf{x} - \mathbf{x}_i}{h} \|^2 \right)} - \mathbf{x}, \tag{19}$$

is the mean-shift at $\mathbf{x}$. The mean-shift vector $\mathbf{m}(\mathbf{x})$ is the normalized gradient of the kernel density estimator obtained with kernel $K$ at $\mathbf{x}$. When searching for modes of the kernel density estimator $\hat{f}(\mathbf{x})$, $\mathbf{x}$ with zero gradient is looked for.

### 3.2 Mean-shift algorithm and IRLS

The mean-shift algorithm works as follows:

1. Let $\mathbf{y}_0$ be the initial point in $\mathbb{R}^d$.
2. For $j = 0, 1, 2, \ldots$, let $\mathbf{y}_{j+1} = \mathbf{y}_j + \mathbf{m}(\mathbf{y}_j) = \frac{\sum_{i=1}^{n} \mathbf{x}_i g\left( \| \frac{\mathbf{y}_j - \mathbf{x}_i}{h} \|^2 \right)}{\sum\limits_{i=1}^{n} g\left( \| \frac{\mathbf{y}_j - \mathbf{x}_i}{h} \|^2 \right)}$.
3. Stop when $\|\mathbf{y}_{j+1} - \mathbf{y}_j\|$ falls below some predefined threshold.

If the kernel $K$ has a convex and monotonically decreasing profile $k$, the sequences $\{\mathbf{y}_j\}_{j=1,2,\ldots}$ and $\{\hat{f}(\mathbf{y}_j)\}_{j=1,2,\ldots}$ converge, and $\{\hat{f}(\mathbf{y}_j)\}_{j=1,2,\ldots}$ is monotonically increasing. Then the sequence $\{\mathbf{y}_j\}_{j=1,2,\ldots}$ converges to some mode of $\hat{f}(\mathbf{y})$, see Ref. [9].

Any kernel $K$ non-increasing in $|x|$ can be written as

$$K(x) = C(1 - \rho(x)), \tag{20}$$

where the function $\rho$ is a bounded $\rho$-function and $C$ is a positive normalization constant.

Therefore, the mean-shift algorithm for kernel $K$ is IRLS algorithm with the bounded $\rho$-function given in Eq.(20) and the residual function

$$R(\theta, \mathbf{x}) = \frac{\|\mathbf{x} - \theta\|}{h}. \tag{21}$$

The $d$-dimensional vector

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} \rho(R(\theta, \mathbf{x}_i)), \tag{22}$$

is a mode of the kernel density estimator with the kernel $K$. Following Eq.(20), corresponding $w$-function, $w(x) = \frac{2}{C} g(x^2)$, see Section 2.1.2, is non-negative and non-increasing for non-increasing and convex kernel profile $k$. The $j$-th iteration of IRLS

$$\theta_{j+1} = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} w(\|\frac{\theta_j - \mathbf{x}_i}{h}\|)\|\frac{\theta - \mathbf{x}_i}{h}\|^2, \tag{23}$$

leads to the weighted mean

$$\theta_{j+1} = \frac{\sum\limits_{i=1}^{n} \mathbf{x}_i g\left(\|\frac{\theta_j - \mathbf{x}_i}{h}\|^2\right)}{\sum\limits_{i=1}^{n} g\left(\|\frac{\theta_j - \mathbf{x}_i}{h}\|^2\right)}. \tag{24}$$

In fact, the mean-shift procedure is equivalent to IRLS method used for an M-estimation of location in a $d$-dimensional space, see Section 2.1.1.

For example, using a uniform kernel for seeking mode is equivalent to the estimation of location with RANSAC $\rho$-function. Similarly, using Epanechnikov kernel is equivalent to MSAC $\rho$-function, triweight kernel corresponds to Tukey $\rho$-function, Gaussian kernel is equivalent to Welsch $\rho$-function and cosine kernel corresponds to Andrew $\rho$-function.

### 3.3 Mean-shift tracker with adaptive scale

The mean-shift tracker suffers from the use of fixed size bounding box if the scale of the target in the video sequence changes. In our study [A04], we proposed the mean-shift tracker that uses kernel profile $k$ with a variable bandwidth $h$. In our approach, the minimization of the distance between a target RGB color histogram and RGB color histograms of target candidates is done not only with respect to the target candidate position but also with respect to its scale $h$. The minimization uses gradient descent method and utilizes the fact that the mean-shift is a normalized gradient of the minimized distance.

In Ref. [A04], we introduce a tracking algorithm using the mean-shift procedure with a variable bandwidth $h$ and call it an adaptive scale mean-shift (ASMS) tracker. In this study, ASMS is compared to state-of-the-art algorithms on a large tracking data set. For example, one of used criterion for comparison is the number of frames the algorithm locates the target correctly, the other is its speed, for other criteria see Ref. [A04]. We observed that in all of these criteria ASMS equals or outperforms the state-of-the-art algorithms.

# 4 Multiple trackers and estimation of their confidence

A large number of diverse tracking methods has been proposed based on different assumption about the target motion, adopted features and optimization techniques. For example, some trackers assume a rigid motion, i.e. the motion preserves the Euclidean distance between every pair of target points. Some trackers use a deep neural network for finding the most similar target candidate to the target in each frame. The mean-shift tracker utilizes the color similarity of the target and a target candidate.

As the situation changes during the video sequence, different trackers are better suited for tracking under different conditions. The natural idea is to exploit multiple different trackers at the same time, in each video frame estimate their probabilities of tracking correctly and then use the most confident ones, i.e. those with the highest estimated probability of being correct. In our study [A06], the hidden Markov model (HMM) is utilized for this estimation and fusion of multiple trackers. This model is briefly introduced in Section 4.1. In addition, an expectation-maximization (EM) algorithm that is commonly used in HMM and its connection to IRLS are mentioned in Section 4.2. Finally, our study [A06] is commented in Section 4.3.

## 4.1 Hidden Markov model and IRLS

Let us assume the HMM with $N$ possible states $\{s_1, s_2, \ldots, s_N\}$, the matrix of state transition probabilities $A = \{a_{ij}\}_{i,j=1}^{N}$, the vector of initial probabilities $\pi = (\pi_1, \pi_2, \ldots, \pi_N)$, a sequence of observations $\mathcal{X} = \{X_t\}_{t=1}^{T}$, $X_t \in \mathbb{R}^m$ and $F = \{f_i(x)\}_{i=1}^{N}$, $x \in \mathbb{R}^m$, using the system of conditional probability densities of observations conditioned on $S_t = s_i$ written

$$f_i(x) = f(x|S_t = s_i) \quad \text{for} \quad 1 \leq i \leq N, 1 \leq t \leq T, x \in \mathbb{R}^m, \tag{25}$$

where $S_t$ are random variables, each representing the hidden state at time $t$, and $\lambda = (A, F, \pi)$ denotes the parameter set of the HMM. Having the observations $\mathcal{X}$, maximum likelihood approach is used to estimate the parameters $\lambda$.

### Maximum likelihood in HMM

Let $\mathcal{S} = \{s_1, s_2, \ldots, s_N\}^T$ be a set of all possible $T$-tuples of states and let $s^* = (s_1^*, s_2^*, \ldots, s_T^*) \in \mathcal{S}$ be one possible sequence of states. Then the likelihood function is

$$P(\mathcal{X}|\lambda) = \sum_{s^* \in \mathcal{S}} P(s^*, \mathcal{X}|\lambda), \tag{26}$$

where

$$P(s^*, \mathcal{X}|\lambda) = \pi_{s_1^*}(\lambda) f_{s_1^*}(\lambda, X_1) \prod_{t=2}^{T} a_{s_{t-1,t}^*}(\lambda) f_{s_t^*}(\lambda, X_t). \tag{27}$$

Maximizing the likelihood function $P(\mathcal{X}|\lambda)$ is a complicated task that usually cannot be solved analytically. In 1970, Baum, Petrie, Soules and Weiss proposed an iterative procedure for maximization of $P(\mathcal{X}|\lambda)$ which is known as Baum-Welch algorithm [13]. The idea is as follows. Let $\lambda \in \Lambda$, where $\Lambda$ is a subset of Euclidean space, and

$$Q(\lambda, \lambda') = \sum_{s^* \in \mathcal{S}} P(s^*|\mathcal{X}, \lambda) \ln[P(s^*, \mathcal{X}|\lambda')]. \tag{28}$$

Then, according to Theorem 2.1 in Ref. [13]

$$Q(\lambda, \lambda') \geq Q(\lambda, \lambda) \Rightarrow P(\mathcal{X}|\lambda') \geq P(\mathcal{X}|\lambda) \tag{29}$$

and the equality holds if and only if $P(s^*|\mathcal{X}, \lambda) = P(s^*|\mathcal{X}, \lambda')$ for $\forall s^* \in \mathcal{S}$.

Here we add that the form of the function $P(s^*, \mathcal{X}|\lambda)$ is not crucial for the inequality (29). Hence, the Baum-Welch iterative procedure introduced below can be used for any hidden variables with finite possible outcomes.

The classical Baum-Welch algorithm [13] repeats the second step in the scheme listed below until convergence:

1. Let $\lambda_0$ be the initial estimate of $\lambda$.
2. For $j = 0, 1, 2, \dots$ let $\lambda_{j+1} = \arg\max_{\lambda} Q(\lambda_j, \lambda)$.

In this way, each step of Baum-Welch algorithm non-decreases the likelihood function $P(\mathcal{X}|\lambda)$. Moreover, its convergence to the likelihood function maximum depends on the choice of the type of density functions $f_i(x)$ for $1 \leq i \leq N$. Specifically, Baum-Welch algorithm converges for the Poisson, binomial, normal and gamma distributions and does not converge for the Cauchy distribution [13].

For discrete distributions $f_i(x) \in (0, 1)$, for $1 \leq i \leq N$, giving $\ln[P(s^*, \mathcal{X}|\lambda)] < 0$ for $\forall s^* \in \mathcal{S}$, the Baum-Welch algorithm can be formulated as IRLS procedure: Let the residual function

$$R(\lambda, s^*) = \sqrt{-\ln[P(s^*, \mathcal{X}|\lambda)]}. \tag{30}$$

Using Welsch $\rho$-function, $\rho(x) = 1 - \exp(-x^2)$ for $x \in \mathbb{R}$, with the $w$-function, $w(x) = 2\exp(-x^2)$, the second step of IRLS algorithm, see Section 2.1.2, gives for $j = 0, 1, 2, \dots$

$$\begin{aligned} \lambda_{j+1} &= \arg\min_{\lambda} \sum_{s^* \in \mathcal{S}} w(R(\lambda_j, s^*)) R^2(\lambda, s^*) \\ &= \arg\min_{\lambda} (-2) \sum_{s^* \in \mathcal{S}} P(s^*, \mathcal{X}|\lambda_j) \ln[P(s^*, \mathcal{X}|\lambda)] \\ &= 2P(\mathcal{X}|\lambda_j) \arg\max_{\lambda} Q(\lambda_j, \lambda), \end{aligned} \tag{31}$$

and is equivalent to the second step of Baum-Welch algorithm mentioned above.

In exactly the same way, one can easily deduce that for continuous bounded densities $f_i(x) \le M$, $\forall x \in \mathbb{R}^m$, $1 \le i \le N$ the Baum-Welch algorithm can be formulated as IRLS procedure with the residual function

$$R(\lambda, s^*) = \sqrt{-\ln \frac{P(s^*, \mathcal{X}|\lambda)}{M^T}}. \tag{32}$$

Let us finally note that the well-known expectation-maximization algorithm proposed by Dempster, Laitd and Rubin in 1977 [14] and briefly mentioned in Section 4.2 is, in fact, a generalization of the Baum-Welch algorithm.

### 4.2 Expectation-maximization algorithm and IRLS

In this section, we remind the expectation-maximization (EM) algorithm and show its connection to IRLS.

Let $\mathcal{X} = \{X_i\}_{i=1}^N$, $X_i \in \mathbb{R}^m$ be observed variables and $\mathcal{Y} = \{Y_i\}_{i=1}^K$, $Y_i \in \mathbb{R}^k$ be unobserved (hidden) variables. In EM algorithm, $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$ is called complete data, $\mathcal{X}$ denotes incomplete data. Let $\theta \in \Theta$, where $\Theta$ is a subset of Euclidean space, and let

1. $f(\mathcal{Z}|\theta) = f(\mathcal{X}, \mathcal{Y}|\theta)$ be a joined density of $\mathcal{X}$ and $\mathcal{Y}$,
2. $g(\mathcal{Y}|\mathcal{X}, \theta)$ be a conditional density of $\mathcal{Y} \in \mathcal{R} = \mathbb{R}^{kK}$ conditioned on $\mathcal{X}$ and
3. $l(\mathcal{X}|\theta)$ be a marginal density of $\mathcal{X}$.

We want to find a MLE of $\theta$, i.e.

$$\hat{\theta} = \arg\max_{\theta \in \Theta} l(\mathcal{X}|\theta). \tag{33}$$

The EM algorithm repeats the second step of the following prescription until convergence:

1. Let $\theta_0$ be the initial estimate of $\theta$.
2. For $j = 0, 1, 2, \ldots$ let $\theta_{j+1} = \arg\max\limits_{\theta \in \Theta} \int\limits_{\mathcal{R}} \ln(f(\mathcal{X}, \mathcal{Y}|\theta))\, g(\mathcal{Y}|\mathcal{X}, \theta_j)\, \mathrm{d}\mathcal{Y}$.

Doing this procedure, each step of EM algorithm non-decreases the likelihood function $l(\mathcal{X}|\lambda)$.

The connection to IRLS follows. Let

$$R(\theta, \mathcal{Z}) = R(\theta, \mathcal{X}, \mathcal{Y}) \tag{34}$$

be a residual function and $\rho(x)$ be a bounded $\rho$-function, with $w$-function $w(x)$, and

$$L(\theta) = \int\limits_{\mathcal{R}} [1 - \rho(R(\theta, \mathcal{X}, \mathcal{Y}))]\, \mathrm{d}\mathcal{Y}, \tag{35}$$

be bounded with respect to $\theta \in \Theta$. Let us propose a slightly generalized IRLS algorithm:

1. Let $\theta_0$ be the initial estimate of $\theta$.
2. For $j = 0, 1, 2, \ldots$ let $\theta_{j+1} = \arg\min\limits_{\theta \in \Theta} \int\limits_{\mathcal{R}} w(R(\theta_j, \mathcal{X}, \mathcal{Y}))\, R^2(\theta, \mathcal{X}, \mathcal{Y})\, \mathrm{d}\mathcal{Y}$.

When the function $w(x)$ is non-increasing in $|x|$ each iteration of this algorithm non-decreases the function $L(\theta)$ in Eq.(35).

Let $f(\mathcal{X}, \mathcal{Y}|\theta) \leq M$ for all $\mathcal{Y} \in \mathcal{R}$. Let $\theta \in \Theta$ and, again, the residual function

$$R(\theta, \mathcal{X}, \mathcal{Y}) = \sqrt{- \ln \frac{f(\mathcal{X}, \mathcal{Y}|\theta)}{M}}. \tag{36}$$

Using Welsch $\rho$-function, $\rho(x) = 1 - \exp(-x^2)$ for $x \in \mathbb{R}$ with the corresponding $w$-function $w(x) = 2\exp(-x^2)$, see Section 2.1.2, the second step of the generalized IRLS algorithm gives for $j = 0, 1, 2, \ldots$

$$\begin{aligned}
\theta_{j+1} &= \arg\min_{\theta \in \Theta} \int\limits_{\mathcal{R}} w(R(\theta_j, \mathcal{X}, \mathcal{Y}))\, R^2(\theta, \mathcal{X}, \mathcal{Y})\, \mathrm{d}\mathcal{Y} \\
&= \arg\min_{\theta \in \Theta} (-2) \int\limits_{\mathcal{R}} \ln[f(\mathcal{X}, \mathcal{Y}|\theta)] f(\mathcal{X}, \mathcal{Y}|\theta_j)\, \mathrm{d}\mathcal{Y} + 2 \ln M \int\limits_{\mathcal{R}} f(\mathcal{X}, \mathcal{Y}|\theta_j)\, \mathrm{d}\mathcal{Y} \\
&= 2l(\mathcal{X}|\theta_j) \arg\max_{\theta \in \Theta} \int\limits_{\mathcal{R}} \ln(f(\mathcal{X}, \mathcal{Y}|\theta))\, g(\mathcal{Y}|\mathcal{X}, \theta_j)\, \mathrm{d}\mathcal{Y},
\end{aligned} \tag{37}$$

i.e. we end up with the second step of the EM algorithm and with $L(\theta) = \frac{1}{M} l(\mathcal{X}|\theta)$.

### 4.3 Online adaptive hidden Markov model for multi-tracker fusion

In our study [A06], we use two or three trackers and a detector. For each video frame each tracker outputs a target candidate and some observables. The detector outputs the verified target pose or nothing. The chosen states of HMM are the correctnesses of individual trackers that are mostly hidden, but they are known if the detector fires. The observables are similarity measures of the estimated target candidates to target bounding box in the first frame. These observables are in the interval $(0, 1)$ or can be normalized in such a way that they range in the interval $(0, 1)$. We assume that the observations follow a beta distribution [A06].

In the maximization step of Baum-Welch algorithm [13], $Q(\lambda_j, \lambda)$ given in Eq.(28) splits up into two summands which can be maximized separately, the first with respect to the state transition probabilities $A = \{a_{ij}\}_{i,j=1}^{N}$ and the second with respect to the parameters of the beta distributions. However, maximization of the beta distribution with respect to its parameters is a complicated task. Therefore, we modify the second step of Baum-Welch algorithm [A06] as follows. The function $Q(\lambda_j, \lambda)$ is maximized only with respect to the probabilities $A = \{a_{ij}\}_{i,j=1}^{N}$. The parameters of the beta distributions are estimated by the method of moments. Due to the method of moments, $Q(\lambda_j, \lambda_{j+1}) \geq Q(\lambda_j, \lambda_j)$ is not guaranteed, however. If $\lambda_{j+1}$ does not satisfy this inequality the $j$-th iteration changes only estimates of the state transition probabilities and the parameters of the beta distributions stay unchanged, implying that $Q(\lambda_j, \lambda_{j+1}) \geq Q(\lambda_j, \lambda_j)$ holds.

In Ref. [A06], we propose a novel algorithm called HMMTxD (hidden Markov model for trackers and detector) for fusion of multiple trackers. We choose fast tracking methods that have different designs and work with different assumptions. One of used trackers is our ASMS tracker [A04]. We have shown that superior performance can be achieved by using simple trackers that may not represent the state-of-the-art. The HMMTxD method achieves the performance of at least the best tracker used or higher and shows the efficiency of the HMM for combination of multiple trackers. The HMMTxD method outperforms the state-of-the-art, often significantly, on many data-sets in almost all criteria.

# 5 On-off problem

In the following, we describe our activities in the development of statistical methods for searching for new phenomena using data that consists of a set of discrete events providing a possible signal that is, in principle, indistinguishable from the background. The on-off experiment and its interpretation is described in Section 5.1. The classical solution to the on-off problem is briefly mentioned in Section 5.2. In Section 5.3, we introduce Bayesian approach to the on-off task that was worked out by us in [A07, A08]. We show our solutions to the on-off problem obtained in term of a difference of unknown on- and off-source intensities [A07] in Section 5.3.1. In order to illustrate pros and cons of Bayesian method we worked out several numerical examples presenting them in Section 5.3.2. Based on our study [A08], in Section 5.3.3, other interesting statistical variables related to the on-off problem within Bayesian settings are mentioned and their usefulness is briefly commented.

## 5.1 Signal detection

The on-off problem arises when measured data consists of two unknown parts, typically in the search for new effects in particle physics or in high-energy astrophysics. The first set of data is due to a signal searched for, and the second one is due to an inseparable background. Thus, the on-off experiment is designed for counting two classes of events registered in two disjoint regions, in the on-source region, where a new phenomenon is searched for, and in the reference off-source region, where only background events contribute.

Specifically, one wants to decide whether the same emitter with a constant but unknown intensity is responsible for the observed counts in both regions or whether a source producing more events in the on-source region is present. The on-off counts, $n_{\text{on}}$ and $n_{\text{off}}$ , are assumed to follow independent Poisson distributions with unknown positive intensities, $\mu_{\text{on}} > 0$ and $\mu_{\text{off}} > 0$. Naturally, exposures of the on- and off-source regions under considerations, $w_{\text{on}}$ and $w_{\text{off}}$ , play the role. It is assumed that their ratio $\alpha = \frac{w_{\text{on}}}{w_{\text{off}}}$ is known from the experimental details or is pre-estimated. The unknown intensity of background counts in the on-source region is derived from the intensity in the off-source region, i.e. $\mu_{\text{b}} = \alpha\mu_{\text{off}}$ .

Apart from less used statistical concept of $p$-values, physicists work mostly with a derived statistics, in a concept of significance $S$. While the $p$-value denotes the probability of obtaining test results at least as extreme as the observed one under the assumption that the hypothesis is true, the significance $S$ is the upper $p$-value quantile of the standard normal distribution, $\Phi^{-1}(S) = 1 - p$, where $\Phi$ is the cumulative distribution function of standard normal distribution. For example, $S = 3$ corresponds to the $p$-value equal to 0.00135.

### 5.2 Li-Ma method and Wilks' theorem

An important paper dealing with the on-off problem was written by Li and Ma in 1983 [15]. It summarizes and criticizes different approaches for analyzing $\gamma$-ray astronomy experiments used so far and introduces a novel method for testing the hypothesis of equality of on- and off-source intensities, $\mu_{\text{on}} = \alpha\mu_{\text{off}}$ . Their approach is based on the likelihood ratio test proposed by Wilks [16]. Famous Wilks' theorem described therein was presented to the American Mathematical Society on 26th March 1937 saying that the asymptotic distribution of a logarithm of likelihood ratio multiplied by $(-2)$ is $\chi^2$.

Monte Carlo simulations presented in Li-Ma paper [15] and subsequent estimation of significances show that, in the case that the observed counts are not too few (say $n_{\text{on}} \gg 10, n_{\text{off}} \gg 10$) and for a reasonable ratio of on- and off-source exposures, i.e. $\alpha \in (0.1, 10)$, the likelihood ratio asymptotics works better than approaches used previously. In the following ten years, Li-Ma formula [15] became famous within a physics community. Nowadays, it is a standard widely used tool claiming detection and setting its significances in different applications.

### 5.3 Bayesian approach

In case the measured events are extremely rare, Li-Ma formula cannot be properly used for its only asymptotic validity. This occurs, for example, in searching for weak signals from distance galaxies in high-energy astrophysics or in searching for still unobserved particles created in unusual interactions in particle physics. It was our motivation for proposing a Bayesian solution of the on-off problem that employs the same set of parameters, namely, the number of registered events in the on-source region, $n_{\text{on}}$ , the number of detected events in the complemented off-source region, $n_{\text{off}}$ , and the ratio of on- and off-source exposures, $\alpha = \frac{w_{\text{on}}}{w_{\text{off}}}$.

With this method, it is possible to infer the signal significance, strength and uncertainty of the measured signal and its upper limit, all in a straightforward way. Bayesian approach is valid without restrictions for count numbers. On the other hand, it relies on the prior choice of parameter distribution. Nevertheless, their prior distributions can be chosen completely uninformative or, on contrary, it can contain information from previous experiments if available. Bayesian method can be utilized in different physics applications, in high-energy gamma astrophysics, cosmic-ray physics or in particle physics, for example.

#### 5.3.1 On Bayesian analysis of on–off measurements

We focused on reformulating the on-off problem within Bayesian settings [A07]. The unknown intensities in the on- and off-source regions, $\mu_{\text{on}}$ and $\mu_{\text{off}}$ , were considered to be independent random variables with different prior gamma distributions, i.e. the conjugate priors for a Poisson distribution. Having a posterior distribution of

$\mu_{\text{on}}$ given a number of events $n_{\text{on}}$ and a posterior distribution of $\mu_{\text{off}}$ given $n_{\text{off}}$ as well, we introduce a difference variable $\delta = \mu_{\text{on}} - \alpha\mu_{\text{off}}$ and construct its posterior distribution given $n_{\text{on}}$ and $n_{\text{off}}$ . This variable estimates the strength of emitter if the source counts are measured in the on-source region. Let us note that $\delta \in$ R, $\delta > 0$ indicate a source is present in the on-source region and $\delta \leq 0$ represents an absence of a source or a sink of events in the on-source region or a possible source in the off-source region.

We analyzed on-off data with the aim to confirm whether an emitter is responsible for observed counts. To this end, we calculated the probability $P^+ = P(\delta > 0)$ and define the Bayesian significance of a source presence in the on-source region by $S_{\text{B}} = \Phi^{-1}(P^+)$, noting that $S_{\text{B}} < 0$ speaks for possible presence of a source in the off-source region or disappearance of events in the on-source area. On top of this, also credible intervals for $\delta$ are constructed in positively identified cases since due to Bayesian approach we have at our disposal not only the significance but the whole distribution of difference variable $\delta$. However, the distribution of $\delta$ has a complicated structure containing Tricomi confluent hypergeometric function [A07]. We showed that if it is guaranteed that a source may be observed just in the on-source region or if the background intensity $\mu_{\text{off}}$ can be assumed to be known from other considerations, the posterior distribution of the difference variable $\delta$ is simply adaptable.

### 5.3.2 Monte Carlo simulations

In order to compare Bayesian and Li-Ma significances, we performed a set of Monte Carlo simulations. According to Li-Ma method [15], for a likelihood ratio $\lambda$ generated by two Poissonian variables $n_{\text{on}}$ and $n_{\text{off}}$ observed in the regions with the ratio of exposures $\alpha = \frac{w_{\text{on}}}{w_{\text{off}}}$ one has

$$\lambda = \left(\frac{\alpha}{1+\alpha}\frac{n_{\text{on}}+n_{\text{off}}}{n_{\text{on}}}\right)^{n_{\text{on}}} \left(\frac{1}{1+\alpha}\frac{n_{\text{on}}+n_{\text{off}}}{n_{\text{off}}}\right)^{n_{\text{off}}}, \tag{38}$$

and $-2\ln\lambda$ asymptotically follows the $\chi^2$-distribution with one degree of freedom under a no-source hypotheses. The Li-Ma significance $S_{\text{LM}}$ is then given by

$$S_{\text{LM}} = \sqrt{-2\ln\lambda} \text{ for } n_{\text{on}} \geq \alpha n_{\text{off}}, \text{ and } S_{\text{LM}} = -\sqrt{-2\ln\lambda} \text{ for } n_{\text{on}} < \alpha n_{\text{off}}. \tag{39}$$

For Bayesian significance $S_{\text{B}}$ we employed distribution of the difference variable $\delta$ as indicated in Section 5.3.1, for more details see [A07].

In order to illustrate usefulness of Bayesian approach, here we worked out several simple examples. Li-Ma significance $S_{\text{LM}}$ is compared to Bayesian significance $S_{\text{B}}$ gained with a totally uninformative prior, i.e. with an improper prior equal to 1 on $(0, +\infty)$. In each Monte Carlo simulation, $10^5$ pairs of independent Poisson random variables with intensities $\mu_{\text{on}}$ and $\mu_{\text{off}}$ were generated. Resultant histograms of significances $S_{\text{LM}}$ and $S_{\text{B}}$ with prescribed values of $\alpha$ are depicted in Figs.4 and 5.

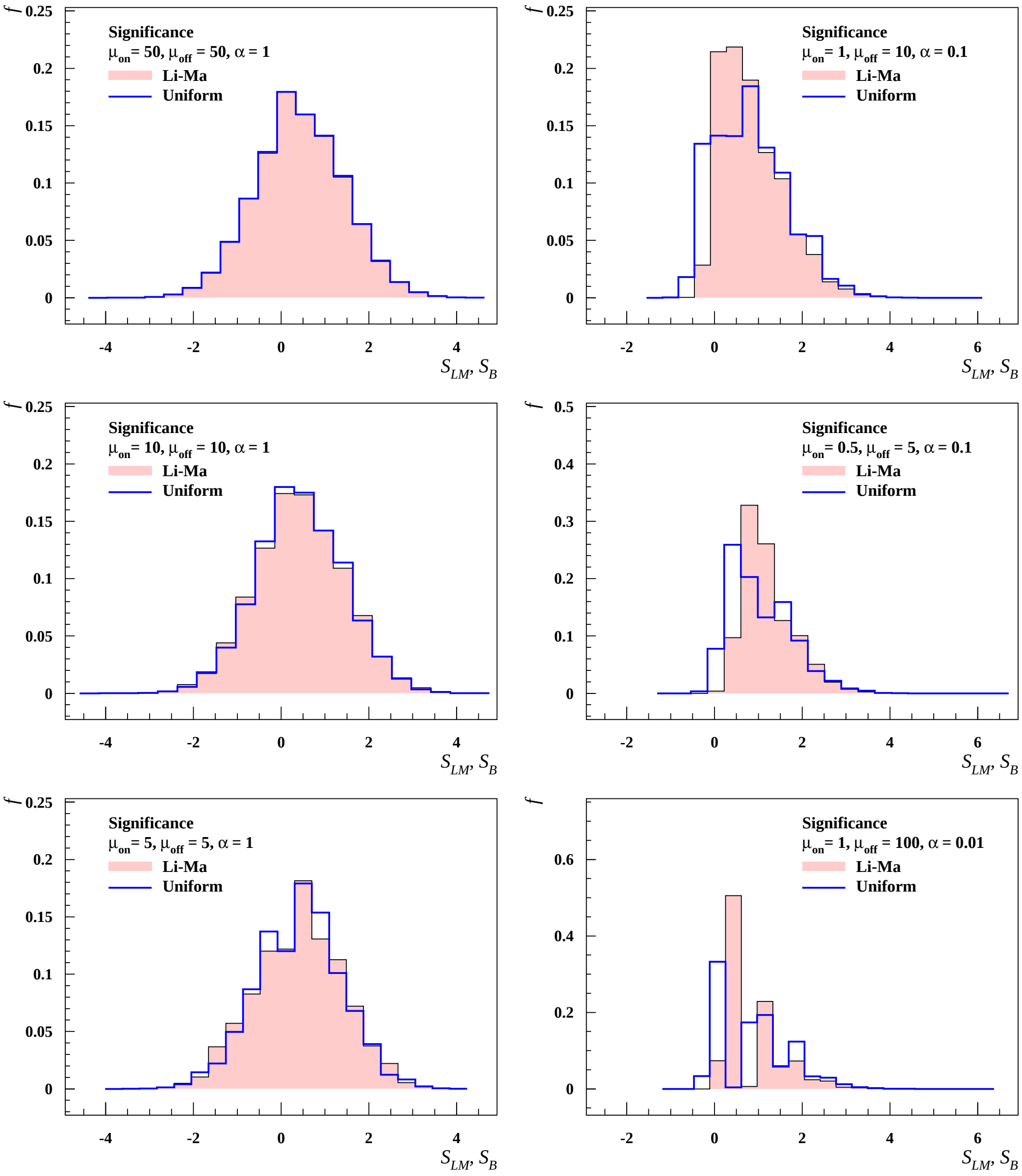


Figure 4: Distributions of significances for the source detection. We show histograms for the Li–Ma significances (red filled areas) and for the Bayes significances (blue lines) using uniform prior distributions. Mean parameters are indicated in the panels.

There is a no source present in the on-source region in examples depicted in Fig.4. It can be seen that for $\mu_{\rm on} = 50$, $\mu_{\rm off} = 50$ and $\alpha = 1$ the histograms of Li-Ma and Bayesian significance are almost the same. Decreasing $\mu_{\rm on}$ , $\mu_{\rm off}$ and $\alpha$ the significance histograms become more different, however. In case of rejecting a no-source

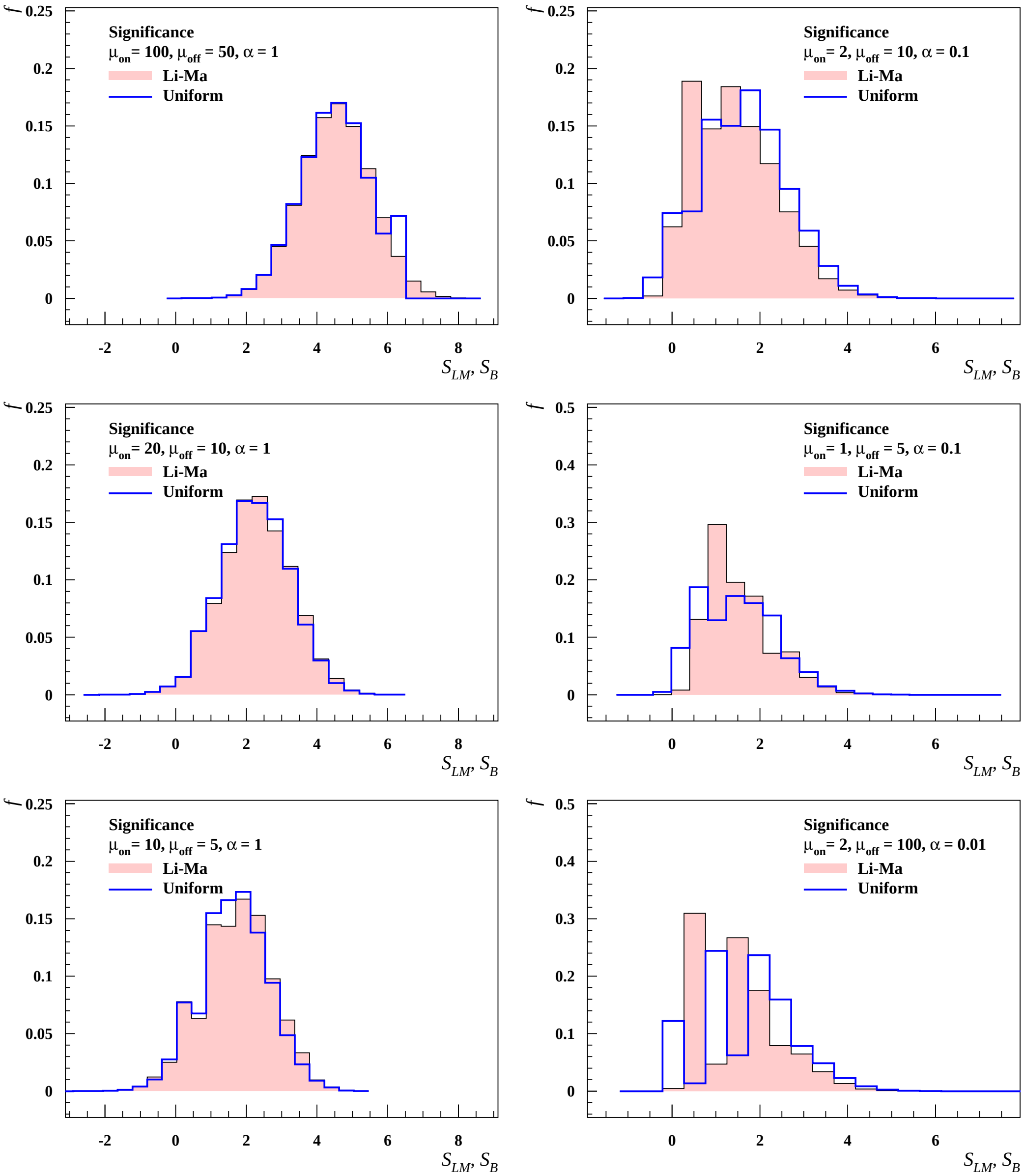


Figure 5: Same as in Fig.4. Here, Li–Ma and Bayes significances using $\mu_{\rm on} = 2\alpha\mu_{\rm off}$ are shown.

hypotheses, if the significance is larger then 3, using $S_{\rm LM}$ or $S_{\rm B}$ leads practically to the same frequencies of disclaims.

Both types of resultant significances for a source present in the on-source region with $\mu_{\rm on} = 2\alpha\mu_{\rm off}$ are shown in Fig.5. Interestingly, for $\alpha < 1$ as it is common in all experiments design for cosmic-ray studies, assuming significance $S_{\rm B}$ within Bayesian approach, no-source hypotheses are rejected more often then using Li-Ma method and corresponding significance $S_{\rm LM}$ , see the rightmost panels in Fig.5.

#### 5.3.3 A Bayesian on-off analysis of cosmic ray data

In the following study [A08], we deal with posterior distributions of some other variables closely related to source identification that are suitable in directional analysis in cosmic-ray physics. Namely, we consider the source flux $j = \delta/a$ , where $a = \frac{\alpha}{\alpha+1}A$ and $A$ denotes the exposure in the on-source region integrated over the period of data taking. We also assumed the intensity registered in the on-source region expressed it terms of a background intensity, $\beta = \frac{\mu_{\rm on}}{\alpha\mu_{\rm off}}$. Finally, we deal with the fraction of the total intensity registered in the on-source zone, $\omega = \frac{\mu_{\rm on}}{\mu_{\rm on} + \mu_{\rm off}}$.

Having in disposal Bayesian posterior distributions of different variables, we focused on predicting waiting time for new events in the on-source region expressed as a total count of events registered in both regions. To this end, we introduce two independent Poisson processes $\{N_{\rm on}(t) : t \geq 0\}$ and $\{N_{\rm off}(t) : t \geq 0\}$ with intensities $\mu_{\rm on}$ and $\mu_{\rm off}$ and expected values $\mu_{\rm on}\, t$ and $\mu_{\rm off}\, t$, respectively. Here, the random variable $N_{\rm on}(t)$ represents new on-source events and the random variable $N_{\rm off}(t)$ stands for new off-source events collected up to and including time $t$. Then, based on previous observations, the random variable $N_{\rm on}(t)$ conditioned on $n = N_{\rm on}(t) + N_{\rm off}(t), n \in$ N, follows a binomial distribution with parameters $n$ and $\omega$ introduced above for any $t > 0$. This result can be used for checking whether new observations are consistent with the previous ones or, on the other hand, whether the intensities of observed processes have changed [A08].

Finally, we also address a question of how to compare two independent on-off measurements [A08]. With the Bayesian posterior distributions of different variables, we were able to quantify statistically which of the measurements indicate a more intense emitter. For this, we choose the unconditional distributions of the difference variable. Then, we calculated the probability that the flux observed in one experiment is less than the flux measured in the second one, both fluxes treated as random variables. Similar results are obtained with other variable related to source identification, for more details see [A08].

# 6 Summary

In this habilitation thesis I specified my contribution to statistical approaches applied in the field of computer vision and cosmic ray physics.

In computer vision I have been working on several different projects. First, I dealt with the RANdom SAmpling Consensus (RANSAC), a popular computer vision method which offers a simple way of fitting parameterized models to data corrupted by outliers. This method is highly sensitive to the user-selected threshold that divides data into inliers and outliers. My idea was not to strictly divide data into inliers and outliers but to assign to each point a weight corresponding to its likelihood to be inlier. This idea led to a novel loss function and the corresponding local optimization procedure MAGSAC++ [A01], a new type of an M-estimator. We have shown, in many large experiments, that MAGSAC++ leads to the most accurate estimate of relative position, significantly smaller sensitivity to the setting of the threshold parameter and similar speed in comparison with state-of-the-art methods. MAGSAC++ is the best RANSAC-type algorithm for challenging problems with ratio of outliers highly above 50% [17].

Second, I dealt with visual object trackers. Visual object tracking is an important research topic in computer vision, where given the initial state of a target i.e. its center and location in the first frame of a video sequence, the aim of tracking is to automatically obtain the positions of the object in the subsequent video frames. Our contribution was an improvement of the mean-shift tracker to be scale adaptive resulting in the adaptive scale mean-shift (ASMS) tracker [A04]. The performance of the mean-shift tracker suffers from the use of a fixed size window if the scale of the target changes which often leads to tracking failure. I proposed the theoretical background which was implemented in ASMS tracker. Although ASMS tracker slows down the standard mean-shift tracker, it is significantly faster than state-of-the-art algorithms which can cope similarly successfully with changing scale of a target during video sequence.

My second post on visual object tracking is a method utilized in the implementation of the HMMTxD-tracker (hidden Markov model for trackers and detector) proposed for cooperation of multiple trackers based on different and complementary assumptions [A06]. The estimation of parameters of the hidden Markov model is done online using partially annotated states by the object detector and utilizing a modified Baum-Welch algorithm. The goal is to use the best performing tracker in each frame. HMMTxD-tracker was among the top three performing trackers at the time of its publication, with speed comparable to other complex tracking methods. Moreover, I have shown here that all the optimization procedures used in the above algorithms for computer vision tasks can be formulated as iteratively reweighted least squares.

In cosmic ray physics, I worked on the Bayesian approach to the on-off problem [A07, A08] that arises when directional data are searched for possible sources the activity of which is submerged in a surrounding background. This scheme is particularly suitable when observing extremely rare events. Bayesian reasoning allows us to gain more statistical characteristics suitable for the study of counting processes, which are widely used in astroparticle physics, but are difficult to obtain in the classical concept. Using distributions of different quantities, we were concern, for example, with confidence intervals of a signal or its upper limits, estimates for the waiting time for the next event and comparison of different on-off measurements. Moreover, we focused on the precise interpretation of the results derived under the validity of various simplifying assumptions.

We also presented several numerical examples that may serve as guides for practical applications. First, a Monte Carlo simulation study comparing Bayesian and classical results was performed at Section 5.3.2. Second, we successfully used Bayesian inference in order to interpret experimental data on very high energy photons from the on-off measurements of gamma ray bursts [A07]. Finally, with the aim to document the use and advantages of the Bayesian reasoning, we dealt with the cosmic ray data collected in currently working observatories [A08]. We have summarized the outputs available in this concept by looking at these data if, as originally suggested, they are associated with a set of positions of active galactic nuclei and the nearest one, Centaurus A.

## 7 Acknowledgments

I am indebted to many individuals over the last ten years for realizing the importance of statistical methods in their field of science and for including me to their scientific team. I thank Jiří Matas who was so courageous to offered me a position of a researcher at the Center for Machine Perception CTU, where I met only enthusiastic people, who helped me to become acquainted with problems in computer vision. I would like express my thanks to all of them, especially to my co-authors Tomáš Vojíř and Daniel Baráth. The papers included in this thesis would not have come into being without their strong interest in it. I also thank Dalibor Nosek who introduced me to the problem of data analysis in cosmic ray physics and spent a lot of time with me in fruitful discussion.

This work was supported by the Czech Science Foundation Project GAČR P103/12/G084, Center for Large Scale Multi-modal Data Interpretation, and by the Ministry of Education OP VVV project CZ.02.1.01/0.0/0.0/16_019/0000765, Research Center for Informatics.

## References

[1] Huber P. J., *Robust estimation of a location parameter*, Annals of Mathematical Statistics 35, pp.73-101, 1964.

[2] Maronna R. A., Martin R. D., Yohai V. J., *Robust Statistics Theory and Methods*, John Wiley @ Sons, Ltd, 2006.

[3] Fischler M. A., Bolles R. C., *Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography*, Communications of the ACM, Vol. 24, pp.381-395, June 1981.

[4] Leroy, A., Rousseeuw, P. J., *PROGRESS: A Program for Robust Regression Analysis*, Technical Report 201, Center for Statistics and O.R., Universiry of Brussels, Belgium, 1984.

[5] Torr P. H. S., Zisserman A., *MLESAC: A new robust estimator with application to estimating image geometry*, Computer Vision and Image Understanding, Volume 78, Issue 1, pp.138-156, April 2000.

[6] Chum O., Matas J., Kittler J., *Locally optimized RANSAC*, Joint Pattern Recognition Symposium, pp.236-243, 2003.

[7] Lebeda K., Chum O., Matas J., *Fixing the Locally Optimized RANSAC*, British machine vision conference. Vol. 2., 2012.

[8] Pajdla T., Kukelova Z., *Minimal Problems in Computer Vision*, http://aag.ciirc.cvut.cz/minimal/, 2022.

[9] Comaniciu D., Ramesh V., Meer P., *Real-Time Tracking of Non-Rigide Objects Using Mean Shift*, Computer Vision and Pattern Recognition Conference Proceedings, vol.2, pp.142-149, 2000.

[10] Fukunaga K., Hostetler L. D., *The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition*, IEEE Transactions on Information Theory 21(1), pp.32-40, January 1975.

[11] Comaniciu D., Meer P., *Mean Shift: A Robust Approach Toward Feature Space Analysis*, IEEE Transactions on Pattern Analysis and Machine Intelligenece 24(5), pp.603-619, May 2002.

[12] Comaniciu D., Visvanathan R., Meer P. *Kernel-based Object Tracking*, IEEE Transactions on Pattern Analysis and Machine Intelligenece 25(5), pp.564-575, May 2003.

[13] Baum L. E., Petrie T., Soules G., Weiss N., *A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains*, Annals of Mathematical Statistics, Vol.41, No.1, pp.164-171, 1970.

[14] Dempster A. P., Laitd N. M., Rubin D. B., *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society, series B 39(1), pp.1-38, 1977.

[15] Li T. P., Ma Y. Q., *Analysis methods for results in gamma-ray astronomy*, Astrophysical Journal, 272, pp.317-324, 1983.

[16] Wilks S. S., *The large-sample distribution of the likelihood ratio for testing composite hypotheses*, Annals of Mathematical Statistics 9(1), pp.60-62, 1938.

[17] Riu C., Nozick V., Monasse P., Dehais J., *Classification Performance of RanSaC Algorithms with Automatic Threshold Estimation*, Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, pp.723-733, 2022.

## Selected publications by the author

[A01] Barath D., Noskova J., Matas J., *Marginalizing Sample Consensus*, IEEE Transactions on Pattern Analysis and Machine Intelligence 44, pp. 8420-8432, November 2022.

[A02] Barath D., Matas J., Noskova J., *MAGSAC: Marginalizing Sample Consensus*, Proceedings of Conference on Computer Vision and Pattern Recognition, USA: IEEE, pp. 10197-10205, 2019.

[A03] Barath D., Noskova J., Ivashechkin M., Matas J., *MAGSAC++, a Fast, Reliable and Accurate Robust Estimator*, Proceedings of Conference on Computer Vision and Pattern Recognition, USA: IEEE, pp. 1304-1312, 2020.

[A04] Vojir T., Noskova J., Matas J., *Robust scale-adaptive mean-shift for tracking*, Pattern Recognition Letters 49, pp. 250-258, 2014.

[A05] Vojir T., Noskova, J., Matas, J., *Robust Scale-Adaptive Mean-Shift for Tracking*, Proceedings of the 18th Scandinavian Conference on Image Analysis, Lecture Notes in Computer Science, pp. 652-663, 2013.

[A06] Vojir T., Matas J., Noskova, J., *Online adaptive hidden Markov model for multi-tracker fusion*, Computer Vision and Image Understanding 153, pp. 109-119, 2016.

[A07] Nosek D., Noskova J., *On Bayesian analysis of on-off measurements*, Nuclear Instruments and Methods in Physics Research A 820, pp. 23-33, 2016.

[A08] Nosek D., Noskova J., *A Bayesian on-off analysis of cosmic ray data*, Nuclear Instruments and Methods in Physics Research A 867, pp. 222-230, 2017.

# A Marginalizing Sample Consensus

Barath D., Noskova J., Matas J., *Marginalizing Sample Consensus*, IEEE Transactions on Pattern Analysis and Machine Intelligence 44, pp. 8420-8432, November 2022.
DOI: 10.1109/TPAMI.2021.3103562

# Marginalizing Sample Consensus

Daniel Barath, Jana Noskova, and Jiri Matas

**Abstract**—A new method for robust estimation, MAGSAC++, is proposed. It introduces a new model quality (scoring) function that does not make inlier-outlier decisions, and a novel marginalization procedure formulated as an M-estimation with a novel class of M-estimators (a robust kernel) solved by an iteratively re-weighted least squares procedure. Instead of the inlier-outlier threshold, it requires only its loose upper bound which can be chosen from a significantly wider range. Also, we propose a new termination criterion and a technique for selecting a set of inliers in a data-driven manner as a post-processing step after the robust estimation finishes. On a number of publicly available real-world datasets for homography, fundamental matrix fitting and relative pose, MAGSAC++ produces results superior to the state-of-the-art robust methods. It is more geometrically accurate, fails fewer times, and it is often faster. It is shown that MAGSAC++ is significantly less sensitive to the setting of the threshold upper bound than the other state-of-the-art algorithms to the inlier-outlier threshold. Therefore, it is easier to be applied to unseen problems and scenes without acquiring information by hand about the setting of the inlier-outlier threshold. The source code and examples both in C++ and Python are available at https://github.com/danini/magsac.

**Index Terms**—Robust model estimation, RANSAC, noise scale, M-estimator, marginalization

## 1 Introduction

The RANdom SAmple Consensus (RANSAC) algorithm proposed by Fischler and Bolles [1] in 1981 has become the most widely used robust estimator in computer vision. RANSAC and its variants have been successfully applied to a wide range of vision tasks, e.g., short baseline stereo [2], [3], wide baseline matching [4], [5], [6], motion segmentation [2], detection of geometric primitives [7], pose-graph initialization for structure-from-motion pipelines [8], [9], image mosaicing [10], and to perform [11] or initialize multi-model fitting algorithms [12], [13]. In brief, RANSAC repeatedly selects random subsets of the input data points, typically minimal, and fits a model, e.g., a 2D line to two points, a fundamental matrix to seven 2D point correspondences, or a 6D pose to three 2D-3D correspondences. The quality of the model is then measured, for instance, as the cardinality of its support, i.e., the number of inlier data points. Finally, the model with the highest quality, polished, e.g., by least-squares fitting or numerical optimization on all inliers, is returned.

We propose a new robust loss, a randomized RANSAC-like robust estimator (MAGSAC++) and a termination criterion which eliminate the need for a hand-picked inlier-outlier threshold by marginalizing over a range of noise scales when determining the model quality and the inlier probabilities of data points.

Since the introduction of RANSAC, a number of modifications have been proposed replacing the components of the original algorithm. For instance, improving the sampler impacts the speed of the robust estimation procedure via selecting a good sample early and, thus, triggering the termination criterion. The NAPSAC [17] sampler assumes that inliers are spatially coherent and, therefore, it draws samples from a hyper-sphere centered at the first, randomly selected, location-defining point. If this point is an inlier, the points sampled in its proximity are more likely to be inliers than the ones outside the ball. While NAPSAC exploits the observation that inliers tend to be “closer” to each other than outliers, the GroupSAC algorithm [18] assumes that inliers are often “similar” to each other and, therefore, data points can be separated into groups according to their similarities. PROSAC [19] exploits an a priori predicted inlier probability rank of each point and starts the sampling with the most promising ones. Progressively, samples that are less likely to lead to the sought model are drawn. P-NAPSAC [20] merges the advantages of local and global sampling by drawing samples from progressively growing neighborhoods. Gradually, the algorithm changes from the fully localized NAPSAC to the global PROSAC sampling.

Regarding speeding up the robust estimation process, one way of avoiding unnecessary calculations is via termination of verification of models which are unlikely to be more accurate than the current so-far-the-best. There has been a number of preemptive model verification strategies proposed. For example, when using the $T_{d,d}$ test [21], the model verification is first performed on $d$ randomly selected points (where $d \ll n$). The remaining $n - d$ ones are evaluated only if the first $d$ points are all inliers to the verified model. The test was extended by the so-called bail-out test [22]. Given a model to be scored, a randomly selected subset of $d$ points is evaluated. If the inlier ratio within this

- *Daniel Barath is with the Visual Recognition Group, Department of Cybernetics, Czech Technical University, Prague and the Machine Perception Research Laboratory, SZTAKI, Budapest and also with the Computer Vision and Geometry Group, Department of Computer Science, ETH Zürich, 8092 Zürich, Switzerland. E-mail: dbarath@inf.ethz.ch.*
- *Jana Noskova and Jiri Matas are with the Visual Recognition Group, Department of Cybernetics, Czech Technical University, 166 36 Prague, Czechia. E-mail: Jana.Noskova@cvut.cz, matas@cmp.felk.cvut.cz.*

*Manuscript received 15 Jan. 2021; revised 19 July 2021; accepted 25 July 2021. Date of publication 10 Aug. 2021; date of current version 3 Oct. 2022.*
*(Corresponding author: Daniel Barath.)*
*Recommended for acceptance by P. Mordohai.*
*Digital Object Identifier no. 10.1109/TPAMI.2021.3103562*

subset is significantly smaller than the current best inlier ratio, it is unlikely that the model will yield a larger consensus set than the current maximum and, thus, is discarded. In [23], [24], an optimal randomized model verification strategy was described. The test is based on Wald's theory of sequential testing [25]. Wald's SPRT test is a solution of a constrained optimization problem, where the user supplies acceptable probabilities for errors of the first type (rejecting a good model) and the second type (accepting a bad model) and the resulting optimal test is a trade-off between the time to decision and the errors committed.

Observing that RANSAC requires in practice more samples than what theory predicts, Chum *et al.* [26] identified a problem that not all all-inlier samples are "good", i.e., lead to a model accurate enough to distinguish all inliers, e.g., due to poor conditioning of the selected random all-inlier sample. They address the problem by introducing the locally optimized RANSAC (LO-RANSAC) that augments the original approach with a local optimization step applied to the *so-far-the-best* models. Lebeda *et al.* [14] showed that, for models with many inliers, the local optimization becomes a computational bottleneck due to the iterated least-squares model fitting where the processing time is a function of the number of used points. In [14], it is proposed to consider only a subset of the inliers in the local optimization. Only the final model polishing process is applied to the whole inlier set.

To improve the accuracy by better modelling the noise in the data, different model quality calculation techniques have been investigated. For instance, MLESAC [27] estimates the model quality by a maximum likelihood procedure with all its beneficial properties, albeit under certain assumptions about data point distributions. In practice, MLESAC results are often superior to the inlier counting of plain RANSAC, and they are less sensitive to the manually set inlier-outlier threshold. In MAPSAC [28], the robust estimation is formulated as a process that estimates both the parameters of the data distribution and the quality of the model in terms of maximum a posteriori.

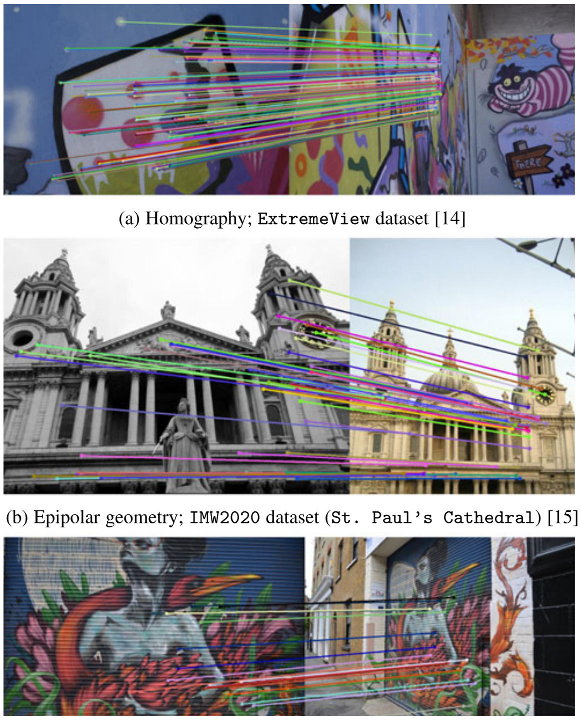
All of the above-mentioned scoring strategies require a manually selected inlier-outlier threshold. Selecting a suitable threshold requires the user to acquire knowledge about the problem and the actual scene, restricting the out-of-the-box applicability of such algorithms. While there are commonly used threshold values for a number of problems, e.g., 2-3 pixels for homography estimation, they rarely lead to highly accurate solutions. Addressing this issue, the dependency on the user-defined inlier-outlier threshold is reduced by its adaptive selection during the model parameter estimation. The MINPRAN [29] algorithm, proposed in 1995, assumes that the outliers are distributed uniformly in the image. For each tested model, MINPRAN tests a number of candidate thresholds and chooses the one with inliers the least likely to have occurred randomly. Moisan *et al.* [30] proposed a contrario RANSAC, AC-RANSAC in short, which follows an approach similar to MINPRAN, but the minimized probability models the consistency of data points with an unknown rigid model. In [31], the best threshold is selected using the Likelihood Ratio Test. While MINPRAN and AC-RANSAC are shown to achieve accurate results, they obtain their solutions using a single adaptively selected threshold. This approach can fail when the background model does not follow the assumed distribution, e.g., the outliers are structured, and it ignores the additional information that other candidate thresholds provide. Also, testing multiple thresholds for each minimal sample model often leads to a deterioration in the processing time. The RECON [32] algorithm assumes that the noisy observations of the sought model have a large amount of common inliers with similar point-to-model residuals. Finding multiple models with similar inlier sets is interpreted as finding the sought model. The RANSAAC [33] algorithm follows a different strategy to eliminate the threshold from the model fitting procedure. RANSAAC estimates models from randomly selected minimal samples similarly as RANSAC. It then converts the models to sets of 2D points, and combines multiple models by averaging the point coordinates used for representing them. Finally, the model is fitted to the averaged point coordinates. Besides the number of drawbacks of RANSAAC, e.g., non-robust model-to-points conversion, it is shown by the authors that it only works inside a local optimization process after a reasonably good model is found. Thus, the inlier-outlier threshold is still required.

As the *main contribution* of this paper, we propose an approach, $\sigma$-consensus++, that eliminates the need for a precise user-defined noise scale $\sigma$ when estimating the model parameters in a robust manner. Instead of $\sigma$, only a loose upper bound $\sigma_{\max}$ is required defining the range of possible threshold values. The $\sigma$-consensus++ algorithm is in fact a new M-estimator (a robust kernel), solved by an iteratively re-weighted least squares procedure. This M-estimator marginalizes over the range of noise scales. As *minor contributions*, we propose a new termination criterion which does not require a $\sigma$ value. Considering the fact that some applications, e.g., structure-from-motion [34], need to know inliers, we propose a way to adaptively determine the set of inliers after the robust estimation finishes. The inliers are selected by thresholding, such that the model to which they lead after least-squares fitting is similar to the model determined by the robust estimation procedure applied without inlier-outlier decisions done.

Preliminary versions of MAGSAC++ with $\sigma$-consensus+ + were published at CVPR 2019 [35] and CVPR 2020 [20]. This paper extends and improves them by (i) combining their "bells and whistles", (ii) proposing a termination criterion applicable for MAGSAC++, (iii) proposing an inlier selection technique after the robust process is applied, (iv) and providing a number of new experiments on homography, fundamental matrix and relative pose estimation. Example results are shown in Fig. 1.

## 2 Notation and Preliminaries

In this paper, the set of input data points is denoted $\mathcal{P} = \{p \mid p \in \mathbb{R}^{\nu}, \nu \in \mathbb{N}_{>0}\}$, where $\nu$ is the dimension, e.g., $\nu = 2$ for 2D points and $\nu = 4$ for point correspondences. The inlier set is $\mathcal{I} \subseteq \mathcal{P}$. The model to fit is represented by its parameter vector $\theta \in \Theta$, where $\Theta = \{\theta \mid \theta \in \mathbb{R}^{d}, d \in \mathbb{N}_{>0}\}$ is the manifold, e.g., of all possible 2D lines, and $d = 2$ is the dimension of the model (angle and offset). Fitting function $F : \mathcal{D} \rightarrow \Theta$, where $\mathcal{D} \subset \mathcal{P}^{*}$ and $|\mathcal{D}| \geq m$, calculates the model



(a) Homography; `ExtremeView` dataset [14]

(b) Epipolar geometry; `IMW2020` dataset (`St. Paul's Cathedral`) [15]

(c) Homography; `HPatches` dataset [16]

Fig. 1. Example image pairs from the datasets used for testing the robust estimators. The inliers of MAGSAC++, selected adaptively by the proposed procedure, are visualized.

parameters from $n \geq m$ data points, where $\mathcal{P}^* = \exp\mathcal{P}$ is the power set of $\mathcal{P}$ and $m \in \mathbb{N}_{>0}$ is the minimum point number for fitting a model, e.g., $m = 2$ for lines. Note that $F$ is a combined function applying different estimators based on the input point set. For instance, for $\mathcal{P}' \in \mathcal{P}^*$

$$F(\mathcal{P}') = \begin{cases} \text{MinimalSolver}(\mathcal{P}') & \text{if } |\mathcal{P}'| = m, \\ \text{LSQ}(\mathcal{P}') & \text{otherwise.} \end{cases} \tag{1}$$

Function $R : \Theta \times \mathcal{P} \rightarrow \mathbb{R}^+$ calculates the point-to-model residual. Function $I : \Theta \times \mathbb{R}^+ \times \mathcal{P}^* \rightarrow \mathcal{P}^*$ selects the set of inliers given model $\theta$ and noise standard deviation $\sigma$. We assume that the inlier-outlier threshold is calculated from the noise $\sigma$ as $\tau(\sigma) = k\sigma$, where $k$ is some constant. For instance, for the original RANSAC approach, $I_{\text{RANSAC}}(\theta, \sigma, \mathcal{P}) = \{p \in \mathcal{P} \mid R(\theta, p) < \tau(\sigma)\}$ and $\tau(\sigma) = \sigma$. The model quality function, measuring how much the actual model interprets the scene, is $Q : \Theta \times \mathbb{R}^+ \times \mathcal{P}^* \rightarrow \mathbb{R}^+$. Higher quality is interpreted as better model. Let $\{R(\theta, p_i)\}_{i=1}^{n}$ be the point-to-model residuals, ordered increasingly, such that $0 \leq R(\theta, p_1) < R(\theta, p_2) < \cdots < R(\theta, p_n)$. For RANSAC, $Q_{\text{RANSAC}}(\theta, \sigma, \mathcal{P}) = |I(\theta, \sigma, \mathcal{P})|$ and for MSAC, it is

$$Q_{\text{MSAC}}(\theta, \sigma, \mathcal{P}) = |I(\theta, \sigma, \mathcal{P})| - \frac{1}{\tau(\sigma)^2} \sum_{i=1}^{|I(\theta,\sigma,\mathcal{P})|} R^2(\theta, p_i).$$

## 3 MAGSAC

First, we describe the idea and design choices of the original MAGSAC [35] approach in brief. We will also discuss its merits and drawbacks.

Symbols used in this paper

| | |
|---|---|
| $\mathcal{P} = \{p \mid p \in \mathbb{R}^\nu, \nu \in \mathbb{N}_{>0}\}$ | - Set of data points |
| $\mathcal{P}^*$ | - Power set of $\mathcal{P}$ |
| $\sigma \in \mathbb{R}^+$ | - Noise standard deviation |
| $\sigma_{\max} \in \mathbb{R}^+$ | - Noise std. upper bound |
| $\tau(\sigma)$ | - Inlier-outlier threshold |
| $\Theta = \{\theta \mid \theta \in \mathbb{R}^d, d \in \mathbb{N}_{>0}\}$ | - Model manifold |
| $R : \Theta \times \mathcal{P} \rightarrow \mathbb{R}^+$ | - Point-to-model residual |
| $F : \mathcal{P}^* \rightarrow \Theta$ | - Model estimator function |
| $I : \Theta \times \mathbb{R}^+ \times \mathcal{P}^* \rightarrow \mathcal{P}^*$ | - Inlier selector function |
| $Q : \Theta \times \mathbb{R}^+ \times \mathcal{P}^* \rightarrow \mathbb{R}^+$ | - Model quality function |

### 3.1 Marginalizing Sample Consensus

*Idea.* In the original marginalizing sample consensus (MAGSAC) algorithm [35], the model quality is defined by marginalizing over the noise scale $\sigma$ as follows:

$$Q^*(\theta, \mathcal{P}) = \int_0^{+\infty} Q(\theta, \sigma, \mathcal{P}) f(\sigma) \mathrm{d}\sigma,$$

where the noise $\sigma$ is a random variable with density function $f(\sigma)$, $Q : \Theta \times \mathbb{R}^+ \times \mathcal{P}^* \rightarrow \mathbb{R}^+$ is a quality function, e.g., the inlier counting of RANSAC, which depends on an input model $\theta \in \Theta$, the inlier-outlier threshold $\tau(\sigma)$, and the set $\mathcal{P}$ of $n$ data points.

Having no prior information, $\sigma$ is assumed to be uniformly distributed within range $(0, \sigma_{\max})$, where $\sigma_{\max}$ is an upper bound for the noise scale ($\sigma_{\max} > 0$). Considering this assumption, the quality calculation becomes

$$Q^*(\theta, \mathcal{P}) = \frac{1}{\sigma_{\max}} \int_0^{\sigma_{\max}} Q(\theta, \sigma, \mathcal{P}) \mathrm{d}\sigma. \tag{2}$$

For instance, using the inlier counting of plain RANSAC $Q_{\text{RANSAC}}(\theta, \sigma, \mathcal{P})$, where $\tau(\sigma) = \sigma$ is the inlier-outlier threshold, we get marginalized quality function

$$Q^*_{\text{RANSAC}}(\theta, \mathcal{P}) = |I(\theta, \sigma_{\max}, \mathcal{P})| - \frac{1}{\sigma_{\max}} \sum_{i=1}^{|I(\theta,\sigma_{\max},\mathcal{P})|} R(\theta, p_i).$$

*Data Interpretation and Design Choices.* In MAGSAC, the choice of the marginalized quality function $Q$ is motivated by the assumption that the residuals are calculated as the square root of a sum of squared normally distributed variables. Typically, the residuals of the inliers are calculated as the euclidean-distance from model $\theta$ in some $\nu$-dimensional space (e.g., the re-projection error). In the case of assuming the distances along each axis of this $\nu$-dimensional space to be independent and normally distributed with the same variance $\sigma^2$, value $(\text{residuals})^2/\sigma^2$ has $\chi^2$- distribution with $\nu$ degrees of freedom. For a given $\sigma$, the residuals of the inliers are described by the trimmed $\chi$-distribution[1] with $\nu$ degrees of freedom multiplied by $\sigma$ with density

$$g(r \mid \sigma) = 2C(\nu)\sigma^{-\nu}\exp(-r^2/2\sigma^2)r^{\nu-1},$$

for $r < \tau(\sigma)$ and $g(r \mid \sigma) = 0$ for $r \geq \tau(\sigma)$. The normalizing constant $C(\nu) = (2^{\nu/2}\Gamma(\nu/2)\alpha)^{-1}$ and, for $a > 0$

[1] The square root of $\chi^2$-distribution.

$$\Gamma(a) = \int_0^{+\infty} t^{a-1}\exp(-t)\mathrm{d}t,$$

is the gamma function, $\nu$ is the dimension of the euclidean space in which the residuals are calculated and $\tau(\sigma)$ is set to $\alpha$-quantile (e.g., $\alpha = 0.99$) of the non-trimmed distribution.

*Note*: the idea of model quality marginalization is general and independent of the choice of the noise distribution, here $\chi^2$.

*Model Polishing.* The last step of RANSAC-like algorithms is the re-fitting of the model to all inliers. However, due to MAGSAC not making a strict inlier-outlier decision, the standard model polishing step is not directly applicable. Therefore, the $\sigma$-consensus algorithm was proposed which, first, assigns an inlier weight to each point and, finally, applies weighted least-squares fitting.

Suppose an input point set $\mathcal{P}$ and model $\theta$ estimated from a minimal sample as in RANSAC. Let $\theta_\sigma = F(I(\theta, \sigma, \mathcal{P}))$ be the model estimated from the inlier set

$$I(\theta, \sigma, \mathcal{P}) = \{p \mid p \in \mathcal{P} \wedge R(\theta, p) < \tau(\sigma)\}, \tag{3}$$

selected using threshold $\tau(\sigma)$ around the input model $\theta$. Scalar $\tau(\sigma)$ is the threshold which $\sigma$ implies; function $F$ estimates the model parameters from a set of data points; function $I$ returns the set of data points for which the point-to-model residuals are smaller than $\tau(\sigma)$.

For each possible $\sigma$ value, the likelihood of point $p \in \mathcal{P}$ being inlier is calculated as

$$\mathrm{P}(p \mid \theta_\sigma, \sigma) = 2C(\nu)\sigma^{-\nu}R^{\nu-1}(\theta_\sigma, p)\exp\left(\frac{-R^2(\theta_\sigma, p)}{2\sigma^2}\right),$$

if $R(\theta_\sigma, p) \leq \tau(\sigma)$, where $R(\theta_\sigma, p)$ is the point-to-model residual. If $R(\theta_\sigma, p) > \tau(\sigma)$, likelihood $\mathrm{P}(p \mid \theta_\sigma, \sigma)$ is 0. For each point $p$, likelihood $\mathrm{P}(p \mid \theta_\sigma, \sigma)$ is marginalized over $\sigma$ and the obtained probability is used as an inlier weight in the final weighted least-squares fitting. The objective function $Q(\theta, \sigma, \mathcal{P})$ is the log-likelihood with inlier density $g(r \mid \sigma)$ and outliers assumed uniformly distributed.

*Issues.*There are two main issues with the MAGSAC approach, a practical and a theoretical one. In practice, the procedure of marginalizing $\mathrm{P}(p \mid \theta_\sigma, \sigma)$ over $\sigma$ calculates $\mathrm{P}(p \mid \theta_\sigma, \sigma)$ a number of times with different $\sigma$ values. Each calculation requires to select the set of inliers and obtain $\theta_\sigma$ by LS fitting on them. *This step is time consuming* even with the number of speedups proposed in the original paper [35]. The theoretical issue is that the objective function does not have its maximum at zero. Consequently, in the case of having perfect data, i.e., no noise, MAGSAC fails to return the sought model parameters. As a minor issue, both the quality function and the likelihood can only be calculated approximately for non piece-wise constant objective functions, e.g., $\chi^2$-based or truncated $L_2$ loss. The exact calculation can only be done for the RANSAC-like inlier counting.

## 4 MAGSAC++

The MAGSAC++ algorithm is proposed here via reformulating the previously described MAGSAC problem as an iteratively re-weighted least-squares (IRLS) approach. To do so, a new model quality function and a procedure to polish the model parameters without making strict inlier-outlier decisions and doing a number of LS fittings are proposed.

The proposed MAGSAC++ is based on an iteratively reweighted least squares (IRLS) approach where the model parameters in the $(i+1)$th step are calculated as follows:

$$\theta_{i+1} = \arg\min_\theta \sum_{p \in \mathcal{P}} w(R(\theta_i, p))R^2(\theta, p), \tag{4}$$

where the weight of point $p$ is

$$w(R(\theta_i, p)) = \int_0^{+\infty} \mathrm{P}(p \mid \theta_i, \sigma)f(\sigma)\mathrm{d}\sigma, \tag{5}$$

and $\theta_0 = \theta$, i.e., the initial model from the minimal sample.

*Data Interpretation and Design Choices.*Similarly as in MAGSAC, the inlier residuals are euclidean-distances of points assumed to be corrupted by Gaussian noise and, thus, have $\chi$-distribution. The noise standard deviation $\sigma$ is assumed to be uniformly distributed within $(0, \sigma_{\max})$. However, we make no assumptions about the outlier distributions. Note that the proposed quality and inlier weight functions can be modified straightforwardly when considering differently distributed inliers.

### 4.1 Inlier Weight Calculation

The weight function defined in (5) is the marginal density of the inlier residuals as follows:

$$w(r) = \int_0^{+\infty} g(r \mid \sigma)f(\sigma)\mathrm{d}\sigma. \tag{6}$$

Let $\tau(\sigma) = k\sigma$ be the chosen quantile of the $\chi$-distribution. For residual $0 \leq r \leq k\sigma_{\max}$

$$w(r) = \frac{1}{\sigma_{\max}}\int_{r/k}^{\sigma_{\max}} g(r \mid \sigma)\mathrm{d}\sigma = \frac{1}{\sigma_{\max}}C(\nu)2^{\frac{\nu-1}{2}}$$
$$\left(\Gamma\left(\frac{\nu-1}{2}, \frac{r^2}{2\sigma_{\max}^2}\right) - \Gamma\left(\frac{\nu-1}{2}, \frac{k^2}{2}\right)\right)$$

and, for $r > k\sigma_{\max}$, weight $w(r) = 0$. Function

$$\Gamma(a, x) = \int_x^{+\infty} t^{a-1}\exp(-t)\mathrm{d}t,$$

is the upper incomplete gamma function. Due to the design choices, weight $w(r)$ is positive and decreasing on interval $[0,\ \tau(\sigma_{\max})]$. Thus there is a $\rho$-function of an M-estimator which is minimized by IRLS using $w(r)$ and each iteration guarantees a non-increase in its loss function (chapter 9 of [36]). Consequently, it converges to a local minimum. If different noise distribution is assumed, this property does not necessarily hold. In those cases, a different algorithm should be used to solve the problem, e.g., Levenberg-Marquardt optimization [37].

IRLS (4) where $w(r)$ is defined by (6) with $\tau(\sigma) = 3.64\sigma$, where 3.64 is the 0.99 quantile of the $\chi$-distribution with $\nu = 4$, will be called $\sigma$-consensus++ for problems using point correspondences. Parameter $\sigma_{\max}$ is the same user-defined maximum noise level parameter as in MAGSAC, usually, set to a fairly high value, e.g., 10 pixels for homography fitting. The $\sigma$-consensus++ algorithm is applied for fitting to a non-minimal sample and, also, as a post-processing to improve the output of any robust estimator.

### 4.2 Model Quality Function

In order to select the model interpreting the data, a quality function has to be defined. Let

$$Q_{\mathrm{M++}}(\theta,\mathcal{P}) = n - \frac{1}{\rho(k\sigma_{\max})} L(\theta,\mathcal{P}) = |I(\theta,\sigma_{\max},\mathcal{P})| - \frac{1}{\rho(k\sigma_{\max})}\sum_{i=1}^{|I(\theta,\sigma_{\max},\mathcal{P})|}\rho(R(\theta,p_i)),$$

where

$$L(\theta,\mathcal{P}) = \sum_{p\in\mathcal{P}}\rho(R(\theta,p)), \tag{7}$$

is a loss function of the M-estimator defined by our weight function $w(r)$. Function

$$\rho(r) = \int_0^r x w(x)\mathrm{d}x = \int_0^{+\infty}\left(\int_0^r x g(x \mid \sigma)\mathrm{d}x\right) f(\sigma)\mathrm{d}\sigma,$$

for $r \in [0,+\infty)$. For any point $p$ with residual $r$, the loss function is the mean of the residual values lower then $r$ of a random variable with $\chi$-distribution, i.e., the assumed distribution of the inlier residuals. Thus, the $\rho$-function is some type of a reasonable distance. It can be formulated in the same way for each $\sigma$ and then marginalized over $\sigma$ as in MAGSAC.

Due to assuming that the $\sigma$ values are uniformly distributed within range $[0,\sigma_{\max}]$ for $0 \leq r \leq \tau(\sigma_{\max})$

$$\rho(r) = \frac{1}{\sigma_{\max}}\int_0^{\sigma_{\max}}\left[C(\nu)2^{\frac{\nu+1}{2}}\sigma \gamma\left(\frac{\nu+1}{2},\frac{r^2}{2\sigma^2}\right) - \frac{r^2}{2}g(k\sigma_{\max}|\sigma)\right]\mathrm{d}\sigma$$

and the integral can be removed as follows:

$$\rho(r) = \frac{1}{\sigma_{\max}}C(\nu)2^{\frac{\nu+1}{2}}\left[\frac{\sigma_{\max}^2}{2}\gamma\left(\frac{\nu+1}{2},\frac{r^2}{2\sigma_{\max}^2}\right) + \frac{r^2}{4}\left(\Gamma\left(\frac{\nu-1}{2},\frac{r^2}{2\sigma_{\max}^2}\right) - \Gamma\left(\frac{\nu-1}{2},\frac{k^2}{2}\right)\right)\right].$$

For $r > \tau(\sigma_{\max})$

$$\rho(r) = \rho(k\sigma_{\max}) = \sigma_{\max}C(\nu)2^{\frac{\nu-1}{2}}\gamma\left(\frac{\nu+1}{2},\frac{k^2}{2}\right),$$

where

$$\gamma(a,x) = \int_0^x t^{a-1}\exp(-t)\mathrm{d}t,$$

is the lower incomplete gamma function. Weight $w(r)$ can be calculated precisely or approximately as in MAGSAC. However, the precise calculation can be done very fast by *storing the values* of the complete and incomplete gamma functions *in a lookup table*. Then the weight and quality calculation becomes merely a few operations per point. MAGSAC++ algorithm uses (7) as quality function and $\sigma$-consensus++ for estimating the model parameters. Function $w(r)$ is visualized in Fig. 2 together with other weightings which are often used for robust model fitting.

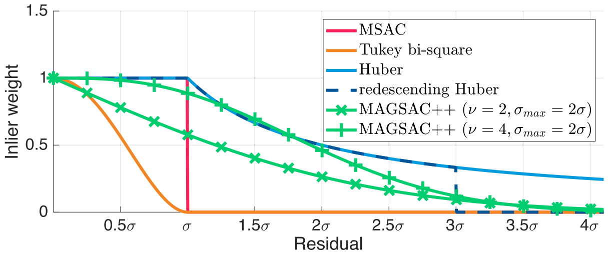


Fig. 2. Weighting functions for robust fitting. For MAGSAC++, we use $\sigma_{\max} = 2\sigma$ as an example and degrees-of-freedom $\nu = 2$ (e.g., 2D line fitting) and 4 (e.g., problems with point correspondences).

### 4.3 Termination Criterion

The number of inliers during the robust estimation is unknown due to not making strict inlier-outlier decisions. It is thus not possible to apply the standard termination criterion of RANSAC [38]

$$k(\theta,\sigma,\mathcal{P}) = \frac{\ln(1-\mu)}{\ln\left(1-\left(\frac{|I(\theta,\sigma,\mathcal{P})|}{|\mathcal{P}|}\right)^m\right)}, \tag{8}$$

where $k$ is the iteration number, $\mu$ is a manually set confidence in the results (typical values are 0.95 or 0.99), $m$ is the size of the minimal sample needed for the estimation, and $|I(\theta,\sigma,\mathcal{P})|$ is the inlier number of the so-far-the-best model.

In order to determine $k$ without using a particular value for $\sigma$, it is a straightforward choice to marginalize over the noise scale $\sigma$. Let us assume that the points are ordered by their residuals as $0 = \tau(\sigma_0) \leq R(\theta,p_1) = \tau(\sigma_1) \leq R(\theta,p_2) = \tau(\sigma_2) \leq \cdots \leq R(\theta,p_k) = \tau(\sigma_k) \leq \tau(\sigma_{\max}) < R(\theta,p_{k+1}) = \tau(\sigma_{k+1}) \leq \cdots \leq R(\theta,p_n) = \tau(\sigma_n)$. The iteration number is calculated as

$$k^*(\theta,\mathcal{P}) = \frac{1}{\sigma_{\max}}\int_0^{\sigma_{\max}} k(\theta,\sigma,\mathcal{P})d\sigma = \tag{9}$$

$$\frac{1}{\sigma_{\max}}\int_0^{\sigma_{\max}}\frac{\ln(1-\mu)}{\ln\left(1-\left(\frac{|I(\theta,\sigma,\mathcal{P})|}{|\mathcal{P}|}\right)^m\right)}d\sigma. \tag{10}$$

Due to the fact that function $|I(\theta,\sigma,\mathcal{P})|$, measuring the number of inliers given a noise scale $\sigma$, is piece-wise constant, and that is the only part of (10) depending on $\sigma$, the integral can be replaced by a weighted summation. It is as follows:

$$k^*(\theta,\mathcal{P}) = \frac{1}{\sigma_{\max}}\sum_{i=1}^{k}\frac{(\sigma_i - \sigma_{i-1})\ln(1-\mu)}{\ln\left(1-\left(\frac{|I(\theta,\sigma_{i-1},\mathcal{P})|}{|\mathcal{P}|}\right)^m\right)}. \tag{11}$$

The function is, however, problematic when there are no points with zero residual. In that case, the denominator becomes $\ln(1) = 0$ and the iteration number $\infty$. We, thus, shift the inlier number by one and introduce a slight and artificial approximation as

$$k^*(\theta,\mathcal{P}) \approx \frac{1}{\sigma_{\max}}\sum_{i=1}^{k}\frac{(\sigma_i - \sigma_{i-1})\ln(1-\mu)}{\ln\left(1-\left(\frac{i}{|\mathcal{P}|}\right)^m\right)}. \tag{12}$$

Thus the number of iterations required for MAGSAC++ is calculated during the procedure and updated whenever a new so-far-the-best model is found, similarly as in RANSAC.

## 5 Inlier Selection

For some applications, the knowledge of what is inlier and outlier is a requirement. For instance, in structure-from-motion algorithms, the inlier correspondences are triangulated in 3D after the relative pose estimation and used for the reconstruction. Given the estimated model parameters $\theta$ after applying MAGSAC++, the objective is to find a reasonable set of inliers without introducing new parameters, e.g., a threshold. The *idea* is to return the set of points on which a least-squares fitting leads to a model which is similar to the one determined by the robust estimator. The problem is formalized as follows:

$$\mathcal{I}^* = \underset{\mathcal{I} \subseteq \mathcal{P}}{\arg\min} |F(\mathcal{I}) - \theta|, \tag{13}$$

where function $F$ estimates the model parameters from a set of data points, and norm $|.|$ is some distance function defined over the model manifold. Note that this formulation allows to consider as inliers points with large point-to-model residuals. Besides, the problem introduced in (13) is NP-hard. Therefore, we weaken (13) by assuming that there exists a noise scale $\sigma^*$ and, thus, an inlier-outlier threshold $\tau(\sigma^*)$ such that the points with residuals smaller than $\tau(\sigma^*)$ are the elements of $\mathcal{I}^*$. Consequently, it is enough to find $\sigma^*$. The problem becomes the following:

$$\sigma^* = \underset{\sigma \in \Sigma}{\arg\min} |F(I(\theta, \sigma, \mathcal{P})) - \theta|, \tag{14}$$

where $\Sigma = \{\sigma_i\}_{i=1}^{k} \subset [0, \sigma_{\max}]$ as introduced above (10). Note that it is straightforward to see that there are no other threshold values leading to different sets of inliers [29].

In the algorithm, we define the model-to-model distance as the sum of $L^1$ point-to-model residual distances as follows:

$$|\theta_1 - \theta_2| = \sum_{p \in \mathcal{P}} |R(\theta_1, p) - R(\theta_2, p)|. \tag{15}$$

Since the sought model should be of the same distance from both the inliers and outliers as the initial one, distance $|\theta_1 - \theta_2|$ can measured on all points without differentiating inliers and outliers. Since we measure the $L^1$ residual differences, outlier points with large residuals do not have higher impact on the model-to-model distance than inliers with small residuals. Also, distance $|\theta_1 - \theta_2|$ is enough to be measured only on a subset of points to speed up the procedure when needed. The pseudo-code of the algorithm is shown in Algorithm 3. Parameter $n_{\min}$ is the minimum number of points required to return, depending on the current application. If there is no requirement, $n_{\min} = m$, where $m$ is the minimal sample size. Note that for models which are estimated from a larger-than-minimal sample by using SVD decomposition, e.g., fundamental/essential matrix, homography, using an incremental version of SVD, e.g., [39], speeds up the procedure significantly when a large number of points falls closer than $\sigma_{\max}$. Also, the procedure is straightforwardly parallelizable on multiple CPU cores.

**Algorithm 1.** The MAGSAC++ Algorithm

```
Input: $\mathcal{P}$ – data points; $\epsilon_{\max}$ – max. threshold
       $\mu$ – confidence;
Output: $\theta^*$ – model parameters; $\mathcal{I}^*$ – inliers (optional)
$q^* \leftarrow 0$.
while $\neg$ Terminate($\mu, q^*$) do                         ▷ Section 4.3
    $S \leftarrow$ Sample($\mathcal{P}$).            ▷ default: P-NAPSAC sampler [20]
    if $\neg$ TestSample($S$) then          ▷ Degen. and cheirality tests
        continue
    $\theta \leftarrow$ EstimateModel($S$)
    if $\neg$ TestModel($\theta$) then          ▷ Degen. and cheirality tests
        continue
    $\theta' \leftarrow \sigma$-consensus++($\mathcal{P}, \theta, \tau^{-1}(\epsilon_{\max})$)        ▷ Algorithm 2
    if $\neg$ TestModel($\theta'$) then         ▷ Degen. and cheirality tests
        continue
    $q \leftarrow$ Scoring($\mathcal{P}, \theta', \tau^{-1}(\epsilon_{\max})$)                  ▷ Eq. (7)
    if $q > q^*$ then
        $q^*, \theta^* \leftarrow q, \theta'$
$\mathcal{I}^* \leftarrow$ SelectInliers($\theta^*, \mathcal{P}$)             ▷ Section 5 (optional)
```

**Algorithm 2.** The $\sigma$-Consensus++ Algorithm

```
Input: $\mathcal{P}$ – data points; $\sigma_{\max}$ – max. noise scale
       $\theta$ – initial model;
Output: $\theta^*$ - model parameters
$\theta_0, i \leftarrow \theta, 0$.
repeat
    $\{r_j\}_{j=1}^{|\mathcal{P}|} \leftarrow \{R(\theta_i, p) \mid p \in \mathcal{P}\}$
    $\{\widehat{r}_j\}_{j=1}^{|\mathcal{P}|} \leftarrow$ Sort($\{r_j\}_{j=1}^{|\mathcal{P}|}$)
    $\{w_j\}_{j=1}^{|\mathcal{P}|} \leftarrow \{w(\widehat{r}_j)\}_{j=1}^{|\mathcal{P}|}$                       ▷ Eq. (6)
    $\theta_{i+1} \leftarrow$ WLS($\mathcal{P}, \{w_j\}_{j=1}^{|\mathcal{P}|}$)            ▷ Weighted least-squares
    if $\neg$ TestModel($\theta_{i+1}$) then          ▷ Degen. and cheir. tests
        break
    $i \leftarrow i + 1$
until Terminate($\theta_{i-1}, \theta_i, i$)
$\theta^* \leftarrow \theta_i$
```

## 6 Algorithmic Choices

To achieve state-of-the-art results, we combine the proposed MAGSAC++ with the components discussed in USAC [40]. We consider three popular vision problems, i.e., fundamental matrix, homography and relative pose (i.e., essential matrix) estimation. The included components for each problem are as follows: 1. *Sample degeneracy*. The degeneracy tests of minimal samples are for rejecting clearly bad samples to avoid the sometimes expensive model estimation. For homographies, samples consisting of collinear points are rejected. 2. *Sample cheirality*. The test is for rejecting samples based on the assumption that both of the cameras observing a 3D surface must be on its same side. For homography fitting, we check if the ordering of the four point correspondences – along their convex hulls – in both images are the same. If not, the sample is rejected. 3. *Model degeneracy*.

TABLE 1
The Average Processing Times (in Milliseconds) and Errors (in Pixels) in the Estimated Homographies (H), Fundamental (F) and Essential (E) Matrices Using Different Methods for Solving the Linear Systems in Their Solvers When Estimating the Model Parameters From a Minimal ($m$) or a Larger-Than-Minimal ($> m$) Sample. Each test is repeated 100 000 times. The size of the larger-than-minimal sample is selected uniformly randomly from range $[m+1, 1000]$. For error calculation, the re-projection was used for homographies, and Sampson-distance for fundamental and essential matrices. The tested methods solving linear systems are the ones implemented in the Eigen library.

| | Average processing time (milliseconds) | | | | | | Average error (pixels) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | H | | F | | E | | H | | F | | E | |
| | $m$ | $> m$ | $m$ | $> m$ | $m$ | $> m$ | $m$ | $> m$ | $m$ | $> m$ | $m$ | $> m$ |
| LLT | **0.002** | – | – | – | – | – | $10^{-8}$ | – | – | – | – | – |
| LDLT | 0.003 | – | – | – | – | – | $10^{-8}$ | – | – | – | – | – |
| PartialPivLU | 0.003 | – | – | – | – | – | **$10^{-11}$** | – | – | – | – | – |
| FullPivLU | 0.003 | – | **0.011** | – | **0.060** | – | **$10^{-11}$** | – | **$10^{-12}$** | – | **$10^{-14}$** | – |
| HouseholderQR | 0.005 | 0.099 | 0.014 | 0.028 | 0.067 | 0.081 | **$10^{-11}$** | **$10^{-7}$** | $10^{-9}$ | **$10^{-7}$** | $10^{-12}$ | $10^{-6}$ |
| ColPivHouseholderQR | 0.006 | **0.085** | 0.015 | 0.027 | 0.069 | 0.077 | $10^{-10}$ | **$10^{-7}$** | $10^{-10}$ | **$10^{-7}$** | $10^{-13}$ | **$10^{-8}$** |
| FullPivHouseholderQR | 0.006 | 0.103 | 0.014 | **0.026** | 0.066 | **0.075** | **$10^{-11}$** | **$10^{-7}$** | **$10^{-12}$** | **$10^{-7}$** | **$10^{-14}$** | $10^{-3}$ |
| JacobiSVD | 0.023 | 22.356 | 0.028 | 0.039 | 0.079 | 0.088 | $10^{-6}$ | $10^{-6}$ | $10^{-4}$ | $10^{-6}$ | $10^{-13}$ | $10^{-7}$ |
| BDCSVD | 0.024 | 27.954 | 0.028 | 0.040 | 0.080 | 0.089 | $10^{-6}$ | $10^{-6}$ | $10^{-4}$ | $10^{-6}$ | $10^{-13}$ | $10^{-7}$ |

The purpose of this test is to reject models early to avoid verifying them unnecessarily. For fundamental matrices, DEGENSAC [41] is applied to determine if the epipolar geometry is affected by a dominant plane. For relative pose estimation, improper rotation matrices [42], i.e., the ones with negative determinant, are rejected. We observed that, for epipolar geometry estimation, symmetric epipolar distance tends to be more robust to degenerate models. In contrast, Sampson distance leads to higher accuracy – when using Sampson distance some degenerate models have lots of inliers. Therefore, we use Sampson distance as residual function when estimating fundamental and essential matrices and reject all models where the inlier number is significantly lower with symmetric epipolar distance. In practice, we found that a model can be rejected if it does not have at least half as many inliers with symmetric epipolar distance as with Sampson distance. 4. *Model cheirality*. The test is for rejecting models considering that the cameras must be on the same side of the observed surface. For fundamental and essential matrix estimation, we apply the oriented epipolar constraint [43]. 5. *Sampling*. We use the P-NAPSAC sampler [20]. It requires an a priori determined ordering of the input data points for its PROSAC [19] part. We used the scoring coming from the ratio-test [44]. The neighborhoods were determined by a multi-layer grid as proposed in [20] to minimize the computational overhead. 6. *Solvers*. One of the most time-sensitive parts of RANSAC-like robust estimation is the solver estimating the model parameter from a minimal or larger-than-minimal sample. It is time-sensitive since it runs *at least* once in every iteration. In many popular vision problems, e.g., homography estimation, the solution includes homogeneous or inhomogeneous linear systems. We thus tested the ways of solving such systems by the algorithms implemented in the Eigen library and chose the actual solvers in our MAGSAC++ implementation accordingly. Homographies are estimated by the standard normalized 4PT algorithm [38]. In the minimal case, the correspondences were not normalized since the system is not over-determined – the solution is exact. For fundamental matrices, the 7PT algorithm [38] runs to estimate from a minimal sample. In the over-determined case, we applied the normalized 8PT algorithm [45]. Essential matrices are estimated by the solver of Stewenius *et al.* [46]. When selecting the actual method applied to solve a linear system, our strategy was the following.

Table 1 reports the accuracy in pixels and processing time in milliseconds of methods solving the linear systems in the solvers for homography, fundamental and essential matrix estimation. Each test is repeated 100 000 times on randomly generated point correspondences. In each test, the size of the larger-than-minimal sample is selected uniformly randomly from range $[m+1, 1000]$, where $m$ is the sample size.

In the minimal case, we chose the fastest methods from Table 1 since the accuracy is not crucial – the model is always improved later on more inliers. Also, this solver runs the most times. For fitting homographies to minimal samples, we solve the normal equations of the implied linear system via the Cholesky decomposition (LLT in the table). For estimating fundamental matrices, the null-space from the coefficient matrix is calculated by the LU decomposition with complete pivoting since that is one of the fastest solutions when we are given a $7 \times 8$ coefficent matrix (FullPivLU). For essential matrices, we chose the LU decomposition with complete pivoting (FullPivLU).

In the over-determined case, we selected the methods leading to the lowest errors. If there are multiple ones leading to the same error, the fastest one is applied. For fitting homographies, we apply the QR decomposition with column pivoting (ColPivHouseholderQR) – all tested types of QR decomposition lead to similarly low error, but column pivoting is the fastest. For estimating fundamental matrices, the null-space from the coefficient matrix is calculated by the QR decomposition with full pivoting (FullPivHouseholderQR). For essential matrices, we chose the QR decomposition with column pivoting (ColPivHouseholderQR).

The pseudo-code of MAGSAC++ and $\sigma$-consensus++ are shown in Algorithms 1 and 2, respectively. In the algorithm, TestSample refers to the degeneracy and cheirality checks applied to minimal samples. Function TestModel is the degeneracy and cheirality checks applied to the estimated models.

**Algorithm 3.** Inlier Selection

```
Input: $\mathcal{P}$ – data points; $\theta$ – initial model
$n_{\min}$ – min. # of required points ▷ default: sample size
Output: $\mathcal{I}^*$ – inliers
$\{r_j\}_{j=1}^k \leftarrow$ Sort($\{R(\theta, p) \mid p \in \mathcal{P} \wedge R(\theta, p) \leq \tau(\sigma_{\max})\}$)
$\epsilon^* \leftarrow \infty$.
for $i = n_{\min} \dots k$ do
    $\theta' \leftarrow$ LS($\{r_j\}_{j=1}^i$)        ▷ Least-squares fitting
    $\epsilon \leftarrow |\theta - \theta'|$        ▷ Eq. (15)
    if $\epsilon < \epsilon^*$ then
        $\epsilon^*, \mathcal{I}^* \leftarrow \epsilon, \{r_j\}_{j=1}^i$
```

## 7 Experiments

For testing the proposed methods, we used the problems and datasets from CVPR tutorial *RANSAC in 2020* [47]. The datasets and codes used are available at https://github.com/ducha-aiki/ransac-tutorial-2020-data. The hyper-parameters of all compared methods were tuned on the provided training set to maximize the accuracy. The reported errors were then calculated on the set which was not used when setting the hyper-parameters.

The error metric used is the mean Average Accuracy (mAA). This metric was originally introduced in [48], where it was called mean Average Precision (mAP). Later, Jin *et al.* [49] argued that "accuracy" is the correct terminology, due to simply evaluating how many of the predicted poses are accurate, as determined by thresholding the acceptance threshold, i.e., the threshold which decides if a particular result is accurate or not.

In order to determine which method is the least sensitive to the setting of either $\sigma$ or $\sigma_{\max}$, we also measure the insensitivity to the inlier-outlier threshold (or upper limit in the case of MAGSAC, MAGSAC++ and AC-RANSAC). The methods were run multiple times using different threshold values from $t_1, \dots, t_n$. For fundamental matrix and relative pose estimation, $t_{1..8} = (0.1, 0.25, 0.5, 1.0, 1.5, 3.0, 5.0, 10.0)$. For homography estimation, the following threshold values are used $t^{\mathrm{H}}_{1..12} = (0.1, 0.25, 0.5, 1.0, 1.5, 3.0, 5.0, 10.0, 25.0, 50.0, 75.0, 100.0)$. For each run, we calculated the mAA score of the results. The insensitivity of a method is measured as the weighted average of the mAA scores as follows:

$$\frac{\sum_{i=1}^n (t_i - t_{i-1})\mathrm{mAA}(t_i)}{\sum_{i=1}^n (t_i - t_{i-1})} = \frac{1}{t_n}\sum_{i=1}^n (t_i - t_{i-1})\mathrm{mAA}(t_i), \tag{16}$$

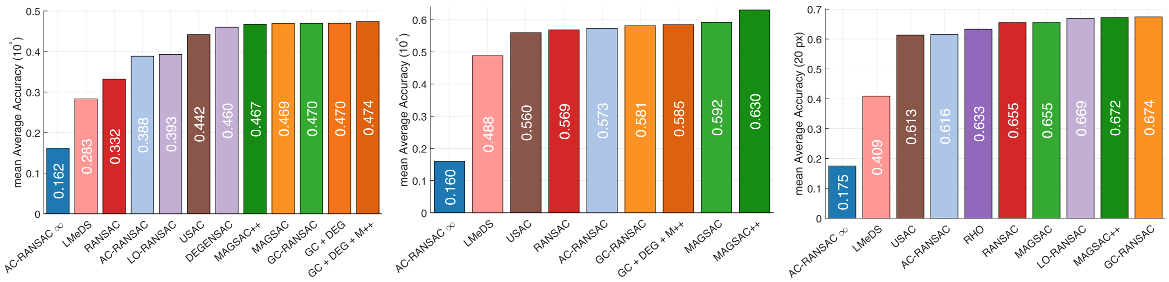
where $t_0 = 0$ and $\mathrm{mAA}(t_i)$ is the mAA score of a method after running it with threshold $t_i$. Formula (16) approximates the area under the mAA curve when plotted as the function of the inlier-outlier threshold used for the estimation.

In the rest of the paper, we call (16) the *insensitivity* measure. Note that measuring purely the insensitivity without including the accuracy of a method would require normalizing (16) by the maximum mAA value. We avoid this to make the insensitivity scores interpretable on their own. For example, (16) equals to 1 only if the method returns the perfect solution independently of the threshold.

### 7.1 Fundamental Matrix Estimation

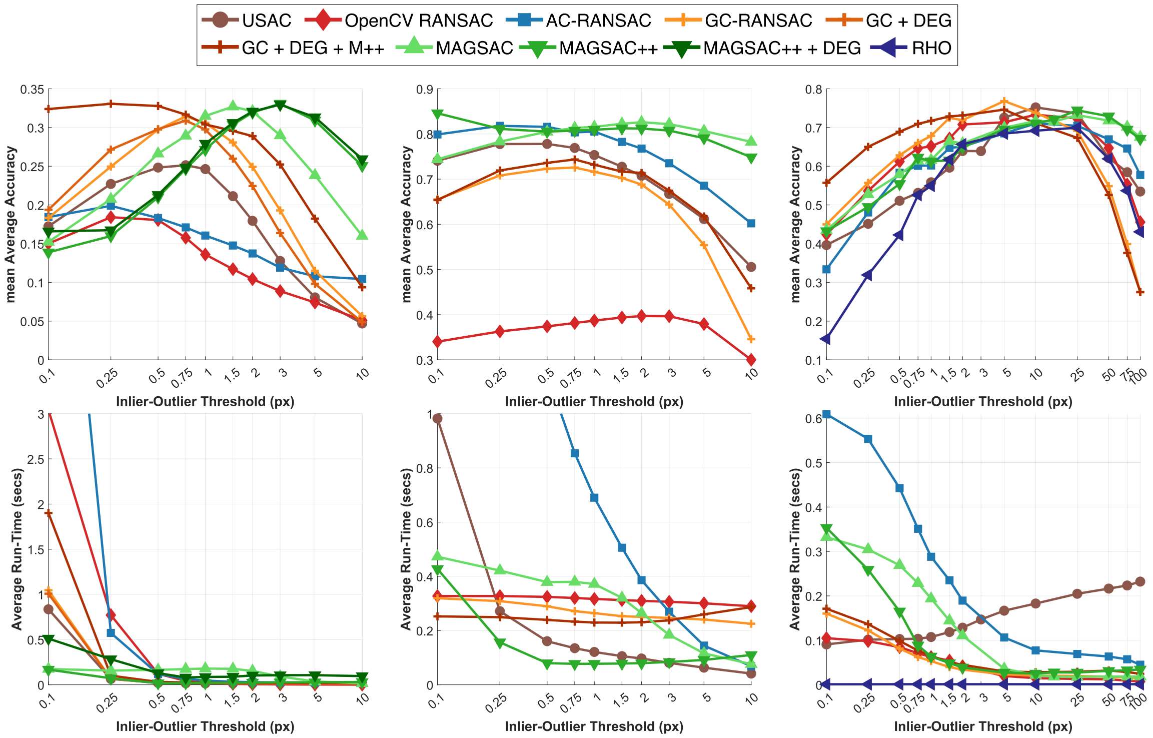
The methods compared for fundamental matrix estimation are OpenCV RANSAC [1], OpenCV LMedS [50], LO-RANSAC [26], LO-RANSAC + DEGENSAC [41], GC-RANSAC [51], GC-RANSAC + DEGENSAC, USAC [40], AC-RANSAC [30], MAGSAC, MAGSAC++, and GC-RANSAC with MAGSAC++ quality function and DEGENSAC. AC-RANSAC is a method setting the threshold adaptively. We tested two settings, i.e., with (AC-RANSAC) and without (AC-RANSAC $\infty$) an upper bound on threshold. The upper bound was tuned on the test set similarly as the parameters of the other tested methods.

The data are from the CVPR IMW 2020 PhotoTourism challenge. Correspondences were obtained using RootSIFT features and mutual nearest neighbour matching. We used all scenes from the test set, i.e., Sacre Coeur, St Peters Square, Brandenburg Gate, Buckingham Palace, Colosseum Exterior, Grand Place Brussels, Notre Dame Front Facade, Palace of Westminster, Pantheon Exterior, Prague Old Town Square, Taj Mahal, Temple Nara Japan, Trevi Fountain, Westminster Abbey. From the validation set, we used only scene British Museum to tune the hyper-parameters of the methods. Each scene contains 4950 image pairs. The reported accuracy is



(a) Fundamental matrix estimation (b) Relative pose estimation (c) Homography estimation

Fig. 3. The mean Average Accuracy of the tested robust estimators on fundamental matrix, relative pose and homography estimation. For each problem, the methods are ordered according to their scores. We used all scenes from the test set of the CVPR IMW 2020 PhotoTourism challenge. For F and E estimation, the methods were tested on a total of 54450 image pairs. Abbrevations used: OpenCV RANSAC (RANSAC), GC-RANSAC + DEGENSAC (GC + DEG), GC-RANSAC + DEGENSAC + MAGSAC++ scoring (GC + DEG + M++). *Higher value is better.*



(a) Fundamental matrix estimation (b) Relative pose estimation (c) Homography estimation

Fig. 4. The mean Average Accuracy (top row; higher is better) and average processing time (bottom; in seconds; lower is better) plotted as the function of the inlier-outlier threshold (or its upper limit; horizontal axis) parameter. For fundamental matrix and relative pose estimation, only scene British Museum was used. Homographies were estimated from both the EVD and HPatches datasets. The threshold (horizontal axis) is shown on a *logarithmic scale* – the right half of the plots covers a significantly larger area than the left one.

calculated on the total of 54450 image pairs from the test set using the parameters tuned on scene British Museum.

The results on the test set are shown in Fig. 3a. It can be seen that MAGSAC, MAGSAC++, GC-RANSAC, GC-RANSAC + DEGENSAC, and GC+RANSAC + DEGENSAC with MAGSAC++ quality function leads to similar accuracy. The maximum mAA difference between their results is 0.007. The most accurate results are obtained by GC-RANSAC with DEGENSAC and the proposed MAGSAC++ quality function. The other methods which do not need to a set a single threshold value, i.e., AC-RANSAC and LMeDS, are significantly less accurate. AC-RANSAC when applied without an upper bound (AC-RANSAC $\infty$) fails to return reasonable solutions in most of the cases. With an upper bound, it is more accurate than the RANSAC implemented in OpenCV.

The first row of Fig. 4a plots the mAA scores on scene British Museum as the function of the inlier-outlier threshold used for the estimation. We chose this scene since it is the first one in the validation set when the scene names are ordered alphabetically. All methods expect for MAGSAC and MAGSAC++, have a similar trend, i.e., their results increase slightly in the beginning while the threshold approaches its optimal value – for example, 0.75 px for USAC. Then their accuracy starts dropping dramatically. The trend of MAGSAC and MAGSAC++ is different. If the maximum threshold is set to a too low value, e.g., $< 1$ px, the results are inaccurate as it is expected. Between 1 and 10 pixels, the results are reasonably stable. This range is much wider than for the other methods which are only stable in-between $0.5 - 1.5$ pixels. Graph-Cut RANSAC with DEGENSAC and the proposed MAGSAC++ scoring shows an interesting trend, since it leads to almost constant mAA score in-between $0.1 - 1.5$ px threshold, then it starts deteriorating, however, less significantly than most of the other methods. The second row of Fig. 4a shows the processing time as the function of the threshold. It can be seen that MAGSAC++ is faster than MAGSAC as it is expected. It leads to similar processing time to its other less accurate alternatives.

TABLE 2
The Insensitivity (16) to the Inlier-Outlier Threshold (or its Upper Bound) is Shown on Fundamental Matrix (F), Essential Matrix (E), and Homography (H) Fitting. The best values are shown in red, the second best ones are in blue.

| | F | E | H | AVG |
|---|---|---|---|---|
| OpenCV RANSAC | 0.076 | 0.342 | 0.358 | 0.259 |
| OpenCV RHO | – | – | 0.329 | – |
| USAC | 0.096 | 0.590 | 0.452 | 0.379 |
| GC + DEG | 0.113 | – | – | – |
| AC-RANSAC | 0.118 | 0.670 | 0.421 | 0.403 |
| GC-RANSAC | 0.125 | 0.489 | 0.261 | 0.292 |
| GC (+ DEG) + M++ | 0.170 | 0.564 | 0.275 | 0.336 |
| MAGSAC | 0.215 | **0.797** | **0.519** | 0.510 |
| MAGSAC++ | 0.273 | 0.776 | 0.514 | **0.521** |
| MAGSAC++ + DEG | **0.279** | – | – | – |

The first column of Table 2 reports the threshold-insensitivity score on scene British Museum calculated as proposed in (16). MAGSAC++ combined with DEGENSAC yields the highest score and, thus, that method is the least sensitive to the setting of the inlier-outlier threshold – it is the easier to be used when applying robust estimation to a yet unseen scene.

### 7.2 Essential Matrix Estimation

The methods compared on relative pose (i.e., essential matrix) estimation are OpenCV RANSAC [1], OpenCV LMedS [50], LO-RANSAC [26], GC-RANSAC [51], USAC [40], AC-RANSAC [30], MAGSAC, MAGSAC++, and GC-RANSAC with MAGSAC++ quality function. DEGENSAC is not included in these tests since it is for recovering the fundamental matrix from scenes with dominant planar structures. For the five-point algorithm [46], planar scenes are not degenerate. Since the datasets used for fundamental matrix estimation contain the intrinsic camera parameters as well, we used the same scenes.

Fig. 3b shows that the most accurate essential matrices are clearly obtained by MAGSAC++ which achieves $\approx$4% higher mAA score than the second best MAGSAC. The other methods which do not need to a set a single threshold value, i.e., AC-RANSAC and LMeDS, are significantly less accurate, however, they are better than for fundamental matrix estimation. AC-RANSAC without an upper bound (AC-RANSAC $\infty$) fails to return reasonable solutions in most of the cases. With an upper bound, it is more accurate than OpenCV RANSAC and USAC.

The top row of Fig. 4b shows similar trend as for fundamental matrix estimation. All methods but MAGSAC and MAGSAC++ have a clear "best" threshold. If it is exceeded, their accuracy deteriorates dramatically. The results of MAGSAC and MAGSAC++ are *almost constant* throughout the range of thresholds. Interestingly, MAGSAC++ is the most accurate when the threshold upper bound is set to a small value, e.g., 0.1. Its results are just slightly less accurate for other threshold values. AC-RANSAC performs better here than for fundamental matrix estimation. The processing times are shown in the bottom row of Fig. 4b. MAGSAC++ is significantly faster for most of the threshold values than the other robust estimators. While AC-RANSAC leads to reasonable accuracy, it is significantly slower than the other methods.

### 7.3 Homography Estimation

For homography estimation, we used the `ExtremeView` [14] (`EVD`) and `HPatches` [16] datasets partitioned into test and validation sets as done in [47]. They consist of image pairs of different sizes from $329 \times 278$ up to $1712 \times 1712$ with point correspondences provided. The pairs of `EVD` undergo an extreme view change, i.e., wide baseline or extreme zoom. The `HPatches` scenes are extracted from a number of image sequences, where each sequence contains images of some planar object, e.g., a painting or a wall with graffiti. Since the datasets contain significantly fewer images then the ones used for epipolar geometry estimation, we repeated every method 100 times on each image pair. Besides the methods used for epipolar geometry estimation, we included the RHO [52] method implemented in OpenCV. The validation set was used to tune the hyperparameters of the methods. The accuracy is measured on the test set.

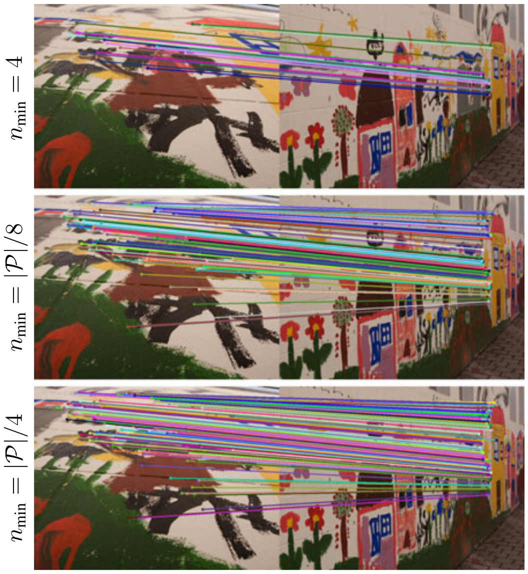


Fig. 5. The inliers of the estimated homography selected by the proposed adaptive strategy with varying the parameter $n_{\min}$ which controls the minimum number of required inliers.

It can be seen in Fig. 3c that the most accurate results are estimated by GC-RANSAC, MAGSAC++ and LO-RANSAC with a marginal difference of 0.002 – 0.005 in their mAA scores. The same trend can be observed for AC-RANSAC and LMeDS as before. AC-RANSAC with its threshold upper bound tuned works reasonably well. LMeDS fails to return accurate results. The mAA scores on the test set using varying threshold values are shown in the top row of Fig. 4c. Since the methods do not seem to be as sensitive to the inlier-outlier threshold as when fitting epipolar geometry, we tested a much wider range 0.1 – 100 than previously. The performance of MAGSAC and MAGSAC++ is very stable if $\sigma_{\max}$ is chosen from interval $[5, 100]$, where the accuracy difference is small. They achieve their maximum accuracy at $\sigma_{\max} = 25$, however, the accuracy drops only marginally for higher values. The bottom plot of Fig. 4c shows the processing time in seconds as the function of the inlier-outlier threshold. If the threshold is set to a small value ($\leq 1$) all methods, except RHO, gets slow. However, if $\tau(\sigma)$ or $\tau(\sigma_{\max})$ is greater than 3 the proposed MAGSAC and MAGSAC++ is similarly fast as the other methods running at real-time speed. While RHO is extremely fast for all settings, it is reasonably accurate only for a narrow range of thresholds, where all the other methods are similarly fast.

### 7.4 Iteratively Re-Weighted Least-Squares on 2D Lines

We compare the proposed iterative re-weighting strategy without the other components of MAGSAC++. To do so, we generated 100 2D points stemming from a 2D line and outliers. The outliers were generated uniformly randomly within a window of size $1000 \times 1000$. A 2D line passing through the middle of the window is generated with a random normal.

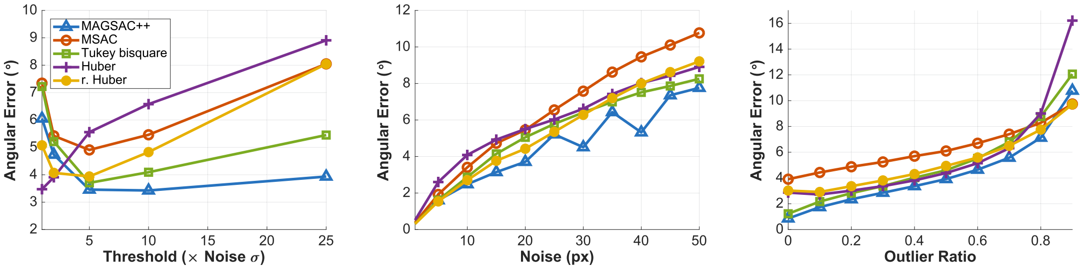


Fig. 6. The average results of iteratively re-weighted least-squares fitting using different robust weights (i.e., the proposed MAGSAC++, MSAC, Tukey bisquare, Huber and redescending Huber weights) when fitting 2D lines. The methods were repeated 10000 times using each parameter setting. (*Left*) The angular error, in degrees, of the estimated lines are plotted as the function of inlier-outlier threshold multiplier. The actual threshold is calculated by multiplying the noise $\sigma$ by the values shown on the horizontal axis. (*Middle*) The angular error is plotted as the function of the noise $\sigma$ added to the point coordinates. (*Right*) The angular error is plotted as the function of the outlier ratio.

Points were sampled from the line uniformly randomly and, then, zero-mean Gaussian-noise was added to their coordinates. We tested the following parameters: noise $\sigma \in \{0, 5, \ldots, 50\}$; outlier ratio $\mu \in \{0.0, 0.1, \ldots, 0.9\}$; threshold multiplier $\tau \in \{1, 2, 5, 10, 25\}$. The actual inlier-outlier threshold is calculated by multiplying $\tau$ with the noise scale $\sigma$. For each configuration, 10000 tests were run.

Fig. 6 plots the average angular errors (in degrees) as the function of the tested parameters. The compared robust weighting techniques are the proposed MAGSAC++; MSAC, assigning weight 1 if the point is closer than the threshold and, otherwise 0; Tukey bi-square weighting; Huber weights and re-descending Huber weights. It be seen that the MAGSAC++ weights guide the IRLS more successfully than the other compared techniques. Thus, the final errors of MAGSAC++ are smaller if threshold is set reasonably large. Also, it is the least sensitive to over-estimating the threshold value – its results are just slightly affected even if the actual threshold is 25 times the noise scale. Note that the offset errors of the estimated lines show a similar trend.

## 7.5 Inlier Selection

To test the proposed inlier selection, we generated a synthetic scene similarly as in the previous section. We compared the proposed technique with MINPRAN [29] and a contrario RANSAC [30]. We measured the average model error (15), in pixels, and the number of returned inliers. All algorithms got the ground truth line as input to select the inliers. Each test was repeated 10000 times. The results are shown in Fig. 7. The average model accuracy (left) and the number of inliers returned (right) of the compared adaptive threshold selection techniques are plotted as the function of the image noise, in pixels. From the left plot, it can be seen that the proposed technique returns inlier sets which lead to significantly more similar models, to the input one, than the other algorithms. The average model error of the proposed method for inlier ratio 0.1 is lower than the error of the other method for inlier ratio 0.9. For the fair comparison, it is important to note that MINPRAN and AC-RANSAC solve a different problem, i.e., selecting the noise scale which minimizes the randomness of the points which fall closer than the threshold. Their objective function is designed to select both the model and noise scale together. In our case, the input model is accurate and, therefore, we only need a set of inliers leading to a similarly accurate model.

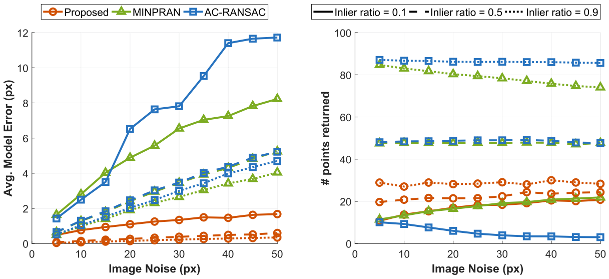


Fig. 7. The avg. model error (left) and the number of returned inliers (right) of adaptive threshold selection techniques are plotted as the function of the image noise (in pixel). Synthetic scene: points from a 2D line with zero-mean Gaussian-noise and uniformly distributed outliers (in total, 100 points), 10000 runs on each setting.

From the right plot of Fig. 7, it can be seen that the proposed inlier selection usually returns fewer points than the other methods if the inlier ratio is higher than 0.1. The number of points that suffices depends on a particular

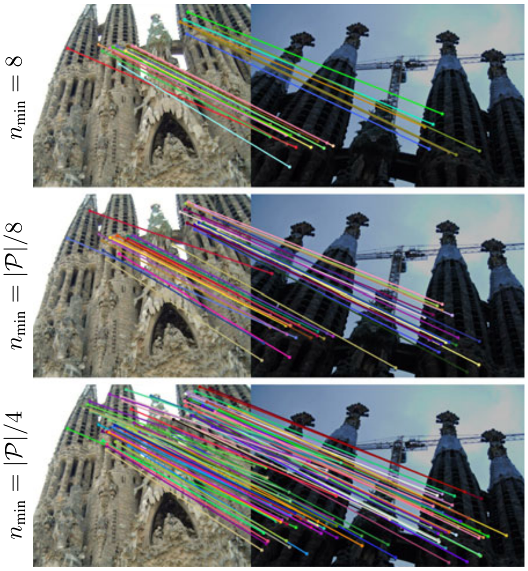


Fig. 8. The inliers of the estimated fundamental matrix selected by the proposed adaptive strategy with varying the parameter $n_{\min}$ which controls the minimum number of required inliers.

application where the proposed method is used. For example, for doing a cheirality check after decomposing an essential matrix, a few correspondences are usually enough, while a scene reconstruction might need many points. Setting the minimum number of points required to $n_{\min}$ is straightforward by initially including the $n_{\min}$ points with the lowest residuals. The algorithm starts adding new points from the $(n_{\min}+1)$th closest one. The upper bound of $n_{\min}$ is the number of points with residuals smaller than $\tau(\sigma_{\max})$.

Example scenes showing the proposed adaptive inlier selection with different values for $n_{\min}$ in the cases of homography and fundamental matrix estimation are shown in Figs. 5 and 8, respectively. Three different values are tested for $n_{\min}$ which are $m$ (4 for homographies; 8 for fundamental matrices), $|\mathcal{P}|/8$ and $|\mathcal{P}|/4$. In these examples, all the selected inliers are correct. Moreover, a reasonable number of inliers are returned even when $n_{\min} = m$. Note that even if the ground truth inlier number is lower than, e.g., $|\mathcal{P}|/4$, the algorithm is guaranteed to return the inliers which lead to an as similar model as possible to the input one.

## 8 Conclusion

We formulate a novel marginalization procedure as an iteratively re-weighted least-squares (IRLS) approach. We introduce a new model quality (scoring) function, that is increased by this IRLS approach, and a termination criterion for RANSAC-like robust estimation that does not require a crisp inlier-outlier decision. Also, a new method for adaptive inlier selection is proposed assuming that an accurate model is known. Combining the proposed techniques, the "bells and whistles" of USAC [40], e.g., pre-emptive verification, degeneracy testing, and a number of technical improvements, we propose MAGSAC++.

To the experiments, MAGSAC++ leads to the most accurate relative pose estimation. When all methods are tested using their "best" inlier-outlier thresholds, the most accurate fundamental matrices are obtained by combining the proposed quality function with GC-RANSAC [51]. For homography estimation, MAGSAC++ is the second most accurate method with only marginally higher errors than first one, i.e., GC-RANSAC. In practice, this "best" threshold is usually unknown. In those cases, both MAGSAC and MAGSAC++ are significantly less sensitive to the setting of the noise scale or its upper limit than the other state-of-the-art robust estimators. The source code and examples implemented both in C++ and Python are available at https://github.com/danini/magsacand in OpenCV.

## Acknowledgments

This work was supported in part by the Czech Science Foundation under Grant GA18-05360S, in part by the Ministry of Education OP VVV Project CZ.02.1.01/0.0/0.0/16 019/0000765 Research Center for Informatics, in part by the Ministry of Innovation and Technology NRDI Office within the framework of the Autonomous Systems National Laboratory Program, and in part by the Artificial Intelligence National Laboratory Program.

## References

[1] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[2] P. H. S. Torr and D. W. Murray, "Outlier detection and motion segmentation," in *Optical Tools for Manufacturing and Advanced Automation*. Bellingham, WA, USA: Int. Soc. Opt. Photon., 1993, pp. 432–443.

[3] P. H. S. Torr, A. Zisserman, and S. J. Maybank, "Robust detection of degenerate configurations while estimating the fundamental matrix," *Comput. Vis. Image Understanding*, vol. 71, no. 3, pp. 312–333, 1998.

[4] P. Pritchett and A. Zisserman, "Wide baseline stereo matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, 1998, pp. 754–760.

[5] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image Vis. Comput.*, vol. 22, no. 10, pp. 761–767, 2004.

[6] D. Mishkin, J. Matas, and M. Perdoch, "MODS: Fast and robust method for two-view matching," 2015, *arXiv:1503.02619*.

[7] C. Sminchisescu, D. Metaxas, and S. Dickinson, "Incremental model-based estimation using geometric constraints," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 727–738, May 2005.

[8] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4104–4113.

[9] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixelwise view selection for unstructured multi-view stereo," in *Eur. Conf. Comput. Vis.*, 2016, pp. 501–518.

[10] D. Ghosh and N. Kaabouch, "A survey on image mosaicking techniques," *J. Vis. Commun. Image Representation*, vol. 34, pp. 1–11, 2016.

[11] M. Zuliani, C. S. Kenney, and B. S. Manjunath, "The multiRANSAC algorithm and its application to detect planar homographies," in *Proc. IEEE Int. Conf. Image Process.*, 2005, pp. III–153.

[12] H. Isack and Y. Boykov, "Energy-based geometric multi-model fitting," *Int. J. Comput. Vis.*, vol. 97, pp. 123–147, 2012.

[13] T. T. Pham, T.-J. Chin, K. Schindler, and D. Suter, "Interacting geometric priors for robust multimodel fitting," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4601–4610, Oct. 2014.

[14] K. Lebeda, J. Matas, and O. Chum, "Fixing the locally optimized RANSAC," in *Proc. Brit. Mach. Vis. Conf.* 2012, pp. 95.1–95.11.

[15] E Trulls, Y. Jun, K. Yi, D. Mishkin, J. Matas, and P. Fua, "Image matching challenge," 2020. [Online]. Available: http://cmp.felk.cvut.cz/cvpr2020-ransac-tutorial/

[16] V. Balntas, K. Lenc, A. Vedaldi, and K. Mikolajczyk, "HPatches: A benchmark and evaluation of handcrafted and learned local descriptors," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3852–3861.

[17] P. H. Torr, S. J. Nasuto, and J. M. Bishop, "NAPSAC: High noise, high dimensional robust estimation-it's in the bag," in *Proc. Brit. Mach. Vis. Conf.*, 2002, pp. 458–467.

[18] K. Ni, H. Jin, and F. Dellaert, "GroupSAC: Efficient consensus in the presence of groupings," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 2193–2200.

[19] O. Chum and J. Matas, "Matching with PROSAC-progressive sample consensus," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 220–226.

[20] D. Barath, J. Noskova, M. Ivashechkin, and J. Matas, "MAGSAC++, a fast, reliable and accurate robust estimator," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1301–1309.

[21] O. Chum and J. Matas, "Randomized RANSAC with TDD test," in *Proc. Brit. Mach. Vis. Conf.*, 2002, vol. 2, pp. 448–457.

[22] D. P. Capel, "An effective bail-out test for RANSAC consensus scoring," in *Proc. Brit. Mach. Vis. Conf.*, 2005, pp. 78.1–78.10.

[23] J. Matas and O. Chum, "Randomized RANSAC with sequential probability ratio test," in *Proc. 10th IEEE Int. Conf. Comput. Vis.* Volume 1, 2005, vol. 2, pp. 1727–1732.

[24] O. Chum and J. Matas, "Optimal randomized RANSAC," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 8, pp. 1472–1482, Aug. 2008.

[25] A. Wald, *Sequential Analysis.* Chelmsford, MA, USA: Courier Corporation, 2004.

[26] O. Chum, J. Matas, and J. Kittler, "Locally optimized ransac," in *Proc. Joint Pattern Recognit. Symp.*, 2003, pp. 236–243.

[27] P. H. S. Torr and A. Zisserman, "MLESAC: A new robust estimator with application to estimating image geometry," *Comput. Vis. Image Understanding*, vol. 78, no. 1, pp. 138–156, 2000.

[28] P. H. S. Torr, "Bayesian model estimation and selection for epipolar geometry and generic manifold fitting," *Int. J. Comput. Vis.*, vol. 50, pp. 35–61, 2002.

[29] C. V. Stewart, "MINPRAN: A new robust estimator for computer vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 10, pp. 925–938, Oct. 1995.

[30] L. Moisan, P. Moulon, and P. Monasse, "Automatic homographic registration of a pair of images, with a contrario elimination of outliers," *Image Process. Line*, vol. 2, pp. 56–73, 2012.

[31] A. Cohen and C. Zach, "The likelihood-ratio test and efficient robust estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2282–2290.

[32] R. Raguram and J.-M. Frahm, "Recon: Scale-adaptive robust estimation via residual consensus," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1299–1306.

[33] M. Rais, G. Facciolo, E. Meinhardt-Llopis , J.-M. Morel, A. Buades, and B. Coll, "Accurate motion estimation through random sample aggregated consensus," 2017, *arXiv: 1701.05268*.

[34] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4104–4113.

[35] D. Barath, J. Noskova, and J. Matas, "MAGSAC: marginalizing sample consensus," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10189–10197.

[36] R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera, *Robust Statistics: Theory and Methods (with R)*. Hoboken, NJ, USA: Wiley, 2019.

[37] J. J. Moré, "The Levenberg-Marquardt algorithm: Implementation and theory," in *Numerical Analysis*. Berlin, Germany: Springer, 1978, pp. 105–116.

[38] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[39] C. G. Baker, K. A. Gallivan, and P. Van Dooren , "Low-rank incremental methods for computing dominant singular subspaces," *Linear Algebra Appl.*, vol. 436, no. 8, pp. 2866–2888, 2012.

[40] R. Raguram, O. Chum, M. Pollefeys, J. Matas, and J.-M. Frahm, "USAC: A universal framework for random sample consensus," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 2022–2038, Aug. 2013.

[41] O. Chum, T. Werner, and J. Matas, "Two-view geometry estimation unaffected by a dominant plane," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 772–779.

[42] H. Haber. Three-dimensional proper and improper rotation matrices. 2011. [Online]. Available: http://scipp.ucsc.edu/haber/ph116A/rotation_11.pdf

[43] O. Chum, T. Werner, and J. Matas, "Epipolar geometry estimation via RANSAC benefits from the oriented epipolar constraint," in *Proc. Int. Conf. Pattern Recognit.*, 2004, pp. 112–115.

[44] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. Int. Conf. Comput. Vis.*, 1999, pp. 1150–1157.

[45] R. I. Hartley, "In defense of the eight-point algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 6, pp. 580–593, Jun. 1997.

[46] H. Stewénius, D. Nistér, F. Kahl, and F. Schaffalitzky, "A minimal solution for relative pose with unknown focal length," *Image Vis. Comput.*, vol. 26, no. 7, pp. 871–877, 2008.

[47] D. Barath, T.-J. Chin, O. Chum, D. Mishkin, R. Ranftl, and J. Matas, "RANSAC in 2020 tutorial," 2020. [Online]. Available: http://cmp.felk.cvut.cz/cvpr2020-ransac-tutorial/

[48] K. Moo Yi , E. Trulls, Y. Ono, V. Lepetit, M. Salzmann, and P. Fua, "Learning to find good correspondences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2666–2674.

[49] Y. Jin *et al.*, "Image matching across wide baselines: From paper to practice," 2020, *arXiv:2003.01587*.

[50] P. J. Rousseeuw, "Least median of squares regression," *J. Amer. Statist. Assoc.*, vol. 79, no. 388, pp. 871–880, 1984.

[51] D. Barath and J. Matas, "Graph-cut RANSAC," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6733–6741.

[52] O. Bilaniuk, H. Bazargani, and R. Laganiere, "Fast target recognition on mobile devices: revisiting gaussian elimination for the estimation of planar homographies," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2014, pp. 119–125.

**Daniel Barath** was born in Budapest in 1989. He received the PhD defense degree from Eotvos Lorand University in 2019. He is a member of the Visual Recognition Group, FEE, Czech Technical University, Prague, Czech Republic, and the Machine Perception Research Laboratory, Institute for Computer Science and Control, Budapest, Hungary. His research interests include robust model estimation and minimal methods in computer vision.

**Jana Noskova** received the master's degree in probability and mathematical statistics and the doctoral degree in robust statistics from Charles University, Prague, in 1996. She is currently an assistant professor with the Department of Mathematics and a researcher with Visual Recognition Group, Department of Cybernetics, Czech Technical University, Prague.

**Jiri Matas** is currently a professor with the Center for Machine Perception, Faculty of Electrical Engineering, Czech Technical University, Prague, Czech Republic. He has authored or coauthored more than 250 papers in computer vision and machine learning. His research interests include object recognition, image retrieval, tracking, sequential pattern recognition, invariant feature detection, and Hough transform and RANSAC-type optimization.

# B Robust scale-adaptive mean-shift for tracking

Vojir T., Noskova J., Matas J., *Robust scale-adaptive mean-shift for tracking*, Pattern Recognition Letters 49, pp. 250-258, 2014.
DOI: 10.1016/j.patrec.2014.03.025

Pattern Recognition Letters 49 (2014) 250–258

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

# Robust scale-adaptive mean-shift for tracking ☆

Tomas Vojir [a,*], Jana Noskova [b], Jiri Matas [a]

[a] The Center for Machine Perception, FEE CTU in Prague, Karlovo Namesti 13, 121 35 Prague 2, Czech Republic
[b] Faculty of Civil Engineering, CTU in Prague, Thakurova 7/2077, 166 29 Prague 6, Czech Republic

ARTICLE INFO

*Article history:*
Available online 13 April 2014

*Keywords:*
Object tracking
Mean-shift
Scale estimation
Background weighting

ABSTRACT

The mean-shift procedure is a popular object tracking algorithm since it is fast, easy to implement and performs well in a range of conditions. We address the problem of scale adaptation and present a novel theoretically justified scale estimation mechanism which relies solely on the mean-shift procedure for the Hellinger distance. We also propose two improvements of the mean-shift tracker that make the scale estimation more robust in the presence of background clutter. The first one is a novel histogram color weighting that exploits the object neighborhood to help discriminate the target called background ratio weighting (BRW). We show that the BRW improves performance of MS-like tracking methods in general. The second improvement boost the performance of the tracker with the proposed scale estimation by the introduction of a forward–backward consistency check and by adopting regularization terms that counter two major problems: scale expansion caused by background clutter and scale implosion on self-similar objects. The proposed mean-shift tracker with scale selection and BRW is compared with recent state-of-the-art algorithms on a dataset of 77 public sequences. It outperforms the reference algorithms in average recall, processing speed and it achieves the best score for 30% of the sequences – the highest percentage among the reference algorithms.

## 1. Introduction

The mean-shift (MS) algorithm by Fukunaga and Hostetler [4] is a non-parametric mode-seeking method for density functions. It was introduced to computer vision by Comaniciu et al. [3] who proposed its use for object tracking. The MS algorithm tracks by minimizing the distance between two probability density functions (pdfs) represented by a target and target candidate histograms. Since the histogram distance (or, equivalently, similarity) does not depend on the spatial structure within the search window, the method is suitable for deformable and articulated objects.

The performance of the mean-shift algorithm suffers from the use of a fixed size window if the scale of the target changes. When the projection of the tracked object becomes larger, localization becomes poor since some pixels on the object are not included in the search window and the similarity function often has many local maxima. If the object become smaller, the kernel window includes background clutter which often leads to tracking failure.

The seminal paper by Comaniciu et al. [3] already considered the problem and proposed changing the window size over multiple runs by a constant factor (±10%). The window size maximizing the similarity to the target histogram was chosen. This approach does not cope well with the increase of the object size since the smaller windows usually have higher similarity and therefore the scale is often underestimated.

Collins [2] exploited image pyramids and used an additional mean-shift procedure for scale selection after estimating the position. The method works well for objects with a fixed aspect ratio, but this often does not hold for non-rigid or a deformable objects. Moreover, the method is significantly slower than the standard MS.

Image moments are used in [1,10] to determine the scale and orientation of the target. The second moments are computed from an image of weights that are proportional to the probability that a pixel belongs to the target model. Yang et al. [13] introduced a new similarity measure that estimates the scale by comparison of second moments of the target model and the target candidate.

Pu and Peng [11] assume target rigidity and restrict motion to scaling and translation. The target is first tracked using the mean-shift both in the forward and backward direction to estimate the translation. Scale is then estimated from feature points matched by an M-estimator with outlier rejection. Similarly, [8,15] rely on ”support features” for scale estimation after the mean-shift algorithm solves for position. Liang et al. [8] search for the target boundary by correlating the image with four

☆ This paper has been recommended for acceptance by J.K. Kämäräinen.
* Corresponding author. Tel.: +420 224 355 767; fax: +420 224 357 385.
*E-mail address:* vojirtom@cmp.felk.cvut.cz (T. Vojir).

http://dx.doi.org/10.1016/j.patrec.2014.03.025
0167-8655/© 2014 Elsevier B.V. All rights reserved.

templates. Positions of the boundaries directly determine the scale of the target. Zhao et al. [15] exploit affine structure to recover the target relative scale from feature point correspondences between consecutive frames.

Methods depending on feature matching are able to robustly estimate the scale, but they cannot be seamlessly integrated to the mean-shift framework. Moreover, estimating scale from feature correspondences takes times, requires presence of well-localised features that can be detected with high repeatability, and it has difficulties dealing with a non-rigid or a deformable object.

We present a theoretically justified scale estimation mechanism which, unlike the method listed above, relies solely on the mean-shift procedure for the Hellinger distance. Furthermore, we propose a formulation for background weighting that exploits the tracked object's neighborhood to help discriminate the object from the background. Additionally, we present two mechanisms that make the scale estimation more robust in the presence of background clutter and improve tracker performance to level of the state-of-the-art. The performance is compared to state-of-the-art algorithms on a large tracking dataset.

## 2. Mean-shift tracker with scale estimation

### 2.1. Standard kernel-based object tracking

In the standard mean-shift tracking of [3], the target is modelled as an $m$-bin kernel-estimated histogram in a feature space located at the origin:

$$\hat{\mathbf{q}} = \{\hat{q}_u\}_{u=1\dots m} \quad \sum_{u=1}^{m} \hat{q}_u = 1. \tag{1}$$

A target candidate at location $\mathbf{y}$ in the subsequent frame is described by its histogram

$$\hat{\mathbf{p}}(\mathbf{y}) = \{\hat{p}_u(\mathbf{y})\}_{u=1\dots m} \quad \sum_{u=1}^{m} \hat{p}_u = 1; \tag{2}$$

Let $\mathbf{x}_i$ denote pixel locations, $n$ be the number of pixels of the target and let $\{\mathbf{x}_i^*\}_{i=1\dots n}$ be the pixel locations of the target centered at the origin. Spatially, the target covers a unit circle and an isotropic, convex and monotonically decreasing kernel profile $k(x)$ is used. Function $b : R^2 \to 1\dots m$ maps the value of the pixel at location $\mathbf{x}_i$ to the index $b(\mathbf{x}_i)$ of the corresponding bin in the feature space. The probability of the feature $u \in \{1,\dots,m\}$ is estimated by the target histogram as follows:

$$\hat{q}_u = C \sum_{i=1}^{n} k\left(\|\mathbf{x}_i^*\|^2\right) \delta[b(\mathbf{x}_i^*) - u], \tag{3}$$

where $\delta$ is the Kronecker delta and $C$ is a normalization constant so that $\sum_{u=1}^{m} \hat{q}_u = 1$.

Let $\{\mathbf{x}_i\}_{i=1\dots n_h}$ be pixel locations in the current frame where the target candidate is centered at location $\mathbf{y}$ and $n_h$ be the number of pixels of the target candidate. Using the same kernel profile $k(x)$, but with a scale parameter $h$, the probability of the feature $u = 1\dots m$ in the target candidate is

$$\hat{p}_u(\mathbf{y}) = C_h \sum_{i=1}^{n_h} k\left(\left\|\frac{\mathbf{y} - \mathbf{x}_i}{h}\right\|^2\right) \delta[b(\mathbf{x}_i) - u], \tag{4}$$

where $C_h$ is a normalization constant. The difference between probability distributions $\hat{\mathbf{q}} = \{\hat{q}_u\}_{u=1\dots m}$ and $\{\hat{p}_u(\mathbf{y})\}_{u=1\dots m}$ is measured by the Hellinger distance of probability measures, which is known to be a metric:

$$H(\hat{\mathbf{p}}(\mathbf{y}), \hat{\mathbf{q}}) = \sqrt{1 - \rho[\hat{\mathbf{p}}(\mathbf{y}), \hat{\mathbf{q}}]}, \tag{5}$$

where

$$\rho[\hat{\mathbf{p}}(\mathbf{y}), \hat{\mathbf{q}}] = \sum_{u=1}^{m} \sqrt{\hat{p}_u(\mathbf{y}) \hat{q}_u} \tag{6}$$

is the Bhattacharyya coefficient of $\hat{\mathbf{q}}$ and $\hat{\mathbf{p}}(\mathbf{y})$. Minimizing the Hellinger distance is equivalent to maximizing the Bhattacharyya coefficient $\rho[\hat{\mathbf{p}}(\mathbf{y}), \hat{\mathbf{q}}]$. The search for the new target location in the current frame starts at location $\hat{\mathbf{y}}_0$ of the target in the previous frame using gradient ascent with a step size equivalent to the mean-shift method. The kernel is repeatedly moved from the current location $\hat{\mathbf{y}}_0$ to the new location

$$\hat{\mathbf{y}}_1 = \frac{\sum_{i=1}^{n_h} \mathbf{x}_i w_i g\left(\left\|\frac{(\hat{\mathbf{y}}_0 - \mathbf{x}_i)}{h}\right\|^2\right)}{\sum_{i=1}^{n_h} w_i g\left(\left\|\frac{(\hat{\mathbf{y}}_0 - \mathbf{x}_i)}{h}\right\|^2\right)}, \tag{7}$$

where

$$w_i = \sum_{u=1}^{m} \sqrt{\frac{\hat{q}_u}{\hat{p}_u(\hat{\mathbf{y}}_0)}} \delta[b(\mathbf{x}_i) - u] \tag{8}$$

and $g(x) = -k'(x)$ is the derivative of $k(x)$, which is assumed to exist for all $x \geqslant 0$, except for a finite set of points.

### 2.2. Scale estimation

Let us assume that the scale changes frame to frame in an isotropic manner[1]. Let $\mathbf{y} = (y^1, y^2)^T$, $\mathbf{x}_i = (x_i^1, x_i^2)^T$ denote pixel locations and $N$ be the number of pixels in the image. A target is represented by an ellipsoidal region $\frac{(x_i^{*1})^2}{a^2} + \frac{(x_i^{*2})^2}{b^2} < 1$ in the image and an isotropic kernel with profile $k(x)$ as in [3], restricted by a condition $k(x) = 0$ for $x \geqslant 1$, is used. The probability of the feature $u \in \{1,..,m\}$ is estimated by the target histogram as

$$\hat{q}_u = C \sum_{i=1}^{N} k\left(\frac{(x_i^{*1})^2}{a^2} + \frac{(x_i^{*2})^2}{b^2}\right) \delta[b(\mathbf{x}_i^*) - u], \tag{9}$$

where $C$ is a normalization constant. Let $\{\mathbf{x}_i\}_{i=1\dots N}$ be the pixel locations of the current frame in which the target candidate is centered at location $\mathbf{y}$. Using the same kernel profile $k(x)$, the probability of the feature $u = 1\dots m$ in the target candidate is given by

$$\hat{p}_u(\mathbf{y}, h) = C_h \sum_{i=1}^{N} k\left(\frac{(y^1 - x_i^1)^2}{a^2 h^2} + \frac{(y^2 - x_i^2)^2}{b^2 h^2}\right) \delta[b(\mathbf{x}_i) - u], \tag{10}$$

where

$$C_h = \frac{1}{\sum_{i=1}^{N} k\left(\frac{(y^1 - x_i^1)^2}{a^2 h^2} + \frac{(y^2 - x_i^2)^2}{b^2 h^2}\right)}. \tag{11}$$

The parameter $h$ defines the scale of the target candidate and thus the number of pixels with non-zero values of the kernel function.

For a given kernel and variable $h$, $C_h$ can be approximated in the following way: Let $n_1$ be the number of pixels in the ellipsoidal region of the target model, and let $n_h$ be the number of pixels in the ellipsoidal region of the target candidate with a scale $h$; then $n_h \doteq h^2 n_1$. Using the definition of Riemann integral we obtain:

$$\begin{aligned}\sum_{i=1}^{N} k\left(\frac{(x_i^1)^2}{a^2 h^2} + \frac{(x_i^2)^2}{b^2 h^2}\right) \frac{\pi a b h^2}{n_h} &\approx \int\int_{\left\{\frac{(x^1)^2}{a^2 h^2} + \frac{(x^2)^2}{b^2 h^2} < 1\right\}} k\left(\frac{(x^1)^2}{a^2 h^2} + \frac{(x^2)^2}{b^2 h^2}\right) dx^1 dx^2 \\ &= h^2 a b \int\int_{\|\mathbf{x}\| < 1} k(\|\mathbf{x}\|^2).\end{aligned} \tag{12}$$

[1] Generalization to the anisotropic where $\mathbf{h} = (h^1, h^2)^T$ is straightforward.

Therefore $C_h \approx C\frac{1}{h^2}$ and for any two values $h_0,\ h_1$ $C_{h_1} \approx C_{h_0}\frac{h_0^2}{h_1^2}$. For justification of the approximation see Appendix A.

As in [3] the difference between probability distribution $\hat{\mathbf{q}} = \{\hat{q}_u\}_{u=1\dots m}$ and $\{\hat{p}_u(\mathbf{y},h)\}_{u=1\dots m}$ is measured by the Hellinger distance. Using the approximations above for $C_h$ in some neighborhood of $h_0$ we get

$$\rho[\hat{\mathbf{p}}(\mathbf{y},h),\hat{\mathbf{q}}] \approx \hat{\rho}(\mathbf{y},h) = \sum_{u=1}^{m}\sqrt{C_{h_0}\frac{h_0^2}{h^2}\sum_{i=1}^{N}k\left(\frac{(y^1-x_i^1)^2}{a^2h^2}+\frac{(y^2-x_i^2)^2}{b^2h^2}\right)\delta[b(\mathbf{x}_i)-u]\hat{q}_u} \tag{13}$$

Thus, to minimize the Hellinger distance, function $\hat{\rho}(\mathbf{y},h)$ is maximized using a gradient method. In the proposed procedure, the kernel with a scale parameter $h_0$ is iteratively moved from the current location $\hat{\mathbf{y}}_0$ in direction of $\bigtriangledown\hat{\rho}(\hat{y}_0^1,\hat{y}_0^2,h_0)$ to the new location $\hat{\mathbf{y}}_1$, changing its scale to $h_1$. The basic idea of this procedure is the same as the mean-shift method.

Let us denote

$$w_i = \sum_{u=1}^{m}\sqrt{\frac{\hat{q}_u}{\hat{p}_u(\hat{\mathbf{y}}_0,h_0)}}\delta[b(\mathbf{x}_i)-u], \tag{14}$$

$$G = \sum_{i=1}^{N}w_i g\left(\frac{(\hat{y}_0^1-x_i^1)^2}{a^2h_0^2}+\frac{(\hat{y}_0^2-x_i^2)^2}{b^2h_0^2}\right), \tag{15}$$

and

$$\mathbf{m}_k(\hat{\mathbf{y}}_0,h_0) = \frac{\sum_{i=1}^{N}\mathbf{x}_i w_i g\left(\frac{(\hat{y}_0^1-x_i^1)^2}{a^2h_0^2}+\frac{(\hat{y}_0^2-x_i^2)^2}{b^2h_0^2}\right)}{G} - \hat{\mathbf{y}}_0, \tag{16}$$

where $\mathbf{m}_k(\hat{\mathbf{y}}_0,h_0) = (m_k^1(\hat{\mathbf{y}}_0,h_0), m_k^2(\hat{\mathbf{y}}_0,h_0))^T$. Then we get

$$\frac{\partial\hat{\rho}(\mathbf{y},h)}{\partial y^1}(\hat{\mathbf{y}}_0,h_0) = \frac{C_{h_0}}{a^2(h_0)^2}\cdot G\cdot m_k^1(\hat{\mathbf{y}}_0,h_0), \tag{17}$$

$$\frac{\partial\hat{\rho}(\mathbf{y},h)}{\partial y^2}(\hat{\mathbf{y}}_0,h_0) = \frac{C_{h_0}}{b^2(h_0)^2}\cdot G\cdot m_k^2(\hat{\mathbf{y}}_0,h_0) \tag{18}$$

and

$$\frac{\partial\hat{\rho}(\mathbf{y},h)}{\partial h}(\hat{\mathbf{y}}_0,h_0) = \frac{C_{h_0}}{(h_0)^2}\cdot G\cdot\left[\frac{1}{h_0}\frac{\sum_{i=1}^{N}w_i\cdot\left(\frac{(\hat{y}_0^1-x_i^1)^2}{a^2}+\frac{(\hat{y}_0^2-x_i^2)^2}{b^2}\right)\cdot g\left(\frac{(\hat{y}_0^1-x_i^1)^2}{a^2h_0^2}+\frac{(\hat{y}_0^2-x_i^2)^2}{b^2h_0^2}\right)}{G} - h_0\frac{\sum_{i=1}^{N}w_i\cdot k\left(\frac{(\hat{y}_0^1-x_i^1)^2}{a^2h_0^2}+\frac{(\hat{y}_0^2-x_i^2)^2}{b^2h_0^2}\right)}{G}\right]. \tag{19}$$

Finally, the mean-shift update of $\mathbf{y}$ and $h$ is obtained:

$$\hat{y}_1^1 = \frac{1}{a^2}m_k^1(\hat{\mathbf{y}}_0,h_0)+\hat{y}_0^1,\ \ \hat{y}_1^2 = \frac{1}{b^2}m_k^2(\hat{\mathbf{y}}_0,h_0)+\hat{y}_0^2 \tag{20}$$

$$h_1 = \left[1-\frac{\sum_{i=1}^{N}w_i\cdot k\left(\frac{(\hat{y}_0^1-x_i^1)^2}{a^2h_0^2}+\frac{(\hat{y}_0^2-x_i^2)^2}{b^2h_0^2}\right)}{G}\right]h_0+\frac{1}{h_0} \times\frac{\sum_{i=1}^{N}w_i\cdot\left(\frac{(\hat{y}_0^1-x_i^1)^2}{a^2}+\frac{(\hat{y}_0^2-x_i^2)^2}{b^2}\right)\cdot g\left(\frac{(\hat{y}_0^1-x_i^1)^2}{a^2h_0^2}+\frac{(\hat{y}_0^2-x_i^2)^2}{b^2h_0^2}\right)}{G}. \tag{21}$$

### *2.3. Background ration weighting*

Instead of maximizing the Bhattacharyya coefficient, we formulate the problem as ratio maximization, where the numerator and the denominator are defined as Bhattacharyya coefficients of target candidate and target and background respectively. We call this formulation *background ratio weighting* (BRW). Background histogram $\hat{\mathbf{bg}}$ is computed over the neighborhood of the target in the first frame and the ratio is obtained as follows:

$$R = \frac{\hat{\rho}[\hat{\mathbf{p}}(\mathbf{y},h),\hat{\mathbf{q}}]}{\hat{\rho}[\hat{\mathbf{p}}(\mathbf{y},h),\hat{\mathbf{bg}}]}. \tag{22}$$

Using a gradient ascent method for maximization of $log(R)$ we use the following formula with weights $w_i$ changed to weights $w_i^{bg}$, where

$$w_i^{bg} = \max\left[0,\sum_{u=1}^{m}\left(\frac{1}{\hat{\rho}[\hat{\mathbf{p}}(\hat{\mathbf{y}}_0,h_0),\hat{\mathbf{q}}]}\sqrt{\frac{\hat{q}_u}{\hat{p}_u(\hat{\mathbf{y}}_0,h_0)}} - \frac{1}{\hat{\rho}[\hat{\mathbf{p}}(\hat{\mathbf{y}}_0,h_0),3\hat{\mathbf{bg}}]}\sqrt{\frac{\hat{bg}_u}{\hat{p}_u(\hat{\mathbf{y}}_0,h_0)}}\right)\delta[b(\mathbf{x}_i)-u]\right]. \tag{23}$$

The max operator set the weights $w_i^{bg}$ to be a non-negative. In the case of a non-negative weights, the mean-shift algorithm preserves its convergence properties.

## 3. The tracking algorithm

Introducing scale estimation into the mean-shift procedure reveals two issues: Firstly, there is a difference in the MS behaviour when the position and scale estimation is imprecise. While errors in position are usually corrected later on during the mean-shift iteration, the error in scale estimation has no "self-correcting" ability in the presence of a non-trivial background. Secondly, the scale ambiguity of self-similar objects usually leads to underestimation of the scale and tracking failure (see Fig. 1).

To cope with this problem and make the tracking more robust, we propose a mean-shift algorithm with regularized scale estimation. The algorithm, denoted $\mathrm{MS}_s$, is summarized in Algorithm.

**Algorithm 1.** $\mathrm{MS}_s$ – mean-shift with regularized scale estimation.

```
Input: Target model $\hat{\mathbf{q}}$, starting position $\mathbf{y}_0$ and starting object
  size $\mathbf{s}_0$
Output: Position $\mathbf{y}_t$ and scale $h_t$
$t = 1$;
repeat
  Compute $\hat{p}_u(\mathbf{y}_{t-1},h_{t-1})$ using Eq. (10);
  Compute weights $w_i^{bg}$ according to Eq. (23);
  Update position $\mathbf{y}_t$ according to Eq. (20), neglecting the
    constants $a,\ b$ assuming that $a \approx b$;
  Update scale $h_t$ according to Eq. (21);
  Apply corrections $h_t$ = $h_t$ + Eq. (24) + Eq. (25);
$t = t+1$;
until $\|\mathbf{y}_t-\mathbf{y}_{t-1}\|^2 < \varepsilon$   OR   $t > maxIter$
```

The structure of the algorithm is similar to the standard mean-shift algorithm, except for the scale update step. Two regularization terms are introduced in the scale update step. The first term *rs* reflects our prior assumption that the target scale does not change dramatically; therefore, the change of scale is penalized according to Eq. (24):

$$rs(h) = \begin{cases} -\log(h) & |\log(h)| \leqslant b_2 \\ b_2 & \log(h) < -b_2 \\ -b_2 & \log(h) > b_2 \end{cases} \tag{24}$$

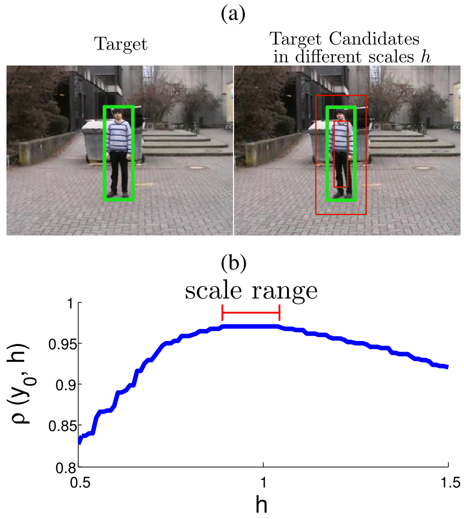


**Fig. 1.** Illustration of the scale ambiguity problem. (a) Target and target candidates at different scales with fixed center location (green rectangle corresponds to $h = 1$), (b) target candidate similarity with target as a function of the scale parameter measured by Bhattacharyya coefficient. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

where the $h$ is scaling factor and the function in absolute value is bounded by the constant $b_2$. The second term $rb$ addresses the problem of scale ambiguity by forcing the search window to include a portion of background pixels. In other words, from the possible range of scales (generated by the object self-similarity), a slight bias towards the largest is introduced. The $rb$ function is defined by Eq. (25):

$$rb(\mathbf{y},h) = \begin{cases} \varrho - B(\mathbf{y},h) & |\varrho - B(\mathbf{y},h)| \leqslant b_1 \\ -b_1 & \varrho - B(\mathbf{y},h) < -b_1 \\ b_1 & \varrho - B(\mathbf{y},h) > b_1 \end{cases} \tag{25}$$

where $(\mathbf{y},h)$ are the position and scaling factor and $\varrho$ define the percentage of weighted background pixels that should be contained in the search window. The function response lies in the interval $(-b_1, b_1)$. The percentage of weighted background pixels is computed as follows:

$$B(\mathbf{y},h) = \sum_{i=1}^{n}\delta[\hat{q}_{b(\mathbf{x}_i)}]\sum_{u=1}^{m}\hat{p}_u\delta[b(\mathbf{x}_i)-u] \bigg/ \sum_{i=1}^{n}\sum_{u=1}^{m}\hat{q}_u\delta[b(\mathbf{x}_i)-u] \tag{26}$$

where the numerator is the sum of bin weights of the target candidate for pixels in which the target model has $\hat{\mathbf{q}}_u = 0$, and the denominator is the sum of bin weights of the target model over all pixels.

The MS$_s$ algorithm works well for sequences with scale change, but for sequences without scale change or with a significant background clutter, the algorithm tends to estimate non-zero scale, which may lead to accumulation of incorrect scale estimates and a tracking failure. Therefore, we adopted a technique to validate the estimated scale change: *the Backward scale consistency check*. The Backward check uses reverse tracking from position $\mathbf{y}_t$ obtained by forward tracking and validates the estimated scale from step $t-1$ to $t$ and $t$ to $t-1$. This validation ensures that in the presence of background clutter the scale estimation does not "grow without bounds" and enables the tracker to recover from erroneous estimates. The algorithm using this technique is summarized in Algorithm 2, and we call it as Adaptive Scale mean-shift (ASMS).

**Algorithm 2.** ASMS – mean-shift with scale and backward consistency check

```
Input: Target model $\hat{\mathbf{q}}$, starting position $\mathbf{y}_0$ and starting object
  size $\mathbf{s}_0$
Output: Position and scale in each frame $(\mathbf{y}_t, \mathbf{s}_t)$, where
  $t \in \{1,\ldots,n\}$
foreach Frame $t \in \{1,\ldots,n\}$ do
  $[\mathbf{y}_t, h]$ = MS$_s(q, \text{image}_t, \mathbf{y}_{t-1}, \mathbf{s}_{t-1})$;
  if $|log(h)| > \Theta_s$ then
    //Scale change - proceed with consistency check
    $[\sim, h_{back}]$ = MS$_s(q, \text{image}_{t-1}, \mathbf{y}_t, h\mathbf{s}_{t-1})$;
      if $|log(h * h_{back})| > \Theta_c$ then
        //Inconsistent scales
  $\mathbf{s}_t = (1-\alpha-\beta)\mathbf{s}_{t-1} + \alpha\mathbf{s}_{\text{default}} + \beta h\mathbf{s}_{t-1}$; where
   $\alpha = c_1(\frac{\mathbf{s}_{\text{default}}}{\mathbf{s}_{t-1}})$;
  else
    $\mathbf{s}_t = (1-\gamma)\mathbf{s}_{t-1} + \gamma h\mathbf{s}_{t-1}$;
```

In the case of a detected scale inconsistency the object size is a weighted combination of three parts: (i) the previous size; (ii) the new estimated size; (iii) "default" size, which in our case is initial size of the object. The parameters for this combination were selected experimentally on the subset of testing sequences as a trade off between scale adaptability of the MS$_s$ and stability of the standard mean-shift algorithm.

We also noticed that mean-shift is more stable if the bandwidth size is biased toward a larger size so that the whole target is included; therefore, the computation of the weight $\alpha$ (Algorithm 2) is not symmetric but it prefers enlarging the object size. The default size is kept constant during tracking, and preliminary experiments with size adaptation show no significant benefit and only introduce error caused by incorrect updates. This can be explained by the character of the data, where the target scale usually oscillate around initial value.

## 4. Experimental protocol

Experiments were conducted on 77 sequences[2] collected from the literature. The sequences vary in length from dozens of frames to thousands, contain diverse object types (rigid, articulated) and have different scene settings (indoor/outdoor, static/moving camera, lightning conditions). Object occlusions and objects that disappear from the field of view are also present in the data.

The proposed mean-shift algorithm ASMS is compared with the standard published mean-shift algorithm (MS) and its scale adaptation (MS$_\pm$) proposed by Comaniciu et al. [3]. All algorithms are evaluated with and without the proposed background weighting.

The proposed method is also compared with the state-of-the-art tracking algorithms that are available as source code, namely SOAMST by Ning et al. [10] base on the mean-shift algorithm, LGT by Čehovin et al. [12], TLD by Kalal et al. [6], CT by Zhang et al. [14] and STRUCK by Hare et al. [5]. Parameters for these algorithms were left default as set by the authors. Note that our results for those algorithms may differ from results reported in other publications since we did not optimize their parameters for the best performance for each sequence as was done, e.g., by Zhang et al. [14], but were fixed for all experiments. Moreover, the target was initialized in the first frame using the ground truth position for all algorithms. Stochastic methods were run multiple times on each sequence and the average result was reported.

[2] http://cmp.felk.cvut.cz/~vojirtom/dataset.

Performance of the algorithms was measured by the recall: the number of correctly tracked frames divided by number of frames where the target is visible. Recall was chosen because some of the algorithms exhibit detector-like behavior; therefore, other frequently used criteria, such as first failure frame or failure frame from which the algorithm does not recover, will not capture the real performance of the algorithm, i.e. in how many frames the algorithm locates the target correctly.

A frame is considered tracked correctly if the overlap with the ground truth is higher than 0.5. The overlap is defined as $o = \frac{area(T \cap G)}{area(T \cup G)}$, where $T$ is object bounding box reported by the tracker and $G$ is ground truth bounding box.

To characterize the speed, the average running time per frame of each algorithm was measured. Note that the algorithms are not implemented in the same programming language (SOAMST, LGT, TLD, CT using matlab with MEX files, STRUCT and mean-shift using C++), which may bias the speed measurement towards the more efficient programming languages.

The proposed mean-shift algorithms are written in C++ without heavy optimization or multithreading. All parameters of the algorithm were fixed for all experiments. Some of the parameters are fairly standard (mean-shift termination criterion) and the rest were chosen empirically as follows: bounds for regularization terms $b_1 = 0.05,\ b_2 = 0.1$ and $\varrho = 0.5$; termination of the mean-shift algorithm $\varepsilon = 0.1$, and $maxIter = 15$; scale consistency check $\Theta_s = 0.05 \approx 5\%$ of the scale change, $\Theta_c = 0.1$; exponential averaging $c_1 = 0.1,\ \beta = 0.1$ and $\gamma = 0.3$. The pdf is represented as a histogram computed over the RGB space and quantized into the $16 \times 16 \times 16$ bins.

## 5. Results

### 5.1. Background weighting evaluation

The experiment evaluates the benefits of different histogram bin weighting based on the background. The proposed BRW method is implemented into a different MS algorithms (i.e. standard MS, the standard scale MS by Comaniciu et al. [3] and the proposed ASMS) and compared to direct histogram weighting (CBWH) proposed by Ning et al. [9].

Fig. 2 shows the recall for 77 sequences. In general, using background weighting improves MS performance. The BRW performs slightly better or equal than CBWH for the standard mean-shift algorithms and dominates for the proposed AMSM. The average recall for the evaluated methods is shown by dashed horizontal lines in the plots. From the experiment, we conclude that ASMS-BRW is superior to other combinations, and therefore, it is used in all subsequent experiments. When not specified otherwise, the abbreviation ASMS refers to ASMS-BRW.

Next, ASMS was compared with the scale adaptation proposed by Comaniciu et al. [3], denoted MS$_\pm$, which runs the MS Algorithm 3 times for different window sizes $(1, 1 \pm 0.1\%)$ and the result with the minimum distance to the target histogram is used. The comparison is included in Fig. 3 which also shows the results of the state of the art methods. ASMS outperforms MS$_\pm$ for average recall. It performs better on 48 sequences.

### 5.2. Comparison with the state-of-the-art methods

Result of the comparison of the ASMS and state-of-the-art algorithms is presented in Fig. 3, which shows that the performance of the ASMS tracker is comparable to the state-of-the-art methods, and on a large fraction of the sequences (30%) it is the top performer. However, Fig. 3 also shows that ASMS performs poorly on some sequences.

The results are summarized in two tables. Results for sequences with at least 30% object scale difference w.r.t the reference size in at least 20% frames of sequences are presented in Table 1. Performance on the remaining "small scale change" sequences is shown in Table 2. The last two rows show the mean performance and the number of sequences where the tracker performed best and second best.

There are some sequences in the set of the 32 sequences with object scale changes where tracking without a re-detection mechanism fails. These "Long-term" sequences with thousands of frames (e.g. *CarChase, Motocross, Panda, Volkswagen*) include object disappearance from the field of view, scene cuts, significant object occlusion and strong background clutter. Some shorter sequences with full object occlusion (e.g. *Vid_F*), cannot be successfully tracked without re-detection too. Since ASMS does not provide any re-detection ability, it can not handle these cases. In these sequences, the TLD tracker achieved the best results.

ASMS achieved the best score on the *Vid_X* sequences of [7]. The sequences contain small amounts of background clutter and out-of-plane or in-plane rotation, which is difficult for many state-of-the-art algorithms whose representation of the object is usually spatial dependent and out or in-plane rotation is not explicitly modeled.

Performance of the mean-shift algorithms, in general, drops in the presence of significant background clutter. This issue is more prominent when the tracker estimates more parameters (such as translation and scale) and the estimation errors induce a larger drift (in scale dimension) than in the case of estimating pure translation. This was mainly the case for the *drunk2 and dinosaur* sequences where the color distribution of the target was similar to the background.

Due to RGB color histogram representation, MS algorithms also perform poorly for grayscale sequences (e.g. *track running, coke, dog1, OccludedFace2, david, shaking, etc.*).

Overall, ASMS achieved the best average performance along with the TLD tracker on the sequences with scale and second best performance on the sequences without scale where the STRUCK tracker perform best. ASMS achieved the best score for 30% (which is the highest amongst other methods) of the sequences and the second best for 13%.

### 5.3. VOT2013 challenge results

The proposed ASMS algorithm was evaluated according to the new Visual Object Tracking (VOT) Challenge[3] methodology. The evaluation protocol, dataset and experiment descriptions are available at the VOT challenge site.

VOT results obtained by the ASMS method are reported in Table 3. The results show that the ASMS is quite robust: it has a low number of reinitializations (robustness column), but lacks in accuracy. Among 27 trackers the ASMS tracker would be ranked around the ninth place. This seems unimpressive, but: (i) most trackers ranked higher are significantly slower; (ii) the results depend on the choice of test sequences and the evaluation methodology; and (iii) as shown in the paper, the MS types of detectors are the best performers for certain sequences.

### 5.4. Speed

To characterize the speed, the average running time per frame of each algorithm was measured across the whole testing dataset. The forward–backward (FB) validation step has been shown to benefit the ASMS, but it comes at the price of slowing the tracking

[3] http://www.votchallenge.net/.

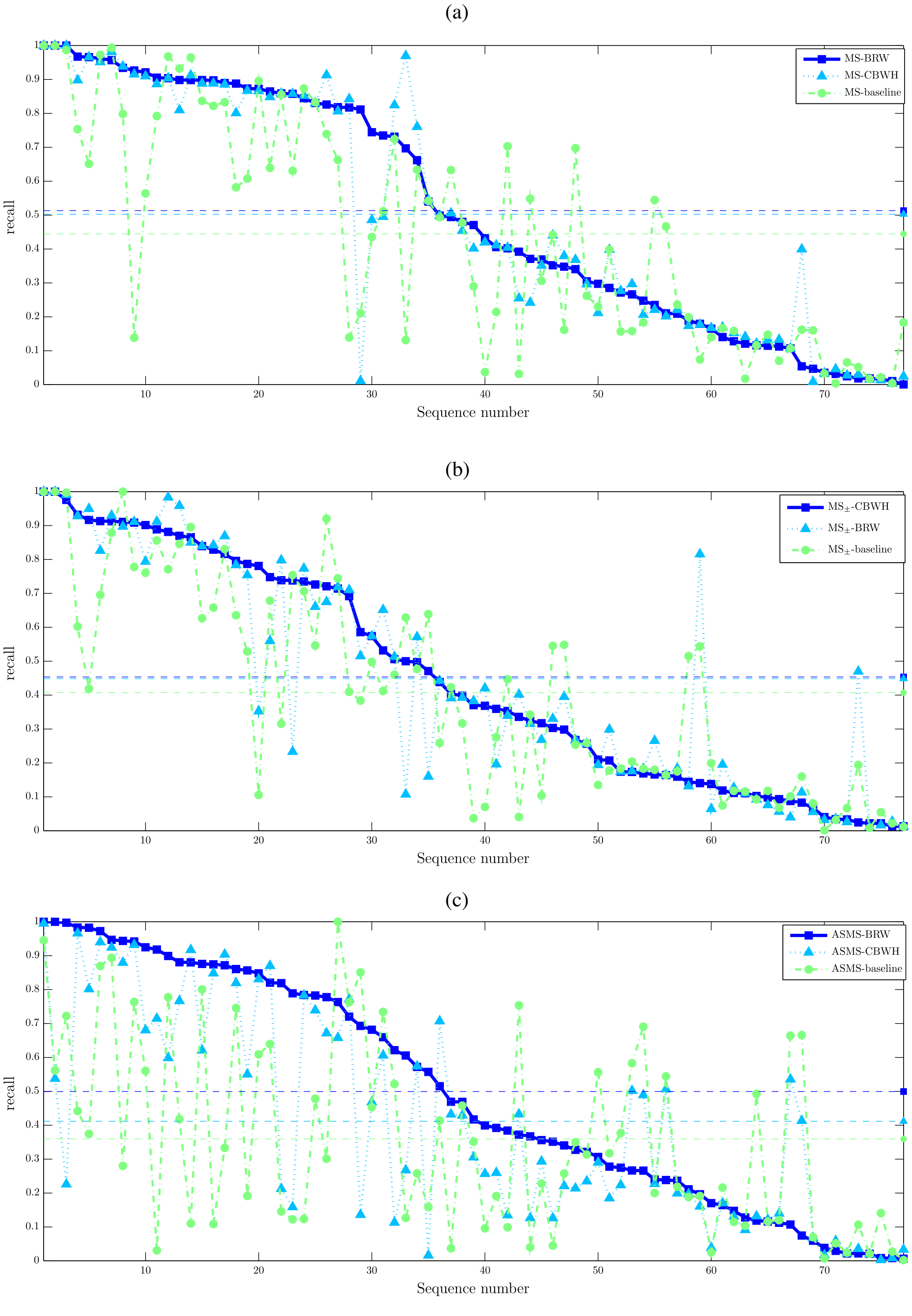


**Fig. 2.** Background weighting methods – a comparison of the standard MS, standard scale MS and adaptive scale MS. CBWH denotes the background weighting of [9]; the proposed background ratio weighting is denoted BRW. In all plots, sequences (*x*-axis) are sorted by the recall of the ASMS-BRW. The legend lists the methods in the order of average performance. The dashed lines show average performance.

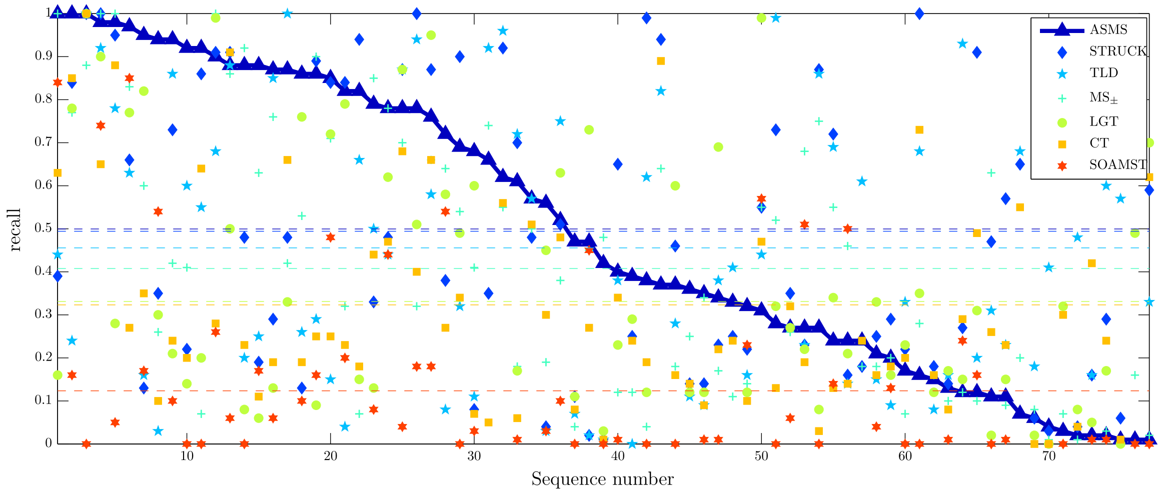


**Fig. 3.** ASMS and the state-of-the-art algorithms - comparison of the recall on 77 sequences. Sequences (*x*-axis) are sorted by the recall measure of the ASMS algorithm. The legend lists the methods in the order of average performance. The dashed lines show average performance.

**Table 1**
Recall on sequences with scale change (target was 30% smaller or larger on at least 20% of frames of the sequence). Bold text – best result for the sequence, underscore – second best. *na* indicates that the algorithm fail to process the whole sequence.

| Sequence | Method | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | MS | $MS_{\pm}$ | ASMS | SOAMST | LGT | TLD | CT | STRUCK |
| girl | 0.70 | 0.55 | 0.24 | 0.14 | 0.34 | 0.69 | 0.13 | **0.72** |
| surfer | 0.17 | 0.14 | **0.32** | 0.23 | 0.12 | 0.16 | 0.10 | 0.22 |
| Vid_A_ball | 0.86 | **1.00** | **1.00** | 0.84 | 0.16 | 0.44 | 0.63 | 0.39 |
| Vid_C_juice | 0.49 | **0.78** | **0.78** | 0.44 | 0.62 | 0.44 | 0.47 | 0.48 |
| Vid_F_person_fully_occluded | **0.40** | 0.26 | 0.27 | 0.06 | 0.27 | 0.27 | 0.32 | 0.35 |
| Vid_I_person_crossing | 0.82 | 0.76 | **0.87** | 0.06 | 0.13 | 0.85 | 0.19 | 0.29 |
| Vid_J_person_floor | **0.93** | 0.85 | 0.79 | 0.08 | 0.13 | 0.50 | 0.44 | 0.33 |
| Vid_L_coffee | 0.24 | **0.68** | 0.27 | 0.51 | 0.22 | 0.23 | 0.19 | 0.23 |
| gymnastics | 0.16 | 0.46 | 0.24 | **0.50** | 0.21 | 0.14 | 0.14 | 0.16 |
| hand | 0.64 | 0.53 | **0.86** | 0.10 | 0.76 | 0.26 | 0.19 | 0.13 |
| track_running | 0.02 | 0.11 | 0.33 | *na* | *na* | **0.41** | 0.24 | 0.25 |
| cliff-dive2 | 0.15 | 0.16 | 0.15 | *na* | 0.12 | 0.08 | 0.16 | **0.18** |
| motocross1 | 0.16 | 0.10 | 0.13 | 0.01 | **0.17** | 0.16 | 0.08 | 0.14 |
| mountain-bike | 0.80 | 0.54 | 0.69 | 0.00 | 0.49 | 0.32 | 0.34 | **0.90** |
| skiing | 0.04 | 0.04 | **0.47** | 0.00 | 0.11 | 0.07 | 0.08 | 0.11 |
| volleyball | 0.54 | 0.50 | **0.57** | *na* | *na* | **0.57** | 0.51 | 0.48 |
| CarChase | 0.07 | 0.08 | 0.06 | *na* | 0.02 | **0.18** | 0.00 | 0.06 |
| Motocross | 0.00 | 0.01 | 0.04 | *na* | 0.00 | **0.41** | 0.00 | 0.03 |
| Panda | 0.07 | 0.07 | 0.17 | 0.00 | 0.23 | **0.33** | 0.20 | 0.22 |
| Volkswagen | 0.00 | 0.00 | 0.01 | *na* | 0.00 | **0.57** | 0.01 | 0.06 |
| pedestrian3 | 0.11 | **0.63** | 0.11 | 0.00 | 0.02 | 0.31 | 0.26 | 0.47 |
| jump | 0.31 | 0.34 | **0.35** | 0.01 | 0.12 | 0.09 | 0.09 | 0.14 |
| animal | 0.63 | 0.18 | 0.61 | 0.01 | 0.17 | **0.72** | 0.06 | 0.70 |
| singer1 | 0.12 | 0.12 | 0.12 | 0.24 | 0.15 | **0.93** | 0.29 | 0.27 |
| singer1(lowfps) | 0.26 | 0.25 | **0.36** | *na* | 0.12 | 0.11 | 0.14 | 0.14 |
| skating2 | 0.83 | 0.26 | **0.94** | 0.54 | 0.30 | 0.03 | 0.10 | 0.35 |
| soccer | 0.20 | 0.20 | 0.20 | 0.13 | 0.16 | 0.09 | 0.18 | **0.29** |
| drunk2 | 0.03 | 0.03 | 0.02 | 0.01 | 0.17 | **0.60** | 0.24 | 0.29 |
| lemming | 0.83 | 0.83 | **0.97** | 0.85 | 0.77 | 0.63 | 0.27 | 0.66 |
| dog1 | 0.18 | 0.20 | 0.07 | *na* | *na* | **0.68** | 0.55 | 0.65 |
| trellis | 0.16 | 0.18 | 0.37 | 0.00 | **0.60** | 0.28 | 0.16 | 0.46 |
| coke | 0.05 | 0.07 | 0.03 | 0.00 | 0.32 | **0.92** | 0.30 | 0.88 |
| Mean | 0.34 | 0.34 | 0.39 | 0.20 | 0.24 | 0.39 | 0.22 | 0.34 |
| Best + Second (out of 32) | 2 + 8 | 4 + 6 | 11 + 3 | 1 + 3 | 2 + 4 | 11 + 2 | 0 + 2 | 4 + 9 |

**Table 2**
Recall on sequences without a scale change. Bold text – best result for the sequence, underscore – second best. *na* indicates that the algorithm fail to process the whole sequence.

| Sequence | Method | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | MS | MS$_\pm$ | ASMS | SOAMST | LGT | TLD | CT | STRUCK |
| OccludedFace2 | 0.44 | 0.28 | 0.16 | 0.00 | 0.35 | 0.68 | 0.73 | **1.00** |
| Vid_B_cup | **1.00** | **1.00** | 0.98 | 0.74 | 0.90 | 0.92 | 0.65 | **1.00** |
| Vid_D_person | 0.97 | **1.00** | 0.98 | 0.05 | 0.28 | 0.78 | 0.88 | 0.95 |
| Vid_E_person_part_occluded | 0.90 | 0.86 | 0.88 | 0.06 | 0.50 | 0.88 | **0.91** | **0.91** |
| Vid_G_rubikscube | 0.63 | 0.77 | **1.00** | 0.16 | 0.78 | 0.24 | 0.85 | 0.84 |
| Vid_H_panda | **1.00** | 0.88 | **1.00** | 0.00 | **1.00** | **1.00** | **1.00** | **1.00** |
| Vid_K_cup | 0.87 | **1.00** | 0.90 | 0.26 | 0.99 | 0.68 | 0.28 | 0.91 |
| dinosaur | 0.23 | 0.17 | 0.34 | 0.01 | **0.69** | 0.38 | 0.22 | 0.23 |
| hand2 | 0.63 | 0.41 | **0.68** | 0.03 | 0.60 | 0.11 | 0.07 | 0.08 |
| torus | 0.65 | 0.60 | **0.95** | 0.17 | 0.82 | 0.16 | 0.35 | 0.13 |
| head_motion | 0.70 | 0.64 | 0.37 | *na* | *na* | 0.82 | 0.89 | **0.94** |
| shaking_camera | 0.44 | 0.74 | 0.66 | *na* | *na* | **0.92** | 0.05 | 0.35 |
| cliff-dive1 | **0.99** | 0.66 | 0.76 | 0.18 | 0.95 | 0.58 | 0.66 | 0.87 |
| motocross2 | 0.74 | 0.70 | 0.78 | 0.04 | **0.87** | **0.87** | 0.68 | **0.87** |
| car | 0.54 | 0.52 | 0.28 | 0.00 | 0.32 | **0.99** | 0.13 | 0.73 |
| david | 0.02 | 0.01 | 0.02 | *na* | 0.08 | **0.48** | 0.04 | 0.04 |
| jumping | 0.51 | 0.75 | 0.27 | 0.00 | 0.08 | 0.86 | 0.03 | **0.87** |
| pedestrian4 | 0.56 | 0.41 | **0.92** | 0.00 | 0.14 | 0.60 | 0.20 | 0.22 |
| pedestrian5 | 0.97 | 0.42 | 0.87 | *na* | 0.33 | **1.00** | 0.66 | 0.48 |
| diving | 0.18 | 0.18 | 0.21 | 0.04 | **0.33** | 0.15 | 0.16 | 0.25 |
| gym | **0.96** | 0.90 | 0.86 | 0.16 | 0.09 | 0.29 | 0.25 | 0.89 |
| trans | 0.55 | 0.55 | 0.31 | 0.57 | **0.99** | 0.44 | 0.47 | 0.55 |
| basketball | 0.48 | 0.45 | 0.47 | 0.45 | **0.73** | 0.02 | 0.27 | 0.02 |
| football | 0.16 | 0.16 | 0.01 | 0.00 | 0.49 | 0.76 | 0.73 | **0.86** |
| shaking | 0.07 | 0.05 | 0.02 | 0.01 | 0.05 | 0.16 | **0.42** | 0.16 |
| singer2 | 0.21 | 0.19 | **0.56** | 0.03 | 0.45 | 0.03 | 0.30 | 0.04 |
| skating1 | 0.16 | 0.12 | 0.40 | 0.01 | 0.23 | 0.38 | 0.34 | **0.65** |
| skating1(lowfps) | 0.14 | 0.09 | 0.11 | 0.01 | 0.15 | 0.23 | 0.23 | **0.57** |
| Asada | 0.66 | 0.64 | **0.72** | 0.54 | 0.58 | 0.08 | 0.27 | 0.38 |
| dudek-face | 0.47 | 0.18 | 0.24 | *na* | *na* | **0.61** | 0.24 | 0.18 |
| faceocc1 | 0.79 | 0.32 | 0.78 | 0.18 | 0.51 | 0.94 | 0.40 | **1.00** |
| figure_skating | 0.61 | 0.32 | 0.82 | 0.20 | 0.79 | 0.04 | 0.23 | **0.84** |
| woman | 0.14 | 0.07 | 0.82 | *na* | 0.15 | 0.66 | 0.18 | **0.94** |
| board | 0.84 | 0.71 | **0.85** | 0.48 | 0.72 | 0.15 | 0.25 | 0.84 |
| box | 0.16 | 0.10 | 0.12 | 0.16 | 0.31 | 0.20 | 0.49 | **0.91** |
| liquor | 0.58 | 0.42 | **0.94** | 0.10 | 0.21 | 0.86 | 0.24 | 0.73 |
| Sylvestr | 0.72 | 0.55 | 0.62 | *na* | *na* | **0.96** | 0.56 | 0.92 |
| car11 | 0.29 | 0.04 | 0.38 | 0.00 | 0.12 | 0.62 | 0.19 | **0.99** |
| person | **0.99** | 0.92 | 0.88 | 0.00 | 0.08 | 0.20 | 0.23 | 0.48 |
| tiger1 | 0.14 | 0.07 | **0.92** | 0.00 | 0.20 | 0.55 | 0.64 | 0.86 |
| tiger2 | 0.02 | 0.02 | 0.01 | 0.00 | **0.70** | 0.33 | 0.62 | 0.59 |
| bird_1 | 0.03 | 0.12 | **0.39** | *na* | 0.29 | 0.00 | 0.24 | 0.25 |
| bird_2 | 0.13 | 0.38 | 0.52 | 0.10 | 0.63 | **0.75** | 0.48 | 0.51 |
| bolt | 0.21 | **0.48** | 0.42 | 0.00 | 0.03 | 0.01 | 0.01 | 0.01 |
| girl_mov | 0.75 | 0.63 | **0.88** | 0.17 | 0.06 | 0.25 | 0.11 | 0.19 |
| Mean | 0.52 | 0.46 | 0.58 | 0.13 | 0.45 | 0.50 | 0.40 | 0.60 |
| Best + Second (out of 45) | 5 + 8 | 4 + 4 | 12 + 7 | 0 + 1 | 7 + 7 | 9 + 9 | 3 + 6 | 15 + 6 |

**Table 3**
VOT2013 Challenge results. Accuracy is abbreviated as *acc* and robustness as *rob*.

| | Experiment baseline | | | Experiment region_noise | | | Experiment grayscale | | |
|---|---|---|---|---|---|---|---|---|---|
| | acc | rob | speed (fps) | acc | rob | speed (fps) | acc | rob | speed (fps) |
| bicycle | 0.51 | 0.00 | 168.30 | 0.50 | 0.07 | 169.24 | 0.53 | 8.00 | 174.82 |
| bolt | 0.65 | 1.00 | 96.90 | 0.65 | 1.00 | 101.73 | 0.42 | 5.00 | 89.38 |
| car | 0.43 | 0.00 | 234.50 | 0.49 | 0.60 | 256.80 | 0.44 | 0.00 | 301.77 |
| cup | 0.72 | 0.00 | 224.59 | 0.70 | 0.00 | 268.81 | 0.57 | 1.00 | 203.13 |
| david | 0.52 | 2.00 | 113.42 | 0.51 | 1.80 | 99.51 | 0.44 | 8.00 | 90.00 |
| diving | 0.37 | 0.00 | 200.06 | 0.38 | 0.00 | 232.96 | 0.38 | 3.00 | 130.10 |
| face | 0.67 | 0.00 | 190.56 | 0.66 | 0.00 | 196.25 | 0.63 | 0.00 | 166.11 |
| gymnastics | 0.47 | 0.00 | 172.89 | 0.45 | 0.00 | 187.58 | 0.41 | 2.00 | 121.01 |
| hand | 0.64 | 0.00 | 159.68 | 0.63 | 0.13 | 193.37 | 0.52 | 1.00 | 169.22 |
| iceskater | 0.61 | 0.00 | 99.02 | 0.60 | 0.00 | 105.35 | 0.49 | 5.00 | 53.66 |
| juice | 0.65 | 0.00 | 218.33 | 0.66 | 0.00 | 237.51 | 0.71 | 1.00 | 201.20 |
| jump | 0.53 | 1.00 | 124.98 | 0.53 | 0.67 | 141.74 | 0.44 | 1.00 | 143.42 |
| singer | 0.39 | 0.00 | 56.45 | 0.38 | 0.07 | 57.39 | 0.42 | 5.00 | 67.09 |
| sunshade | 0.61 | 0.00 | 199.40 | 0.62 | 0.00 | 218.85 | 0.55 | 2.00 | 154.39 |
| torus | 0.70 | 0.00 | 168.12 | 0.69 | 0.00 | 213.55 | 0.55 | 1.00 | 172.86 |
| woman | 0.63 | 2.00 | 168.92 | 0.61 | 2.00 | 166.40 | 0.57 | 7.00 | 138.73 |

**Table 4**
Processing speed in milliseconds. Max (min) are computed as a maximum (minimum) of the average time per sequence; mean is the average time over all sequences.

| Method | MS | MS$_\pm$ | ASMS | SOAMST | LGT | TLD | CT | STRUCK |
|---|---|---|---|---|---|---|---|---|
| max | 14.4 | 61 | 48 | 6107 | 864 | 152 | 36 | 112 |
| min | 0.4 | 0.8 | 0.6 | 207 | 107 | 6 | 11 | 43 |
| mean | 2.9 | 7.3 | 6.1 | 816 | 250 | 51 | 21 | 82 |

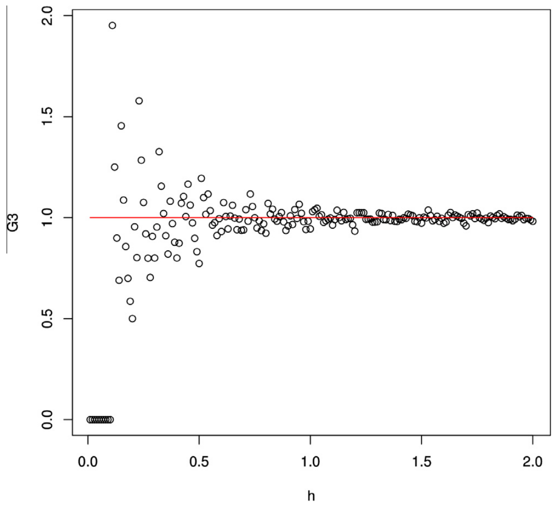


**Fig. A.4.** Behaviour of the *A* term for a target represented by an ellipsoidal region with Epanechnikov kernel and $a = 10$ and $b = 10$ for a variable scale parameter *h*.

two times. The experiment shown (see Table 4) that the slow down factor w.r.t. to standard MS is 2 on average. However, ASMS is still faster then MS$_\pm$ and significantly faster than the state-of-the-art algorithms.

## 6. Conclusion

In this work, a theoretically justified scale estimation for the mean-shift algorithm using Hellinger distance has been proposed. The new scale estimation procedure is regularized, which makes it more robust. Furthermore, we proposed a new formulation of the histogram bin weighting function (BRW) that takes into account background appearance. The formulation is general and can be used in any MS-based algorithm. The increase in performance when using BRW is shown in Fig. 2.

We introduced a scheme (Forward–Backward) for automatic decision to accept the newly estimated scale or to use a more robust weighted combination, which is shown to reduce erroneous scale updates. This technique reduces tracking speed twice, however ASMS is still faster then MS$_\pm$ and outperforms the speed of the state-of-the-art method by a large margin (see Table 4).

The newly proposed ASMS has been compared with the state-of-the-art algorithms on a very large dataset of tracking sequences. It outperforms the reference algorithms in average recall, processing speed and it achieves the best score for 30% of the sequences (the highest percentage among the reference algorithms) and it is the second best performer for 13% of the sequences.

## Acknowledgments

The research was supported by the Czech Science Foundation Project GACR P103/12/G084.

## Appendix A

Let us assume we do not use an approximation for $C_h$. Thus to minimize the Hellinger distance

$$\rho[\hat{\mathbf{p}}(\mathbf{y},h),\hat{\mathbf{q}}] = \sum_{u=1}^{m}\sqrt{C_h\sum_{i=1}^{N}k\left(\frac{(y^1-x_i^1)^2}{a^2h^2}+\frac{(y^2-x_i^2)^2}{b^2h^2}\right)\delta[b(\mathbf{x}_i)-u]\hat{q}_u} \tag{A.1}$$

is maximized using a gradient method. The only difference from the derivation using the approximation (Eq. (13)) is in the partial derivative w.r.t. *h*:

$$\begin{aligned}\frac{\partial\rho(\mathbf{y},h)}{\partial h}(\hat{\mathbf{y}}_0,h_0) = \frac{C_{h_0}}{(h_0)^2}\Bigg[\frac{1}{h_0}\sum_{i=1}^{N}w_i\cdot\left(\frac{(\hat{y}_0^1-x_i^1)^2}{a^2}+\frac{(\hat{y}_0^2-x_i^2)^2}{b^2}\right)\\ \cdot g\left(\frac{(\hat{y}_0^1-x_i^1)^2}{a^2h_0^2}+\frac{(\hat{y}_0^2-x_i^2)^2}{b^2h_0^2}\right)\\ -h_0\sum_{i=1}^{N}w_i\cdot k\left(\frac{(\hat{y}_0^1-x_i^1)^2}{a^2h_0^2}+\frac{(\hat{y}_0^2-x_i^2)^2}{b^2h_0^2}\right)\cdot A\Bigg],\end{aligned} \tag{A.2}$$

where

$$A = \frac{\sum_{i=1}^{N}\left(\frac{(\hat{y}_0^1-x_i^1)^2}{a^2}+\frac{(\hat{y}_0^2-x_i^2)^2}{b^2}\right)\cdot g\left(\frac{(\hat{y}_0^1-x_i^1)^2}{a^2h_0^2}+\frac{(\hat{y}_0^2-x_i^2)^2}{b^2h_0^2}\right)}{\sum_{i=1}^{N}k\left(\frac{(\hat{y}_0^1-x_i^1)^2}{a^2h_0^2}+\frac{(\hat{y}_0^2-x_i^2)^2}{b^2h_0^2}\right)}, \tag{A.3}$$

and *A* tends to 1 for large numbers of pixels in a target candidate. The proposed approximation therefore replaces *A* by 1 and eliminates the noise caused by *A* term for small scales of the objects. It is illustrated in Fig. A.4 for a target represented by an ellipsoidal region with $a = 10$ and $b = 10$ (i.e. object size equal to 20*x*20 px).

## References

[1] G.R. Bradski, Computer vision face tracking for use in a perceptual user interface, Intel Technol. J. 2 (1998) 15–26.
[2] R.T. Collins, Mean-shift blob tracking through scale space, in: Computer Vision and Pattern Recognition, IEEE Computer Society, 2003, pp. 234–240.
[3] D. Comaniciu, V. Ramesh, P. Meer, Real-time tracking of non-rigid objects using mean shift, in: Computer Vision and Pattern Recognition, Proceedings, IEEE Conference on, vol. 2, 2000, pp. 142–149.
[4] K. Fukunaga, L. Hostetler, The estimation of the gradient of a density function, with applications in pattern recognition, Inf. Theory 21 (1975) 32–40.
[5] S. Hare, A. Saffari, P. Torr, Struck: structured output tracking with kernels, in: International Conference Computer Vision, 2011, pp. 263–270.
[6] Z. Kalal, J. Matas, K. Mikolajczyk, P-N learning: bootstrapping binary classifiers by structural constraints, in: Conference on Computer Vision and Pattern Recognition, 2010.
[7] D.A. Klein, D. Schulz, S. Frintrop, A.B. Cremers, Adaptive real-time video-tracking for arbitrary objects, in: Intelligent Robots and Systems, 2010, pp. 772–777.
[8] D. Liang, Q. Huang, S. Jiang, H. Yao, W. Gao, Mean-shift blob tracking with adaptive feature selection and scale adaptation, in: International Conference Image Processing, 2007.
[9] J. Ning, L. Zhang, D. Zhang, C. Wu, Robust mean-shift tracking with corrected background-weighted histogram, IET Comput. Vision 6 (2012) 62–69.
[10] J. Ning, L. Zhang, D. Zhang, C. Wu, Scale and orientation adaptive mean shift tracking, IET Comput. Vision 6 (2012) 52–61.
[11] J.X. Pu, N.S. Peng, Adaptive kernel based tracking using mean-shift, in: International conference on Image Analysis and Recognition, 2006, pp. 394–403.
[12] L. Čehovin, M. Kristan, A. Leonardis, An adaptive coupled-layer visual model for robust visual tracking, in: 13th International Conference on Computer Vision, 2011.
[13] C. Yang, R. Duraiswami, L. Davis, Efficient mean-shift tracking via a new similarity measure, in: Computer Vision and Pattern Recognition, vol. 1, 2005, pp. 176–183.
[14] K. Zhang, L. Zhang, M.H. Yang, Real-time compressive tracking, in: European conference on Computer Vision, 2012, pp. 864–877.
[15] C. Zhao, A. Knight, I. Reid, Target tracking using mean-shift and affine structure, in: ICPR, 2008, pp. 1–5.

# C Online adaptive hidden Markov model for multi-tracker fusion

Vojir T., Matas J., Noskova, J., *Online adaptive hidden Markov model for multi-tracker fusion*, Computer Vision and Image Understanding 153, pp. 109-119, 2016. DOI: 10.1016/j.cviu.2016.05.007

Computer Vision and Image Understanding 153 (2016) 109–119

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

# Online adaptive hidden Markov model for multi-tracker fusion

Tomas Vojir[a,*], Jiri Matas[a], Jana Noskova[b]

[a] The Center for Machine Perception, FEE CTU in Prague, Karlovo namesti 13, 121 35 Prague 2, Czech Republic
[b] Faculty of Civil Engineering, CTU in Prague, Thakurova 7/2077, 166 29 Prague 6, Czech Republic

ARTICLE INFO

*Article history:*
Received 27 July 2015
Revised 21 January 2016
Accepted 15 May 2016
Available online 16 May 2016

*Keywords:*
Visual tracking
On-line learning
Hidden markov model
Object detection

ABSTRACT

In this paper, we propose a novel method for visual object tracking called HMMTxD. The method fuses observations from complementary out-of-the box trackers and a detector by utilizing a hidden Markov model whose latent states correspond to a binary vector expressing the failure of individual trackers. The Markov model is trained in an unsupervised way, relying on an online learned detector to provide a source of tracker-independent information for a modified Baum- Welch algorithm that updates the model w.r.t. the partially annotated data. We show the effectiveness of the proposed method on combination of two and three tracking algorithms. The performance of HMMTxD is evaluated on two standard benchmarks (CVPR2013 and VOT) and on a rich collection of 77 publicly available sequences. The HMMTxD outperforms the state-of-the-art, often significantly, on all data-sets in almost all criteria.

## 1. Introduction

In the last thirty years, a large number of diverse visual tracking methods has been proposed (Smeulder et al., 2013; Yilmaz et al., 2006). The methods differ in the formulation of the problem, assumptions made about the observed motion, in optimization techniques, the features used, in the processing speed, and in the application domain. Some methods focus on specific challenges like tracking of articulated or deformable objects (Cehovin et al., 2013; Godec et al., 2011; Kwon and Lee, 2009), occlusion handling (Grabner et al., 2010), abrupt motion (Zhou and Lu, 2010) or long-term tracking (Kalal et al., 2012; Pernici and Bimbo, 2013).

Three observations motivate the presented research. First, most trackers perform poorly if run outside the scenario they were designed for. Second, some trackers make different and complementary assumptions and their failures are not highly correlated (called complementary trackers in the paper). And finally, even fairly complex well performing trackers run at frame rate or faster on standard hardware, opening the possibility for multiple trackers to run concurrently and yet in or near real-time.

We propose a novel methodology that exploits a hidden Markov model (HMM) for fusion of non-uniform observables and pose prediction of multiple complementary trackers using an on-line learned high-precision detector. The non-uniform observables, in this sense, means that each tracker can produce its own type of "confidence estimate" which may not be directly comparable between each other.

The HMM, trained in an unsupervised manner, estimates the state of the trackers – failed, operates correctly – and outputs the pose of the tracked object taking into account the past performance and observations of the trackers and the detector. The HMM treats the detector output as correct if it is not in contradiction with its current most probable state in which the majority of trackers are correct. This limits the cases where the HMM would be wrongly updated by a false detection. For the potentially many frames where reliable detector output is not available, it combines the trackers. The detector is trained on the first image and interacts with the learning of the HMM by partially annotating the sequence of HMM states in the time of verified detections. The recall of the detector is not critical but it affects the learning rate of the HMM and the long-term properties of the HMMTxD method, i.e., its ability to re-initialize trackers after occlusions or object disappearance.

**Related work.** The most closely related approaches include Santner et al. (2010), where three tracking methods with different rates of appearance adaptation are combined to prevent drift due to incorrect model updates. The approach uses simple, hard-coded rules for tracker selection. Kalal et al. (2012) combine a tracking-by-detection method with a short-term tracker that generates so called P-N events to learn new object appearance. The output is defined either by the detector or the tracker based on visual similarity to the learned object model. Both these methods employ pre-defined rules to make decisions about

[*] Corresponding author.
*E-mail addresses:* vojirtom@cmp.felk.cvut.cz (T. Vojir), matas@cmp.felk.cvut.cz (J. Matas), noskova@fsv.cvut.cz (J. Noskova).
*URL:* http://cmp.felk.cvut.cz/~vojirtom/ (T. Vojir), http://cmp.felk.cvut.cz/~matas/ (J. Matas)

http://dx.doi.org/10.1016/j.cviu.2016.05.007
1077-3142/© 2016 Elsevier Inc. All rights reserved.

object pose and use one type of measurement, a certain form of similarity between the object and the estimated location. In contrary, HMMTxD learns continuously and causally the performance statistics of individual parts of the systems and fuses multiple "confidence" measurements in the form of probability densities of observables in the HMM. Zhang et al. (2014) use a pool of multiple classifiers learned from different time spans and choose the one that maximize an entropy-based cost function. This method addresses the problem of model drifting due to wrong model updates, but the failure modes inherent to the classifier itself remains the same. This is unlike the proposed method which allows to combine diverse tracking methods with different inherent failure modes and with different learning strategies to balance their weaknesses.

Similarly to the proposed method, Wang and yan Yeung (2014) and Bailer et al. (2014) fuse different out-of-the box tracking methods. Bailer et al. combine offline the *outputs* of multiple tracking algorithms. There is no interaction between trackers, which for instance implies that the method avoids failure only if one method correctly tracks the whole sequence. Wang et al., use a factorial hidden Markov model and a Bayesian approach. The state space of their factorial HMM is the set of potential object positions, therefore it is very large. The model contains a probability description of the object motion based on a particle filter. Trackers interact by reinitializing those with low reliability to the pose of the most confident one. The Yuan et al. (2015) use HMM in the same setup, but rather than merging multiple tracking method, they focus on modeling the temporal change of the target appearance in the HMM framework by introducing a observational dependencies. In contrast, the HMMTxD method is online with tracker interaction via a high precision object detector that supervises tracker re-initializations which happen on the fly. The appearance modeling is performed inside of each tracker and the HMMTxD capture the relation of the confidence provided by tracker and its performance, validated by the object detector, by the observable distributions. Moreover, the HMMTxD confidence estimation is motion-model free and this prevents biases towards support of trackers with a particular motion model.

Yoon et al. (2012) combines multiple trackers in a particle filter framework. This approach models observables and transition behavior of individual trackers, but the trackers are self-adapting which makes it prone to wrong model updates. The adaptation of HMMTxD model is supervised by a detector method set to a specific mode of operation – near 100% precision – alleviating the incorrect update problem.

The contributions of the paper are: a novel method for fusion of multiple trackers based on HMMs using non-uniform observables, a simple, and so far unused, unsupervised method for HMMs training in the context of tracking, tunable feature-based detector with very low false positive rate, and the creation of a tracking system that shows state-of-the-art performance.

## 2. Fusing multiple trackers

HMMTxD uses a hidden Markov model (HMM) to integrate pose and observational confidence of different trackers and a detector, and updates its own confidence estimates that in turn define the pose that it outputs. In the HMM, each tracker is modeled as working correctly (1) or incorrectly (0). The HMM poses no constraints on the definition of tracker correctness, we adopted target overlap above a threshold. Having at our disposal **n** trackers, the set of all possible states is $\{s_1, s_2, \ldots, s_N\} = \{0, 1\}^{\mathbf{n}}, N = 2^{\mathbf{n}}$ and the initial state $s_1 = (1, 1, \ldots, 1)$. Note that the trackers are not assumed to be independent, because an independence of tracker correctness is not a realistic assumption. For example, if the tracking problem is relatively easy, all trackers tend to be correct and in the case of occlusion all tend to be incorrect (see the analysis in Kristan et al., 2015). The number of states $2^{\mathbf{n}}$ grows exponentially with the number of trackers. However, we do not consider this a significant issue – due to "real-time" requirements of tracking, the need to combine more than a small number of trackers, say $\mathbf{n} = 4$, is unlikely.

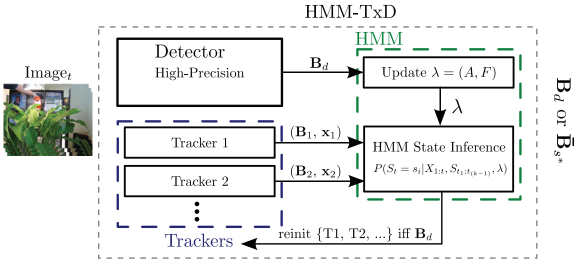

HMM-TxD
Image$_t$
Detector
High-Precision
$\mathbf{B}_d$
HMM
Update $\lambda = (A, F)$
$\lambda$
Tracker 1
Tracker 2
$(\mathbf{B}_1, \mathbf{x}_1)$
$(\mathbf{B}_2, \mathbf{x}_2)$
HMM State Inference
$P(S_t = s_i | X_{1:t}, S_{t_1:t_{(k-1)}}, \lambda)$
reinit {T1, T2, ...} iff $\mathbf{B}_d$
Trackers
$\mathbf{B}_d$ or $\bar{\mathbf{B}}_{s^*}$

**Fig. 1.** The structure of the HMMTxD. For each frame, the detector and trackers are run. Each tracker outputs a new object pose and observables ($\mathbf{B}_i$, $\mathbf{x}_i$) and the detector outputs either the verified object pose $\mathbf{B}_d$ or nothing. If detector fires, HMM is updated and trackers are re-initialized and the final output is the $\mathbf{B}_d$, otherwise, HMM estimate the most probable state $s^*$ and outputs an average bounding box $\bar{\mathbf{B}}_{s^*}$ of trackers that are correct in the estimated state $s^*$.

The HMMTxD method overview is illustrated in Fig. 1. Each tracker provides an estimate of the object pose ($\mathbf{B}_i$) and a vector of observables ($\mathbf{x}_i$), which may contain a similarity measure to some model (such as normalized cross-correlation to the initial image patch, distance of template and current histograms at given position, etc.) or any other estimates of the tracker performance. The $\mathbf{x}_i, i = \{1, 2, \ldots, \mathbf{n}\}$ serve as observables to relate the tracker current confidence to the HMM. Each individual observable depends only on one particular tracker and its correctness, hence, they are assumed to be conditionally independent conditioned on the state of the HMM (which encodes the tracker correctness).

In general, there are no constraints on observable values, however, in the proposed HMM the observable values are required to be normalized to the (0, 1) interval. The observables are modeled as beta-distributed random variables (Eq. 1) and its parameters are estimated online. The beta distribution was chosen for its versatility, where practically any kind of unimodal random variable on (0, 1) can be modeled by the beta distribution, i.e., for any choice of any lower and upper quantiles, a beta distribution exists satisfying the given quantile constraint (Gupta and Nadarajah, 2004).

Learning the parameters of the beta distributions online is crucial for the adaptability to particular tracking scenes, where the observable values from a different trackers may be biased due to scene properties, or to adapt to a different types of observables of trackers and their correlations to the "real" tracker performance. For example, taking correlation with the initial target patch as an observable for one tracker and color histogram distance to a initial target for a second tracker, the correlation between their values and the performance of the tracker may differ depending on object rigidity and color distribution of object and background.

The HMM is parameterized by the pair $\lambda = (A, F)$, where $A$ are the probabilities of state transition and $F$ are the beta distributions of observables with shape parameters $p$, $q > 0$ and density defined for $x \in (0, 1)$

$$f(x|p, q) = \frac{x^{p-1}(1-x)^{q-1}}{\int_0^1 u^{p-1}(1-u)^{q-1} du}. \tag{1}$$

Since the goal is real-time tracking without any specific preprocessing, learning of HMM parameters has to be done online. Towards this goal, the object detector, which is set to operating mode with low false positive rate, is utilized to partially annotate the sequence of hidden states. In contrast to classical HMM, where only a

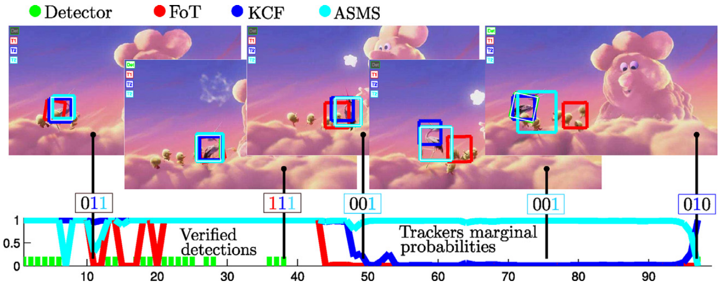


**Fig. 2.** Illustration of HMM state and trackers probability estimation during tracking. The bottom graph shows the marginal probabilities for each tracker being correct and the detection times (green spikes). Above the graph the inferred states $s_t^*$ with color encoded correct trackers (1) are displayed. The final output is defined by the state $s_t^*$ and the bounding box is highlighted by white color. Best viewed zoomed in color.

sequence of observations $\mathbb{X} = \{X_t\}_{t=1}^T, X_t = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x_n})$ is available, we are in a semi-supervised setting and have a time sequence $0 = t_0 < t_1 < t_2 \ldots < t_K \le T$ of observed states of a Markov chain $\mathbb{S} = \{S_{t_k} = s_{i_k}, \{t_k\}_{k=1}^K\}$, and Markov chain starting again in state $s_1$, all trackers correct, at any time $\{t_k + 1, 0 \le k \le K\}$, since there are re-initialized to common object pose. This information is provided by the detector, where $\{t_k\}_{k=1}^K$ is a sequence of detection times. The HMM parameters are learned by a modified Baum-Welch algorithm run on the observations $\mathbb{X}$ and the annotated sequence of states $\mathbb{S}$. The partial annotation and HMM parameter estimation update is done strictly online.

The output of the HMMTxD is an average bounding box of correct trackers of the current most probable state $s_t^*$. For $t_{(k-1)} < t < t_k, 1 \le k \le K$ the forward-backward procedure (Rabiner, 1989) for HMM is used to calculate probability of each state at time $t$ (see Eqs. A.1–A.7) and the state $s_t^* \in \{0, 1\}^{\mathbf{n}} \setminus (0, 0, \ldots, 0)$ is the state for which

$$P(S_t = s_i | X_1, \ldots, X_t, S_{t_1}, \ldots, S_{t_{(k-1)}}, \lambda) \tag{2}$$

is maximal. This equation is computed using Eq. A.5 and maximized w.r.t $i$, $1 \le i \le N$. For $t_K < t \le T$ the Eq. 2 holds with $t_{(k-1)} = t_K$. This ensures that the algorithm outputs a pose for each frame which is required by most benchmark protocols. Illustration of the tracking process and HMM insight is shown in Fig. 2. Theoretically the parameters of HMM could be updated after each frame. However, in our implementation, learning takes place only at frames where the detector positively detects the object, i.e., the sequence of states starting and ending with observed state inferred by the detector.[1] The detector is used only if the detection pose is not in contradiction with the pose of the current most probable state in which the majority of trackers are correct. This ensure that even when the detector makes a mistake, the HMM is not wrongly updated. When we are in the state that one or none of the trackers are correct, the detector get precedence.

## 3. Learning the hidden Markov model

For learning of the parameters $\lambda$ of the HMM a MLE inference is employed, however maximizing the likelihood function $P(\mathbb{X}, \mathbb{S}|\lambda)$ is a complicated task that cannot be solved analytically. In the proposed method, the Baum-Welch algorithm (Baum et al., 1970) is adapted. The Baum-Welch algorithm is a widespread iterative procedure for estimating parameters of HMM where each iteration increases the likelihood function but, in general, the convergence to the global maximum is not guaranteed. The Baum-Welch algorithm is in fact an application of the EM (Expectation-Maximization) algorithm (Dempster et al., 1977).

### 3.1. Classical Baum-Welch algorithm

Let us assume the HMM with $N$ possible states $\{s_1, s_2, \ldots, s_N\}$, the matrix of state transition probabilities $A = \{a_{ij}\}_{i,j=1}^N$, the vector of initial state probabilities $\pi = (1, 0, 0, \ldots, 0)$, the initial state $s_1 = (1, 1, \ldots, 1)$, a sequence of observations $\mathbb{X} = \{X_t\}_{t=1}^T, X_t \in R^m$ and $F = \{f_i(x)\}_{i=1}^N$ the system of conditional probability densities of observations conditioned on $S_t = s_i$

$$f_i(x) = f(x|S_t = s_i) \text{ for } 1 \le i \le N, 1 \le t \le T, x \in R^m \tag{3}$$

where $S_t$ are random variables representing the state at time $t$, and $\lambda = (A, F)$ is denoting the parameter set of the model.

Let us denote

$$Q(\lambda, \lambda') = \sum_{\mathfrak{s} \in \mathfrak{S}} P(\mathfrak{s}|\mathbb{X}, \lambda) \log[P(\mathfrak{s}, \mathbb{X}|\lambda')], \tag{4}$$

where $\mathfrak{S} = \{s_1, s_2, \ldots, s_N\}^T$ is a set of all possible T-tuples of states and
$\mathfrak{s} \in \mathfrak{S}, \mathfrak{s} = (\mathfrak{s}_1, \ldots, \mathfrak{s}_t, \ldots, \mathfrak{s}_T)$ is one sequence of states. According to Theorem 2.1. in Baum et al. (1970)

$$Q(\lambda, \lambda') \ge Q(\lambda, \lambda) \Rightarrow P(\mathbb{X}|\lambda') \ge P(\mathbb{X}|\lambda) \tag{5}$$

and the equality holds if and only if $P(\mathfrak{s}|\mathbb{X}, \lambda) = P(\mathfrak{s}|\mathbb{X}, \lambda')$ for $\forall \mathfrak{s} \in \mathfrak{S}$. The classical Baum-Welch algorithm repeats the following steps until convergence:

1. Compute $\lambda^* = \arg\max_\lambda Q(\lambda_n, \lambda)$
2. Set $\lambda_{n+1} = \lambda^*$.

### 3.2. Modified Baum-Welch algorithm

We propose the modified Baum-Welch algorithm that exploits the partially annotated sequence of states, where the known states are inferred from the detector output. Let $0 = t_0 < t_1 < t_2 \ldots < t_K \le T$ be a sequence of detection times, $\mathbb{S} = \{S_{t_k} = s_{i_k}, \{t_k\}_{k=1}^K\}$ be observed states of Markov chain, marked by the detector, and $S_{t_k+1} = s_1$ for $0 \le k \le K$. So the sequence of observations of the HMM is divided into $K + 1$ independent subsequences, each with a fixed initial state $s_1$, the first $K$ subsequences with a known terminal state defined by the detector and the last subsequence with an unknown terminal state.

[1] If pure online fusion is not required, future observations can also be used to determine the probability of each state.

The following equations are obtained by employing the modification to the Baum-Welch algorithm,

$$\log[P(\mathfrak{s}, \mathbb{X}, \mathbb{S}|\lambda)] = \sum_{t=1}^{T-1} \log a_{\mathfrak{s}_t \mathfrak{s}_{t+1}} + \sum_{t=1}^{T} \log f_{\mathfrak{s}_t}(X_t), \tag{6}$$

$$Q(\lambda_n, \lambda) = \sum_{\mathfrak{s} \in \mathfrak{S}} P(\mathfrak{s}|\mathbb{X}, \mathbb{S}, \lambda_n) \sum_{t=1}^{T-1} \log a_{\mathfrak{s}_t \mathfrak{s}_{t+1}} + \sum_{\mathfrak{s} \in \mathfrak{S}} P(\mathfrak{s}|\mathbb{X}, \mathbb{S}, \lambda_n) \sum_{t=1}^{T} \log f_{\mathfrak{s}_t}(X_t). \tag{7}$$

The maximization of the $Q(\lambda_n, \lambda)$ can be separated to maximization w.r.t. transition probability matrix $A = \{a_{ij}\}_{i,j=1}^{N}$ by maximizing the first term and w.r.t. observable densities $F = \{f_i(x)\}_{i=1}^{N}$ by maximizing the second term.

The maximization of Eq. 7 w.r.t. $A$ constrained by $\sum_{j=1}^{N} a_{ij} = 1$ for $1 \le i \le N$ is obtained by re-estimating the parameters $\hat{A} = \{\hat{a}_{ij}\}_{i,j=1}^{N}$ as follows:

$$\hat{a}_{ij} = \frac{\text{expected number of transitions from state } s_i \text{ to state } s_j}{\text{expected number of transitions from state } s_i}$$

$$= \frac{\sum_{(t=1 \text{ and } t \neq t_k, 1 \le k \le K)}^{T-1} P(S_t = s_i, S_{t+1} = s_j|\mathbb{X}, \mathbb{S}, \lambda)}{\sum_{(t=1 \text{ and } t \neq t_k, 1 \le k \le K)}^{T-1} P(S_t = s_i|\mathbb{X}, \mathbb{S}, \lambda)}. \tag{8}$$

This equation is computed using modified forward and backward variables of the Baum-Welch algorithm to reflect the partially annotated states. For the exact derivation of formulas for computation of $\hat{a}_{ij}$ see the Appendix A.

### 3.2.1. Learning Observable Distributions

The maximization of Eq. 7 w.r.t. $F = \{f_i(x)\}_{i=1}^{N}$ depends on assumptions on the system of probability densities $F$. It is usually assumed (e.g., in Baum et al., 1970; Rabiner, 1989) that $F$ is a system of probability distributions of the same type and differ only in their parameters.

In the HMMTxD the $m$-dimensional observed random variables $X_t = (X_t^1, X_t^2, \ldots, X_t^m) \in R^m$ are assumed conditionally independent and to have the beta-distribution, so $f_i(x)$, $1 \le i \le N$ are products of $m$ one-dimensional beta distributions with parameters of shape $\{(p_i{}^j, q_i{}^j)\}_{j=1}^{m}, 1 \le i \le N$. In this case maximization of the second term of the Eq. 7 is an iterative procedure using inverse digamma function which is very computationally expensive (Gupta and Nadarajah, 2004).

We propose to estimate the shape parameters of the beta distributions with a generalized method of moments. The classical method of moments is based on the fact that sample moments of independent observations converge to its theoretical ones due to the law of large numbers for independent random variables. In the HMMTxD observations $\mathbb{X} = \{X_t\}_{t=1}^{T}$ are not independent. The generalized method of moments is based on the fact that $\{X_t - E(X_t|X_1, X_2, \ldots, X_{t-1})\}_{t=1}^{T}$ is a sequence of martingale differences for which the law of large numbers also holds. Using the generalized method of moments gives estimates of the parameters of shape

$$\hat{p}_i^j = \hat{\mu}_i^j \left( \frac{\hat{\mu}_i^j (1 - \hat{\mu}_i^j)}{(\hat{\sigma}_i^j)^2} - 1 \right) \tag{9}$$

and

$$\hat{q}_i^j = (1 - \hat{\mu}_i^j) \left( \frac{\hat{\mu}_i^j (1 - \hat{\mu}_i^j)}{(\hat{\sigma}_i^j)^2} - 1 \right) \tag{10}$$

where

$$\hat{\mu}_i^j = \frac{\sum_{t=1}^{T} X_t^j P(S_t = s_i|\mathbb{X}, \mathbb{S}, \lambda)}{\sum_{t=1}^{T} P(S_t = s_i|\mathbb{X}, \mathbb{S}, \lambda)} \tag{11}$$

and

$$(\hat{\sigma}_i^j)^2 = \frac{\sum_{t=1}^{T} (X_t^j - \hat{\mu}_i^j)^2 P(S_t = s_i|\mathbb{X}, \mathbb{S}, \lambda)}{\sum_{t=1}^{T} P(S_t = s_i|\mathbb{X}, \mathbb{S}, \lambda)}. \tag{12}$$

Let us denote the system of probability densities with re-estimated parameters as $\hat{F} = \{\hat{f}_i(x)\}_{i=1}^{N}$. The generalized method of moments is described in detail in the Appendix B.

### 3.2.2. Algorithm overview

The complete modified Baum-Welch algorithm is summarized in Algorithm 1, where after each iteration $P(\mathbb{X}, \mathbb{S}|\lambda_{n+1}) \ge P(\mathbb{X}, \mathbb{S}|\lambda_n)$ and we repeat these steps until convergence. Note that $\hat{A}_n$ is a maximum likelihood estimate of $A$ therefore always increases $P(\mathbb{X}, \mathbb{S}|\lambda_n)$ (shown in Rabiner (1989)) but $\hat{F}_n$ is estimated by the method of moments so the test on likelihood increase is required ("if statement" in the Algorithm 1). In fact, this algorithm structure match to the generalized EM algorithm (GEM) introduced in Dempster et al. (1977).

**Algorithm 1:** Algorithm for HMM parameters learning.

```
Input: $\mathbb{X}, \mathbb{S}, \lambda_n = (A_n, F_n)$
Output: $\lambda_{n+1} = (A_{n+1}, F_{n+1})$
repeat
    Compute likelihood $P(\mathbb{X}, \mathbb{S}|\lambda_n)$;
    Estimate $\hat{A}_n$ by Eq. 8 and $\hat{F}_n$ by Eq. 9, 10;
    if $P(\mathbb{X}, \mathbb{S}|\hat{A}_n, \hat{F}_n) < P(\mathbb{X}, \mathbb{S}|A_n, F_n)$ then
        $\lambda_{n+1} = (\hat{A}_n, F_n)$
    else
        $\lambda_{n+1} = (\hat{A}_n, \hat{F}_n)$
    $\lambda_n = \lambda_{n+1} = (A_{n+1}, F_{n+1})$;
until convergence $\vee$ max number of iteration;
```

## 4. Feature-based detector

The requirements for the detector are: adjustable operation mode (e.g., set for high precision but possibly low recall), (near) real-time performance and the ability to model pose transformations up to at least similarity (translation, rotation, isotropic scaling). Basically, any detector-like approach can be used and it may vary based on application. We choose to adapt a feature-based detector which has been shown to perform well in image retrieval, object detection and object tracking (Pernici and Bimbo, 2013) tasks.

There are many possible combinations of features and their descriptors with different advantages and drawbacks. We exploit multiple feature types: specifically, Hessian keypoints with the SIFT (Lowe, 2004) descriptor, ORB (Rublee et al., 2011) with BRISK and ORB with FREAK (Ortiz, 2012). Each feature type is handled separately, up to the point where point correspondences are established. A weight is assigned to each feature type $w^g$ and is set to be inversely proportional to the number of features on the reference template, to balance the disparity in individual feature numbers.

The detector works as follows. In the initialization step, features are extracted from the inside and the outside of the region specifying the tracked object. Descriptors of the features outside of the region are stored as the background model.

Usually, the input region is not 100% occupied by the target; therefore, fast color segmentation (Kristan et al., 2014) attempts to

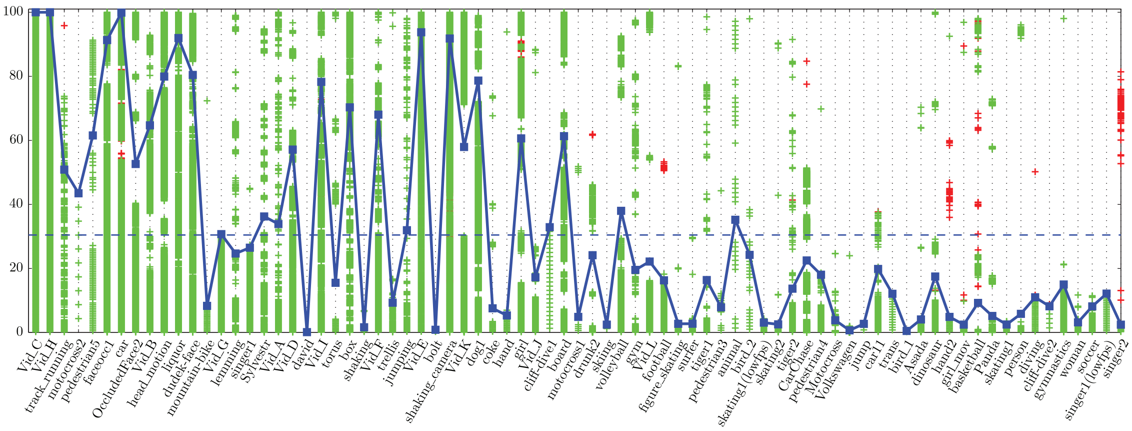


**Fig. 3.** Frames with the detections for 77 sequences data-set. The green marks show the true positive detection and red marks are false positive. The blue line shows the recall of the detector and blue dashed line shows the average recall over all sequences. The length of each sequence is normalized to range (0, 100). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

delineate the object more precisely than the axis-aligned bounding box to remove the features that are most likely not on the target. The step is not critical for the function of the detector, since the bounding box is a fall-back option. We assume that at least 50% of the bounding box is filled with pixels that belong to the target, if the segmentation fails (returns a region containing less than 50% of area of the bounding box), all features in the initial bounding box are used.

Additionally, for each target feature, we use a normal distribution $\mathcal{N}(\mu^f, \sigma^f)$ to model the similarity of the feature to other features. The parameters $\mu^f$ and $\sigma^f$ are estimated in the first frame by randomly sampling 100 features, other than *f*, and computing distances to the feature *f*, from which the mean and variation are computed. This allows defining the quality of correspondence matches in a probabilistic manner for each feature, thus getting rid of global static threshold for the acceptable correspondence distance.

In the detection phase, features are detected and described in the whole image. For each feature $g_i$ from the image the nearest neighbour (in Euclidean space or in Hamming distance metric space, depending on the feature type) feature $b^*$ from the background model and the nearest neighbour feature $f^*$ from the foreground model are computed. A tentative correspondence is formed if the feature match passes the second nearest neighbour test and a probability that the correspondence distance belongs to the outlier distribution is lower than a predefined significance set to 0.1%. So

$$\frac{d(g_i, f^*)}{d(g_i, b^*)} < 0.8 \ \wedge \ \mathcal{F}(d(g_i, f^*)|\mu^{f^*}, \sigma^{f^*}) < 0.1\% \tag{13}$$

where $\mathcal{F}(d|\mu^{f^*}, \sigma^{f^*})$ is a c.d.f. of the normal distribution with parameters $\mu^{f^*}$ and $\sigma^{f^*}$ of a distance distribution of features not corresponding to $f^*$. The 0.1% significance corresponds to the $\mu - 3\sigma$ threshold. Finally, RANSAC estimates the target current pose using a sum of weighted inliers as a cost function for model support

$$\text{cost} = \sum_i w^{g_i} * [g_i == \text{inlier}], \tag{14}$$

which takes into account the different numbers of features per feature type on the target.

The decision whether the detected pose is considered correct depends on the number of weighted inliers that supports the RANSAC-selected transformation and it controls the trade-of between precision and recall of the method. This threshold is automatically computed in the first frame of the sequence as max(5,min(0.03*max_number_of_features_in_target_bbox, 10)). The threshold interval (5,10) and the feature multiplier (0.03) were set experimentally to have the false positive rate close to zero for the most of the testing sequences. Furthermore, majority voting is used to verify that the detection is not in contradiction to the estimated HMM state, i.e. if we are in the state where two or more (majority) trackers are correct and the detector is not consistent with them, the detection is not used. This mitigates the false positive detections, therefore HMM updates, when the trackers works correctly.

The true and false positives for 77 sequences are shown in Fig. 3, where the detector works on almost all sequences with zero false positive rate (0.46% average false positive rate on the dataset) and 30% recall rate. The failure cases of this feature-based detector are mostly caused by the imprecise initial bounding box, which contains large portion of structured background (i.e., background where the detector finds features) and due to the presence of similar object in the scene, e.g., sequences *hand2, basketball, singer2*.

## 5. HMMTxD implementation

To demonstrate the performance of the proposed framework, a pair and a triplet of published short-term trackers were plugged into the framework to show the performance gain by combination of a different number of trackers. As Bailer et al. (2014) pointed out, not all trackers when combined can improve the overall performance (i.e., adding tracking method with similar failure mode will not benefit).

We therefore choose methods that have a different designs and work with different assumptions (e.g., rigid global motion vs. color mean-shift estimation vs. maximum correlation response). These trackers are the Flock of Trackers (FoT) (Vojir and Matas, 2014), scale adaptive mean-shift tracker (ASMS) (Vojir et al., 2013) and kernelized correlation filters (KCF) (Henriques et al., 2015). This choice shows that superior performance can be achieved by using simple, fast trackers (above 100 fps) that may not represent the state-of-the-art. The trackers can be arbitrarily replaced depending on the user application or requirements.

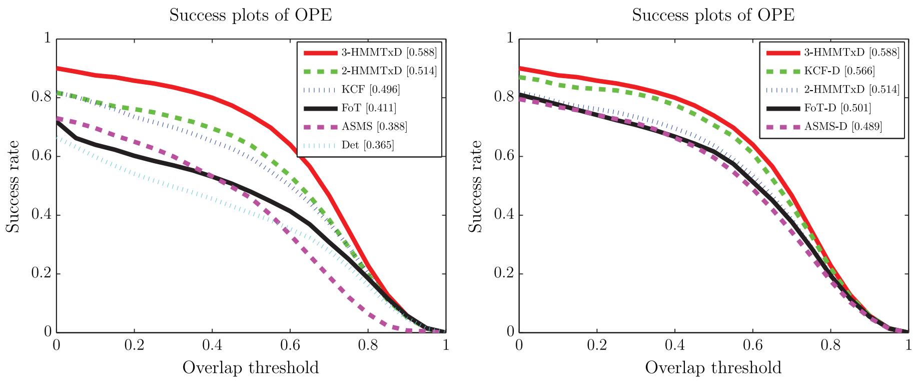


**Fig. 4.** CVPR2013 OPE benchmark comparison of individual trackers and their combination in the proposed HMMTxD. The 2-HMMTxD denotes the combination of FoT and ASMS trackers and 3-HMMTxD is a combination of FoT, ASMS and KCF trackers. Det stands for the proposed detector. The right plot show simple combination of individual trackers with the proposed detector. Suffix "-D" refers to the combination with detector.

*Trackers*

The Flock of Trackers (FoT) (Vojir and Matas, 2014) evenly covers the object with patches and establishes frame-to-frame correspondence by the Lucas-Kanade method (Lucas and Kanade, 1981). The global motion of the target is estimated by RANSAC.

The second tracker is a scale adaptive mean-shift tracker (ASMS) (Vojir et al., 2013) where the object pose is estimated by minimizing the distance between RGB histograms of the reference and the candidate bounding box. The KCF (Henriques et al., 2015) tracker learns a correlation filter by ridge regression to have high response to target object and low response on background. The correlation is done in the Fourier domain which is very efficient.

These three trackers have been selected since they are complementary by design. FoT enforces a global motion constrain and works best for rigid object with texture. On the other hand, ASMS does not enforce object rigidity and is well suited for articulated or deformable objects assuming their color distribution is discriminative w.r.t. the background. KCF can be viewed as a tracking-by detection approach using sliding window like scanning.

For each tracker position, two global observable measurements are computed, namely the Hellinger distance between the target template histogram and the histogram of the current position and normalized cross-correlation score of the current patch and the target model patch. These target models are initialized in the first frame and then updated exponentially with factor of 0.5 during each positive detection of the detector part. Additionally, each tracker produces its own estimate of performance. For FoT it is the number of predicted correspondences (for details please see Vojir and Matas, 2014) that support the global model. For ASMS it is the Hellinger distance between its histogram model and current neighbourhood background (i.e., color similarity of the object and background) and for KCF it is a correlation response of the tracking procedure.

## 6. Experiments

The HMMTxD was compared with state-of-the-art methods on two standard benchmarks and on a dataset TV77[2] containing 77 public video sequences collected from tracking-related publications. The dataset exhibits wider diversity of content and variability of conditions than the benchmarks.

Parameters of the method were fixed for all the experiments. In the HMM, the initial beta distribution shape parameters $(p, q)$ were set to (2, 1) for correct state (1) and (1, 2) for fail state (0) for all observations and the transition matrix was set to prefer staying in the current state. The transition matrix has 0.98 on diagonal, 0 in fist column, 0.001 in last column, $1e-10$ in last row and 0.05 otherwise. The matrix is normalized so that rows sum to one. States in the matrix are binary encoded starting from the left column which corresponds to the state $s_1 = (1, \ldots, 1)$. The number of iteration for Baum-Welch algorithm was set to 3.

The processing speed on the VOT2015 dataset is (in frames per second) minimum 1.03, maximum 33.72 and average 10.83 measured on a standard notebook with Intel Core-i7 processor. This speed is mostly affected by the number of features detected in the images which correlates to the resolution of the image (in the dataset the range is from 320x180 to 1280x720).

First, we compare the performance of individual parts of the HMMTxD framework (i.e., KCF, ASMS, FoT trackers) and their combination via HMM as proposed in this paper. Two variants of HMMTxD are evaluated – 2-HMMTxD refers to combination of FoT and ASMS trackers and the 3-HMMTxD to combination of all mentioned trackers. We also show the benefit of the proposed detector when simply combined with the individual trackers in such way that if detector fires the tracker is re-initialized. The Fig. 4 shows the benefit gained from the detector and further consistent improvement achieved by the combination of the trackers. More detailed per sequence analysis on the TV77 dataset (Figs. 5 and 6) shows more clearly the efficiency of learning tracker performance online. In almost all sequences the HMMTxD is able to identify and learn which trackers works correctly and achieve the performance of at least the best tracker or higher (e.g., *motocross1, skating1(low), Volkswagen, singer1, pedestrian3, surfer*). Most notable failure cases are caused by the detector failure, e.g., in sequences *singer2, woman, skating1, basketball, girl_mov*.

In all other experiments, the abbreviation HMMTxD refers to the combination of all 3 trackers.

**Evaluation on the CVPR2013 Benchmark** (Wu et al., 2013) that contains 50 video sequences. Results on the benchmark have been published for about 30 trackers. The benchmark defines three types of experiments: (i) one-pass evaluation (OPE) – a tracker initialized in the first frame is run to the end of the sequence, (ii) temporal robustness evaluation (TRE) – the tracker is

[2] http://cmp.felk.cvut.cz/~vojirtom/dataset/index.html

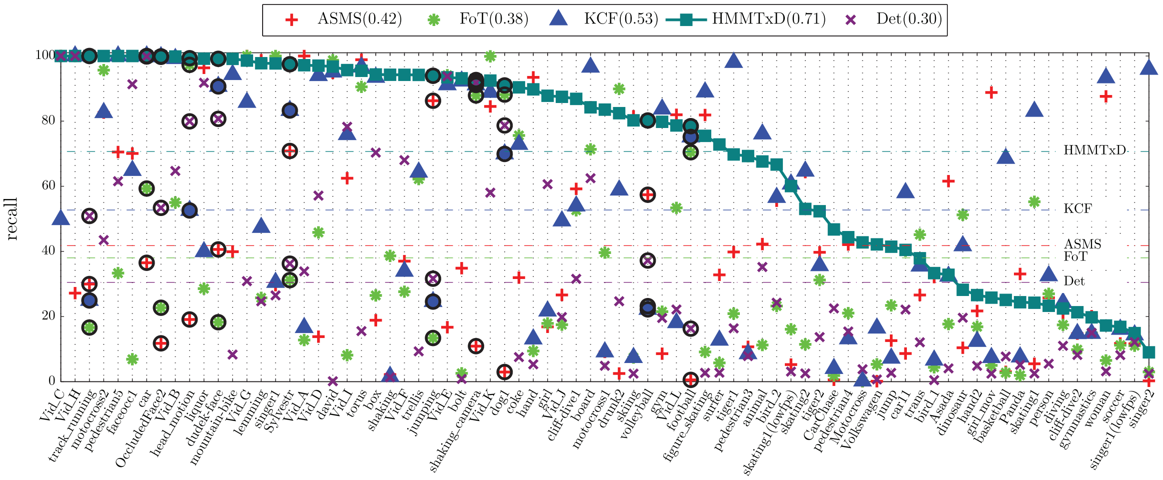


**Fig. 5.** Per sequence analysis of the single trackers (i.e., KCF, ASMS, FoT) and the proposed HMMTxD. The average recall is shown by the dashed lines (precise number is in the legend). Black circles mark gray-scale sequences. The sequences are ordered by HMMTxD performance.

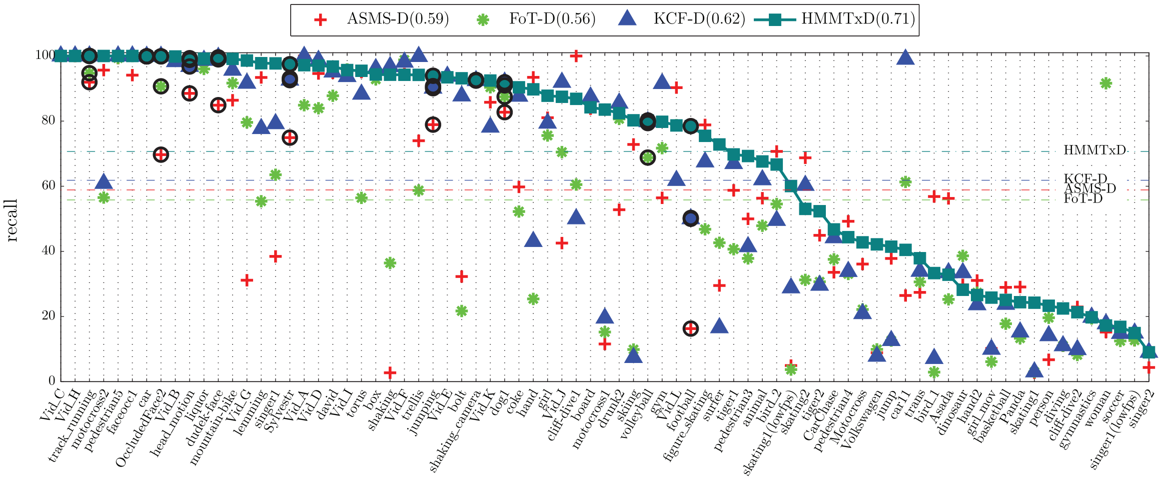


**Fig. 6.** Per sequence analysis of the single trackers combined with the detector (i.e., KCF-D, ASMS-D, FoT-D) and the proposed HMMTxD. The average recall is shown by the dashed lines (precise number is in the legend). Black circles mark grayscale sequences. The sequences are ordered by HMMTxD performance.

initialized and starts at a random frame, and (iii) spatial robustness evaluation (SRE) – the initialization is perturbed spatially. Performance is measured by precision (spatial accuracy, i.e., center distance of ground truth and reported bounding box) and success rate (the number of frames where overlap with the ground truth was higher than a threshold). The results are visualized in Fig. 7 where only results of the 10 top performing trackers are plotted. Together with the tracker from this benchmark, we also added the MEEM (Zhang et al., 2014) tracker, which is a recent state-of-the-art tracker. The proposed HMMTxD outperforms all trackers in the success rate in all three experiments. Its precision is comparable to MEEM (Zhang et al., 2014) the top performing tracker in terms of precision. HMMTxD outperforms significantly the OPE results reported in Wang and yan Yeung (2014), where five top performing trackers from this particular benchmark were used for combination (other experiments were not reported in the paper).

**VOT2013 benchmark** (Kristan et al., 2013) evaluates trackers on a collection containing 16 sequences carefully selected from a large pool by a semi- automatic clustering method. For comparison, results of 27 tracking methods are available and the added MEEM tracker was evaluated by us using default setting from the publicly available source code. The performance is measured by accuracy, average overlap with the ground truth, and robustness, the number of re-initialization of the tracker so that it is able to track the whole sequence. Average rank of trackers is used as an overall performance indicator.

In this benchmark, the proposed HMMTxD achieves clearly the best accuracy (Fig. 8). With less than one re-initialization per

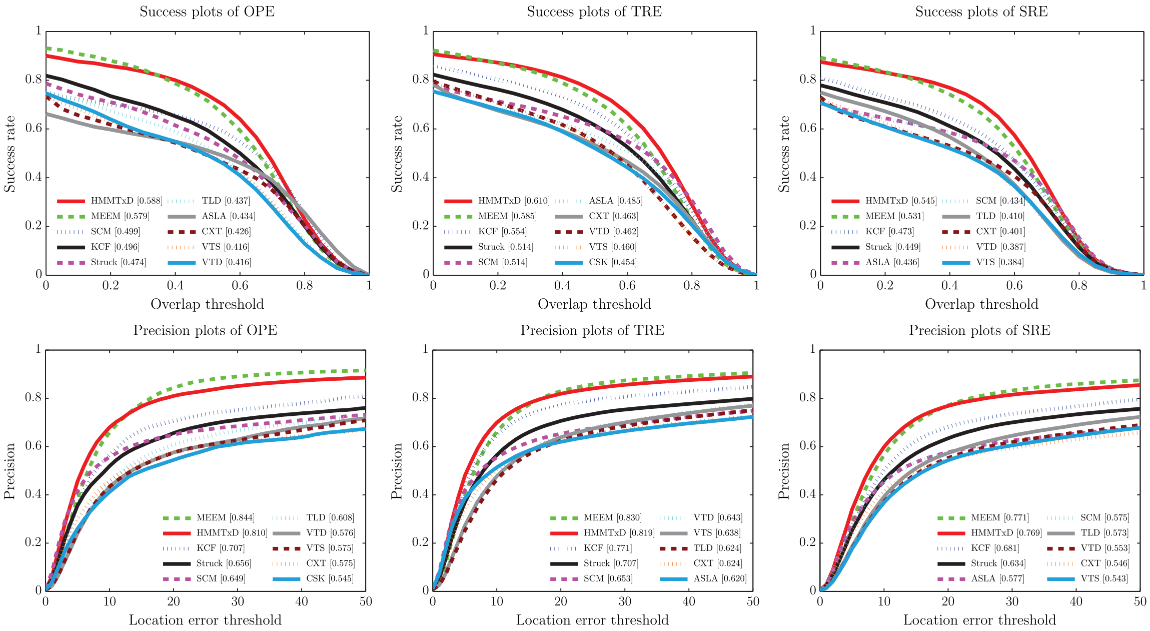
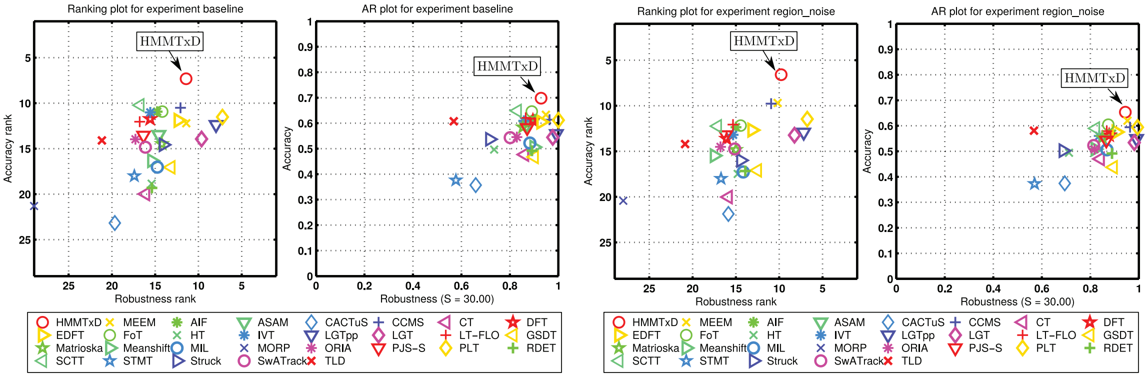


**Fig. 7.** Evaluation of HMMTxD on the CVPR2013 Benchmark (Wu et al., 2013). The top row shows the success rate as a function of the overlap threshold. The bottom row shows the precision as a function of the localization error threshold. The number in the legend is AUC, the area under ROC-curve, which summarizes the overall performance of the tracker for each experiment.



(a) baseline experiment

(b) region-noise experiment

**Fig. 8.** Evaluation of HMMTxD on the VOT 2013 Benchmark (Kristan et al., 2013). HMMTxD result is shown as the red circle. The left plot shows the ranking in accuracy (vertical axis) and robustness (horizontal axis) and the right plot shows the raw average values of accuracy and robustness (normalized to the (0, 1) interval). For both plots the top right corner is the best performance.

sequence it performs slightly worse in terms of robustness due to two reasons.

Firstly, the HMM recognizes a tracker problem with a delay and switching to other tracker (here even one frame where the overlap with ground truth is zero leads to penalization) and secondly the VOT evaluation protocol, which require re-initialization after failure and to forget all previously learned models (the VOT2013 refer to this as causal tracking), therefore the learned performance of the trackers is forgotten and has to be learned from scratch.

The results for the baseline and region-noise experiments are shown in Fig. 8. Note that the ranking of the methods differs from the original publication since two new methods (HMMTxD and MEEM) were added and the relative ranking of the methods changed. The top three performing trackers and their average ranks are HMMTxD (8.77), PLT (9.24), LGTpp (Xiao et al., 2013) (10.11). MEEM tracker ends up at the fifth place with average rank 10.87. The rankings were obtained by the toolkit provided by the VOT in default settings for baseline and region noise experiments.

The second best performing method on the VOT2013 is the unpublished PLT for which just a short description is available in Kristan et al. (2013). PLT is a variation of structural SVM that uses multiple features (color, gradients). STRUCK (Hare et al., 2011) and

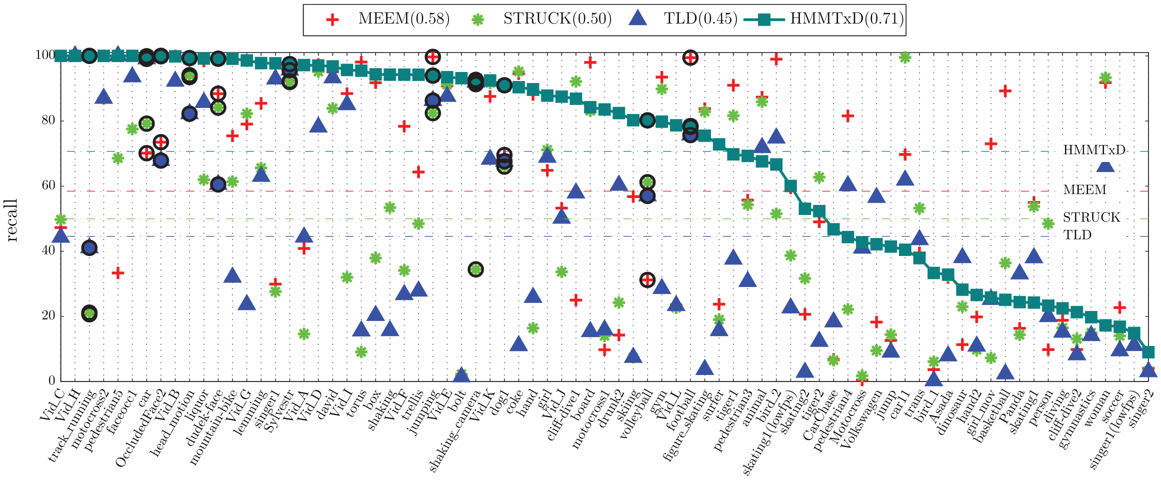


**Fig. 9.** Evaluation of state-of-the-art trackers on the TV77 data-set in terms of recall, i.e. number of correctly tracked frames. The average recall is shown by the dashed lines (precise number is in the legend). Black circles mark gray-scale sequences. The sequences are ordered by HMMTxD performance.

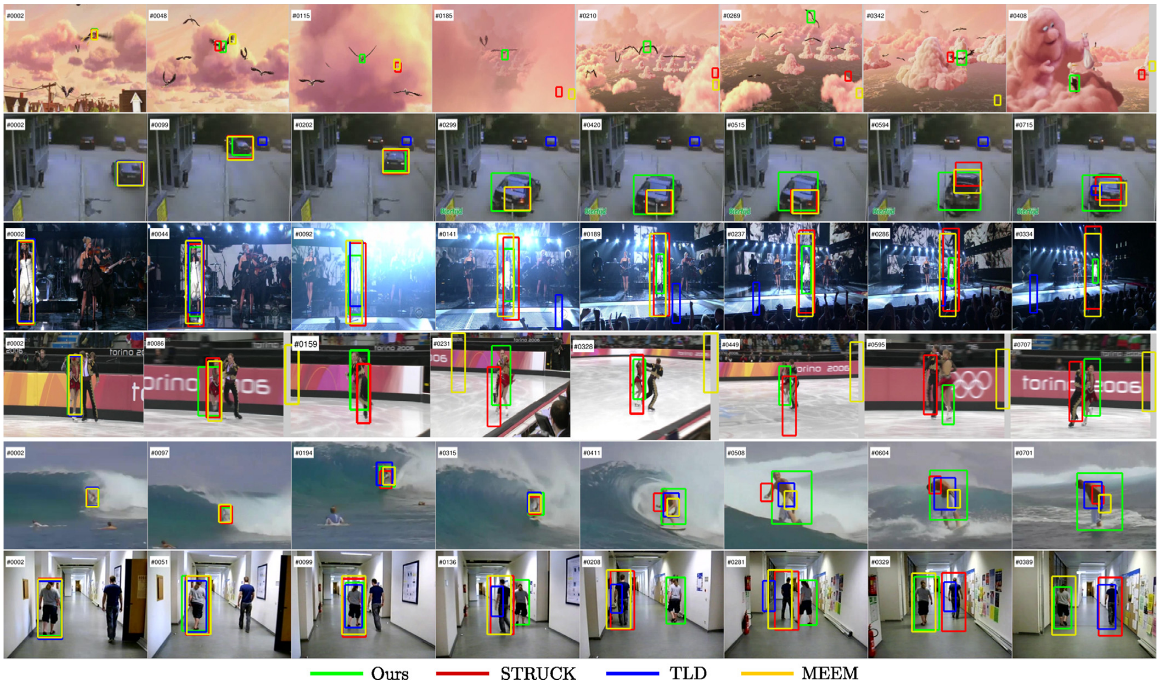


**Fig. 10.** Qualitative comparison of the state-of-the-art trackers on challenging sequences from the TV77 data-set (from top *bird_1, drunk2, singer1, skating2, surfer, Vid_J*).

MEEM (Zhang et al., 2014) are similar method to the PLT based on SVM classification. We compared these method with HMMTxD on the diverse 77 videos along with the TLD (Kalal et al., 2012) which has a similar design as HMMTxD. HMMTxD outperforms all these methods by a large margin on average recall – measured as number of frames where the tracker overlap with ground truth is higher than 0.5 averaged over all sequences. Results are shown in Fig. 9. Qualitative comparison of these state-of-the-art methods is shown in Fig. 10. Even for sequences with lower recall (e.g. *bird_1, skating2*), the HMMTxD is able to follow the object of interest.

## 7. Conclusions

A novel method called HMMTxD for fusion of multiple trackers has been proposed. The method utilizes an on-line trained HMM to estimate the states of the individual trackers and to fuse a different types of observables provided by the trackers. The HMMTxD outperforms its constituent parts (FoT, ASMS, KCF, Detector and its combinations) by a large margin and shows the efficiency of the HMM with combination of three trackers.

HMMTxD outperforms all methods included in the CVPR2013 benchmark and perform favorably against most recent state-of-the-art tracker. The HMMTxD also outperforms all method of the VOT2013 benchmark in accuracy, while maintaining very good robustness, and ranking in the first place in overall ranking. Experiments conducted on a diverse dataset TV77 show that the HMMTxD outperforms state-of-the-art MEEM, STRUCK and TLD methods, which are similar in design, by a large margin. The processing speed of the HMMTxD is $5 - 10$ frames per second on average, which is comparable with other complex tracking methods.

## Acknowledgements

The research was supported by the Czech Science Foundation Project GACR P103/12/G084 and by the Technology Agency of the Czech Republic project TE01020415 (V3C – Visual Computing Competence Center).

## Appendix A. Forward-backward procedure for modified Baum-Welch algorithm

Let us assume the HMM with $N$ possible states $\{s_1, s_2, \dots, s_N\}$, the matrix of state transition probabilities $A = \{a_{ij}\}_{i,j=1}^{N}$, the vector of initial state probabilities $\pi = (1, 0, 0, \dots, 0)$, the initial state $s_1 = (1, 1, \dots, 1)$, a sequence of observations $\mathbb{X} = \{X_t\}_{t=1}^{T}, X_t \in R^m$ and $F = \{f_i(x)\}_{i=1}^{N}$ the system of conditional probability densities of observations conditioned on $S_t = s_i$.

Let $0 = t_0 < t_1 < t_2 \dots < t_K \le T$ be a sequence of detection times, $\mathbb{S} = \{S_{t_k} = s_{i_k}, \{t_k\}_{k=1}^{K}\}$ be observed states of Markov chain, marked by the detector, and $S_{t_k+1} = s_1$ for $0 \le k \le K$.

The forward variable for the Baum-Welch algorithm is defined as follows. Let $1 \le i \le N, 1 \le k \le K, t_{(k-1)} < t \le t_k$ and

$$\alpha_t(i) = P(X_{t_{(k-1)}+1}, \dots, X_t, S_t = s_i|\lambda) \text{ then} \tag{A.1}$$

$$\alpha_{t_{(k-1)}+1}(1) = f_1(X_{t_{(k-1)}+1}), \tag{A.2}$$

$$\alpha_{t_{(k-1)}+1}(i) = 0 \text{ for } i \neq 1 \tag{A.3}$$

and for $t_{(k-1)} < t < t_k$

$$\alpha_{(t+1)}(i) = \sum_{j=1}^{N} \alpha_t(j) a_{ji} f_i(X_{t+1}), \tag{A.4}$$

$$P(S_t = s_i|X_1, \dots X_t, S_{t_1}, S_{t_2}, \dots, S_{t_{(k-1)}}, \lambda) = \frac{\alpha_t(i)}{\sum_{j=1}^{N} \alpha_t(j)}. \tag{A.5}$$

For $t_K < t < T$ the forward variable is in principle the same as above with $t_{(k-1)} = t_K$. So

$$P(X_{t_K+1}, \dots, X_T|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{A.6}$$

$$P(\mathbb{X}, \mathbb{S}|\lambda) = \prod_{k=1}^{K} \alpha_{t_k}(i_k) * \sum_{i=1}^{N} \alpha_T(i) \quad \text{where } S_{t_k} = s_{i_k}. \tag{A.7}$$

The backward variable for $t_{(k-1)} < t < t_k$ is

$$\beta_t(i) = P(X_{t+1}, \dots, X_{t_k}, S_{t_k}|S_t = s_i, \lambda), \tag{A.8}$$

where $\beta_{t_k}(i_k) = 1$ and $\beta_{t_k}(i) = 0$ for $i \neq i_k$ and

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} f_j(X_{t+1}) \beta_{t+1}(j). \tag{A.9}$$

For $t_K < t < T$ the backward variable is in principle the same as above where $\beta_T(i) = 1$ for $1 \le i \le N$.

Given the forward and backward variables, we get the following probabilities, that are used to update parameters of HMM. For $0 < t < T$ and $t \neq t_k$, $1 \le k \le K$

$$P(S_t = s_i, S_{t+1} = s_j|\mathbb{X}, \mathbb{S}, \lambda) = \tag{A.10}$$

$$\frac{\alpha_t(i) a_{ij} f_j(X_{t+1}) \beta_{(t+1)}(j)}{\sum_{k=1}^{N} \sum_{l=1}^{N} \alpha_t(k) a_{kl} f_l(X_{t+1}) \beta_{t+1}(l)} \tag{A.11}$$

and for $0 < t \le T$

$$P(S_t = s_i|\mathbb{X}, \mathbb{S}, \lambda) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)}. \tag{A.12}$$

The final equation for the update of transition probabilities $A$ of HMM is as follows.

$$\hat{a}_{ij} = \frac{\text{expected number of transitions from state } s_i \text{ to state } s_j}{\text{expected number of transitions from state } s_i} \tag{A.13}$$

$$= \frac{\sum_{(t=1 \text{ and } t \neq t_k, 1 \le k \le K)}^{T-1} P(S_t = s_i, S_{t+1} = s_j|\mathbb{X}, \mathbb{S}, \lambda)}{\sum_{(t=1 \text{ and } t \neq t_k, 1 \le k \le K)}^{T-1} P(S_t = s_i|\mathbb{X}, \mathbb{S}, \lambda)}. \tag{A.14}$$

## Appendix B. Generalized method of moments

For a simplification let us assume HMM with one-dimensional observed random variables $\{X_t\}_{t=1}^{+\infty}, X_t \in R$. The sequence $\{X_t - E(X_t|X_1, X_2, \dots, X_{t-1})\}_{t=1}^{+\infty}$ is a martingale difference series where

$$E(X_t|X_1, X_2, \dots, X_{t-1}) = \sum_{i=1}^{N} E(X_t|X_1, X_2, \dots, X_{t-1}, S_t = i) P(S_t = i) \tag{B.1}$$

$$= \sum_{i=1}^{N} E(X_t|S_t = i) P(S_t = i). \tag{B.2}$$

Under the assumption that $\{X_t\}_{t=1}^{+\infty}$ are uniformly bounded random variables i.e. $|X_t| < c, c \in (0, +\infty)$ for all $t \ge 1$, the strong law of large numbers for a sum of martingale differences can be used(see Theorem 2.19 in Hall and Heyde (1980)). So

$$\lim_{T \to +\infty} \frac{1}{T} \sum_{t=1}^{T} \left[ X_t - \sum_{i=1}^{N} E(X_t|S_t = i) P(S_t = i) \right] = 0 \text{ almost surely.} \tag{B.3}$$

Let us denote $\mu_i = E(X_t|S_t = i)$ for $1 \le t \le T$ and $\hat{\mu}_i$ the estimate of $\mu_i$ based on the modified method of moments. The estimate $\hat{\mu}_i$ is a solution of a following equation w.r.t. $\mu_i$

$$\frac{1}{T} \sum_{t=1}^{T} X_t = \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{N} \mu_i P(S_t = i). \tag{B.4}$$

Having one equation for $N$ unknown variables $\mu_i$, $1 \le i \le N$ it is necessary to add some constrains to get a unique solution. We propose to minimize

$$\sum_{t=1}^{T} \sum_{i=1}^{N} (X_t - \mu_i)^2 P(S_t = i), \tag{B.5}$$

w.r.t. $\mu_i$, $1 \le i \le N$ giving

$$\hat{\mu}_i = \frac{\sum_{t=1}^{T} X_t P(S_t = s_i)}{\sum_{t=1}^{T} P(S_t = s_i)} \tag{B.6}$$

which satisfy the moment Eq. (B.4). The same way of reasoning can be used for higher moments of $\{X_t\}_{t=1}^{T}$. For example using

$\{(X_t)^2\}_{t=1}^T$ we get estimates $\hat{\sigma}_i^2$ for $\sigma_i^2 = \text{var}(X_t|S_t = i)$ for $1 \le t \le T$,

$$\hat{\sigma}_i^2 = \frac{\sum_{t=1}^T (X_t - \hat{\mu}_i)^2 P(S_t = s_i)}{\sum_{t=1}^T P(S_t = s_i)}. \tag{B.7}$$

In the HMMTxD $m$-dimensional observed random variables $X_t = (X_t^1, X_t^2, \ldots, X_t^m)$ are assumed, each of them having beta- distribution and being conditionally independent. There are well-known relations for a mean value $EX$ and a variance $varX$ of a random variable $X$ having beta distribution and its shape parameters $(p, q)$

$$p = EX\left(\frac{EX(1 - EX)}{\text{var}X} - 1\right) \tag{B.8}$$

and

$$q = (1 - EX)\left(\frac{EX(1 - EX)}{\text{var}X} - 1\right). \tag{B.9}$$

Using the modified method of moments gives

$$\hat{\mu}_i^j = \frac{\sum_{t=1}^T X_t^j P(S_t = s_i|\mathbb{X}, \mathbb{S}, \lambda)}{\sum_{t=1}^T P(S_t = s_i|\mathbb{X}, \mathbb{S}, \lambda)} \tag{B.10}$$

and

$$(\hat{\sigma}_i^j)^2 = \frac{\sum_{t=1}^T (X_t^j - \hat{\mu}_i^j)^2 P(S_t = s_i|\mathbb{X}, \mathbb{S}, \lambda)}{\sum_{t=1}^T P(S_t = s_i|\mathbb{X}, \mathbb{S}, \lambda)}. \tag{B.11}$$

Then

$$\hat{p}_i^j = \hat{\mu}_i^j\left(\frac{\hat{\mu}_i^j(1 - \hat{\mu}_i^j)}{(\hat{\sigma}_i^j)^2} - 1\right) \tag{B.12}$$

and

$$\hat{q}_i^j = (1 - \hat{\mu}_i^j)\left(\frac{\hat{\mu}_i^j(1 - \hat{\mu}_i^j)}{(\hat{\sigma}_i^j)^2} - 1\right). \tag{B.13}$$

If we assume in our model $\lambda = (A, F)$ that for some $\{(i_r, j_r) \in \{1, 2, \ldots, N\} \times \{1, 2, \ldots, m\} : p_{i_r}^{j_r} = p, q_{i_r}^{j_r} = q\}_{r=1}^R$ then

$$\hat{p} = \hat{\mu}\left(\frac{\hat{\mu}(1 - \hat{\mu})}{\hat{\sigma}^2} - 1\right) \tag{B.14}$$

and

$$\hat{q} = (1 - \hat{\mu})\left(\frac{\hat{\mu}(1 - \hat{\mu})}{\hat{\sigma}^2} - 1\right) \tag{B.15}$$

where

$$\hat{\mu} = \frac{\sum_{r=1}^R \sum_{t=1}^T X_t^{j_r} P(S_t = s_{i_r}|\mathbb{X}, \mathbb{S}, \lambda)}{\sum_{r=1}^R \sum_{t=1}^T P(S_t = s_{i_r}|\mathbb{X}, \mathbb{S}, \lambda)} \tag{B.16}$$

and

$$\hat{\sigma}^2 = \frac{\sum_{r=1}^R \sum_{t=1}^T (X_t^{j_r} - \hat{\mu})^2 P(S_t = s_{i_r}|\mathbb{X}, \mathbb{S}, \lambda)}{\sum_{r=1}^R \sum_{t=1}^T P(S_t = s_{i_r}|\mathbb{X}, \mathbb{S}, \lambda)}. \tag{B.17}$$

## References

Bailer, C., Pagani, A., Stricker, D., 2014. A superior tracking approach: Building a strong tracker through fusion. In: European Conference on Computer Vision.

Baum, L.E., Petrie, T., Soules, G., Weiss, N., 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Annal. Math. Stat. 41 (1), 164–171.

Cehovin, L., Kristan, M., Leonardis, A., 2013. Robust visual tracking using an adaptive coupled-layer visual model. IEEE Trans. Pattern Anal. Mach. Intell. 35 (4), 941–953.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J.R. Stat. Soc. Ser. B 39 (1), 1–38.

Godec, M., Roth, P.M., Bischof, H., 2011. Hough-based tracking of non-rigid objects. In: International Conference on Computer Vision.

Grabner, H., Matas, J., Van Gool, L., Cattin, P., 2010. Tracking the invisible: Learning where the object might be. In: Computer Vision and Pattern Recognition, pp. 1285–1292.

Gupta, A., Nadarajah, S., 2004. Handbook of Beta Distribution and Its Applications. Statistics: A Series of Textbooks and Monographs. CRC Press, Boca Raton, Florida.

Hall, P., Heyde, C., 1980. Martingale limit theory and its application. Probability and mathematical statistics. Academic Press, New York.

Hare, S., Saffari, A., Torr, P.H.S., 2011. Struck: Structured output tracking with kernels. In: International Conference on Computer Vision, pp. 263–270.

Henriques, J.F., Caseiro, R., Martins, P., Batista, J., 2015. High-speed tracking with kernelized correlation filters. IEEE Trans. Pattern Anal. Mach. Intell. 37 (3), 583–596.

Kalal, Z., Mikolajczyk, K., Matas, J., 2012. Tracking-learning-detection. IEEE Trans. Pattern Anal. Mach. Intell. 34 (7), 1409–1422.

Kristan, M., Matas, J., Leonardis, A., Vojir, T., Pflugfelder, R.P., Fernández, G., Nebehay, G., Porikli, F., Cehovin, L., 2015. A novel performance evaluation methodology for single-target trackers. arXiv 2015 abs/1503.01313.

Kristan, M., Perš, J., Sulic, V., Kovacic, S., 2014. A graphical model for rapid obstacle image-map estimation from unmanned surface vehicles. In: Asian Conference on Computer Vision. Accepted, to be published.

Kristan, M., Pflugfelder, R., Leonardis, A., Matas, J., Porikli, F., Cehovin, L., Nebehay, G., Fernandez, G., Vojir, T., et al., 2013. The visual object tracking vot2013 challenge results. In: The IEEE International Conference on Computer Vision (ICCV) Workshops.

Kwon, J., Lee, K.M., 2009. Tracking of a non-rigid object via patch-based dynamic appearance modeling and adaptive basin hopping monte carlo sampling.. In: Computer Vision and Pattern Recognition, pp. 1208–1215.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60 (2), 91–110.

Lucas, B.D., Kanade, T., 1981. An iterative image registration technique with an application to stereo vision. In: International Joint Conference on Artificial Intelligence.

Ortiz, R., 2012. Freak: Fast retina keypoint. In: Conference on Computer Vision and Pattern Recognition, pp. 510–517. Washington, DC, USA.

Pernici, F., Bimbo, A.D., 2013. Object tracking by oversampling local features. IEEE Trans. Pattern Anal. Mach. Intell. 99 (PrePrints), 1.

Rabiner, L., 1989. A tutorial on hidden markov models and selected applications in speech recognition. Proc. IEEE 77 (2), 257–286.

Rublee, E., Rabaud, V., Konolige, K., Bradski, G., 2011. Orb: An efficient alternative to sift or surf. In: International Conference on Computer Vision, pp. 2564–2571. Washington, DC, USA.

Santner, J., Leistner, C., Saffari, A., Pock, T., Bischof, H., 2010. PROST Parallel Robust Online Simple Tracking. Computer Vision and Pattern Recognition. San Francisco, CA, USA.

Smeulder, A., Chu, D., Cucchiara, R., Calderara, S., Deghan, A., Shah, M., 2013. Visual tracking: an experimental survey. IEEE Trans. Pattern Anal. Mach. Intell. (2013).

Vojir, T., Matas, J., 2014. The enhanced flock of trackers. In: Registration and Recognition in Images and Videos. In: Studies in Computational Intelligence, 532, pp. 113–136.

Vojir, T., Noskova, J., Matas, J., 2013. Robust scale-adaptive mean-shift for tracking. In: Image Analysis. In: Lecture Notes in Computer Science, 7944, pp. 652–663.

Wang, N., yan Yeung, D., 2014. Ensemble-based tracking: Aggregating crowdsourced structured time series data. In: Jebara, T., Xing, E.P. (Eds.), Proceedings of the Thirty-First International Conference on Machine Learning, pp. 1107–1115.

Wu, Y., Lim, J., Yang, M.-H., 2013. Online object tracking: A benchmark. In: Computer Vision and Pattern Recognition, pp. 2411–2418.

Xiao, J., Stolkin, R., Leonardis, A., 2013. An enhanced adaptive coupled-layer lgtracker++. In: Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on, pp. 137–144.

Yilmaz, A., Javed, O., Shah, M., 2006. Object tracking: A survey. ACM Comput. Surv. 2006.

Yoon, J.H., Kim, D.Y., Yoon, K.-J., 2012. Visual tracking via adaptive tracker selection with multiple features. In: Proceedings of the Twelth European Conference on Computer Vision - Volume Part IV, pp. 28–41.

Yuan, Y., Yang, H., Fang, Y., Lin, W., 2015. Visual object tracking by structure complexity coefficients. Multim., IEEE Trans. 17 (8), 1125–1136.

Zhang, J., Ma, S., Sclaroff, S., 2014. MEEM: robust tracking via multiple experts using entropy minimization. In: Proceedings of the European Conference on Computer Vision.

Zhou, X., Lu, Y., 2010. Abrupt motion tracking via adaptive stochastic approximation Monte Carlo sampling. In: Computer Vision and Pattern Recognition, pp. 1847–1854.

## D On Bayesian analysis of on-off measurements

D.Nosek, J.Noskova, *On Bayesian analysis of on-off measurements*, Nuclear Instruments and Methods in Physics Research A 820, 23-33, 2016 [arXiv:1603.03386]. DOI: 10.1016/j.nima.2016.02.094

Nuclear Instruments and Methods in Physics Research A 820 (2016) 23–33

Contents lists available at ScienceDirect

Nuclear Instruments and Methods in Physics Research A

journal homepage: www.elsevier.com/locate/nima

# On Bayesian analysis of on–off measurements

Dalibor Nosek [a],[*], Jana Nosková [b]

[a] Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic
[b] Czech Technical University, Faculty of Civil Engineering, Prague, Czech Republic

ARTICLE INFO

Article history:
Received 27 June 2015
Received in revised form
26 February 2016
Accepted 28 February 2016
Available online 2 March 2016

Keywords:
Bayesian inference
On–off problem
Source detection
Gamma-ray astronomy

ABSTRACT

We propose an analytical solution to the on–off problem within the framework of Bayesian statistics. Both the statistical significance for the discovery of new phenomena and credible intervals on model parameters are presented in a consistent way. We use a large enough family of prior distributions of relevant parameters. The proposed analysis is designed to provide Bayesian solutions that can be used for any number of observed on–off events, including zero. The procedure is checked using Monte Carlo simulations. The usefulness of the method is demonstrated on examples from γ-ray astronomy.

## 1. Introduction

We consider an on–off experiment that is designed for counting two classes of events, source and background events, the type of which cannot be distinguished in principle. These events are registered in two disjoint regions characterized by some sets of coordinates. We deal with small numbers of events the detection rates of which are modeled as independent Poisson processes with unknown means.

The problem of the on–off measurement we want to solve is whether the same emitter with a constant but unknown intensity is responsible for the observed counts in both studied regions. Any inconsistency between the numbers of events collected in these regions, when they are properly normalized, then speaks in favor of the predominance of a source producing more events in one of the explored region over the other.

Techniques for addressing these issues from a classical point of view have been presented in the literature. The likelihood ratio test together with the Wilks' theorem [1] are often utilized to characterize asymptotically the level of agreement between the data and the assumption of new phenomena, see e.g. Refs. [2–4]. A widely discussed problem is of how to establish upper bounds of the source intensity for small numbers of detected counts [4–7]. A Bayesian solution of the on–off problem was proposed in Refs. [8,9] when analyzing multichannel spectra in nuclear physics. More general solutions can be found in Refs. [10–13]. Very recently, two specific Bayesian solutions to the on–off problem have been presented [14,15].

In this study, we focus on how to confirm the presence of a weak source and on the determination of credible intervals for its intensity at a given level of significance. We address these issues from a Bayesian point of view. Our original intent is to provide different insights pertaining to the on–off problem that benefit from its simplicity. We do not follow Bayesian alternatives to classical hypothesis testing that deal with priors for models (hypotheses) and compare competing models in terms of Bayes factors. In our concept, the plausibility of possible models for the on–off data is assessed by parametrizing the space of models using a suitable parameter. Here we use the difference between the on-source and background means inferred from the on–off measurement, treating these quantities as random variables within the Bayesian setting. Information on the various aspects of the investigated phenomena are accessible in the posterior distribution of this difference.

Our à priori knowledge about the underlying Poisson processes is consistently improved by using only the on–off data without any external assumptions. By finding the extent to which the on-source Poisson mean is greater or less than the background one, this option allows us to obtain a new well-reasoned formula for the Bayesian probability of the source presence in the on-source zone. As other Bayesian approaches to the on–off problem, we also receive solutions in the case of small numbers, including the null experiment or the experiment with no background, when classical methods based on the asymptotic properties of the likelihood ratio statistic [2–4] are not easily applicable. In addition, our strategy allows us to establish limits on parameters for the processes that are responsible for the observed phenomena. There are no problems with the discreteness of counting experiments or with unphysical likelihood estimates, see e.g. Refs. [5–7]. We provide credible intervals that are

[*] Corresponding author.
*E-mail address:* nosek@ipnp.troja.mff.cuni.cz (D. Nosek).

http://dx.doi.org/10.1016/j.nima.2016.02.094
0168-9002/© 2016 Elsevier B.V. All rights reserved.

likely to be very similar to the way in which the experimental results are usually communicated.

The proposed method is particularly suitable when dealing with peculiar sources whose observational conditions cannot be set in an optimal way. Examples include transient γ-ray sources and gamma-ray bursts when searching for an accompanying signal on targets in their space-time vicinity. This approach can also advantageously be used to assess possible sources of charged cosmic rays with characteristics hypothesized in previous measurements.

The structure of the paper is as follows. The essential features of our approach are described in detail in Section 2. We present and discuss general formulae for assessing the existence of the source and for estimating its activity. Particular attention is paid to cases with uninformative priors. Examples are presented and discussed in Section 3. The paper is concluded in Section 4.

## 2. Bayesian solutions to on–off problem

In a typical on–off analysis, a measurement of a physical quantity of interest is set by comparing the number of events $n_{\mathrm{on}}$ recorded in a signal (on-source) region, where a source is expected, with the number of events $n_{\mathrm{off}}$ detected in a reference (off-source) region. The on–off data, especially when only a few events are recorded, are modeled as discrete random variables generated in two independent Poisson processes with unknown on- and off-source means, $\mu_{\mathrm{on}}$ and $\mu_{\mathrm{off}}$, i.e. $n_{\mathrm{on}} \sim \mathrm{Po}(\mu_{\mathrm{on}})$ and $n_{\mathrm{off}} \sim \mathrm{Po}(\mu_{\mathrm{off}})$,[1] respectively.

The relationship between both the on- and off-source regions is given by the ratio of on- and off-source exposures, by the on–off parameter $\alpha > 0$. This parameter includes, for example, the ratio of the observational time for the two kinds of events and the ratio of their collecting areas modified by corresponding experimental efficiencies. Its value is assumed to be known from the experimental details. It can be estimated from additional measurements or extracted from a model of the detection. Relying upon that, the unknown mean of background counts in the on-source region is $\mu_{\mathrm{b}} = \alpha\mu_{\mathrm{off}}$.

In our treatment, the on–off problem consists in the assessment of the relationship between the two unknown on- and off-source means, $\mu_{\mathrm{on}}$ and $\mu_{\mathrm{off}}$. To solve this task we utilize Bayesian reasoning. It is worth pointing out that we do not use often adopted scheme, whereby the source and background parameters, that are responsible for observed on-source counts, are chosen as the basic independent variables, see e.g. [13–15]. We proceed quite differently. In our concept, the Bayesian inference is applied to improve our knowledge about the observed phenomena without any external assumptions about the relationship of the underlying processes. We do not compare models with and without a source in the on-source zone, as usually proposed, i.e. no hypotheses about the source presence are tested nor Bayes factors for on-source model selection are examined. In addition to our à priori notion derived from our previous experience or just selected with respect to our ignorance, for example, we use only experimental data in order to assess whether a source may be identified in the on-source region. We also show how new information may be incorporated in our treatment. This scheme is not only backed by a compelling statistical motivation, but also fairly simple to implement, yet sufficiently general. Nonetheless, our results may deviate from the results obtained under the assumptions used in other Bayesian inference methods aiming to analyze the on–off problem [10–15]. Based on that, our findings are to be interpreted differently in some cases.

In the first step of our analysis, we focus on what kind of information about the on- and off-source means can be obtained from the on–off measurements provided that observed counts in both zones follow the Poisson distribution. Since we have no à priori knowledge whether or in what way these means are related, we examine their independent prior distributions. Using the on–off data and our prior information about the on- and off-source means, we derive their marginal posterior distributions. These distributions summarize our state of knowledge and remaining uncertainty about the on-source mean $\mu_{\mathrm{on}}$ and separately about the off-source mean $\mu_{\mathrm{off}}$, given the data. Thus, the probability that the on-source mean acquires a certain value is given without reference to values of the off-source mean, and vice versa.

In the second step, we compare information we have about both inferred means. Using a known on–off parameter $\alpha$, we normalize the off-source mean in order to obtain a parameter that corresponds to the on-source exposure, i.e. we construct the marginal distribution for the parameter $\mu_{\mathrm{b}} = \alpha\mu_{\mathrm{off}}$. Then, we determine which of the observed on–off processes is more substantial without any assumptions about the relationship of the underlying processes. In order to get the most unbiased value of the source probability, we assume a maximally uninformative joint distribution of the on-source and background means, given the on–off data. For this purpose, we examine the product of their marginal posterior distributions, as dictated by the principle of maximum entropy, and construct the distribution of their difference $\delta = \mu_{\mathrm{on}} - \mu_{\mathrm{b}}$ with a real valued domain. The probability with which $\delta > 0$, as inferred from this distribution, tells us what is the probability that a larger intensity is detected in the on-source region than expected from the off-source measurement.

In more detail, the posterior distribution of the difference $\delta$ allows us to decide whether a source is or is not present in the on-source zone. The presence of a source in the on-source region is validated if the on–off data prefers $\delta > 0$ at a given level of significance. On the other hand, if the data indicates that $\delta \le 0$ at a chosen level of significance, we state that a source is not present in the on-source region. Instead, we infer that the data suggests more activity observed in the off-source region than in the on-source zone.

Although, in the Bayesian context, the assumption of a non-negative source rate ($\delta \ge 0$) is often taken into account by noting that its negative values are unphysical [10–15], we initially do not require that a source is present in the on-source region. In this step, our choice of the on- and off-source region is regarded as purely formal considering the fact that it is not clear in advance what the data will reveal. This feature brings our analysis closer to widely used classical methods that strive for knowledge about the source activity without constraints on the properties of the underlying processes and benefit just from the maximum likelihood estimates of relevant parameters, see e.g. Refs. [2–7]. In this sense, we adopt more precise information about the same parameters including their uncertainties. This information is contained in their marginal posterior distributions obtained with the help of the on–off data using the Bayesian reasoning. We end up with simple expressions that, in agreement with our initial knowledge and without any additional assumptions, describe what can be learned about the source presence in the on–off experiment.

In the final step of our analysis, we focus on the important case when it is known before the measurement is carried out that a source may be present only in the on-source region. Under this condition, our initial ignorance about the relationship between the on- and off-source processes is updated. With the prior requirement that only the joint distribution of the on-source and background means satisfying $\mu_{\mathrm{on}} \ge \mu_{\mathrm{b}} (\delta \ge 0)$ is admissible, the dependence

[1] Throughout this study, we use the same symbols for random variables and their sample values.

between on- and off-source processes reappears in the posterior probability distribution of the non-negative difference. Using this distribution, we finally obtain credible intervals or upper bounds of the intensity of a possible source that is expected in the on-source zone. By their construction, these limits are to be non-negative. In special cases, we obtain the posterior distributions of the non-negative difference that are in agreement with the distributions of the source rate that have been derived using a joint prior distribution with dependent on- and off-source parameters ($\mu_{\rm on} \geq \alpha\mu_{\rm off}, \mu_{\rm off} > 0$) within different Bayesian approaches [8–11,13,14].

### 2.1. On–off means

We consider that $n_{\rm on}$ and $n_{\rm off}$ counts were registered independently in the on- and off-source regions, respectively. We treat the on- and off-source data separately on an equal footing and construct the marginal posterior distributions of the on- and off-source means $\mu_{\rm on}$ and $\mu_{\rm off}$. This way, information about the on-source mean contained in its posterior distribution is given without reference to what is known about the off-source mean, and vice versa.

For both these random variables, we adopt a sufficiently large family of conjugate prior distributions for the Poisson likelihood function. Specifically, we assume that the prior probability distribution of the on-source mean is $p(\mu_{\rm on}) = f_{\rm Ga}(\mu_{\rm on}|s_p, \gamma_p - 1)$ and, in a similar way, the prior distribution of the off-source mean is $p(\mu_{\rm off}) = f_{\rm Ga}(\mu_{\rm off}|s_q, \gamma_q - 1)$, where the Gamma distribution $f_{\rm Ga} = f_{\rm Ga}(\mu|s, \lambda)$ is introduced in Appendix A. The prior shape parameters, $s_p > 0$ and $s_q > 0$, and the prior rate parameters, $\gamma_p > 1$ and $\gamma_q > 1$, characterize our information about the on- and off-source zones before the measurement began.

Here we allow that the prior parameters for the on- and off-source means can acquire different values. This freedom is due to the fact that we can have in principle different initial knowledge about the on- and off-source zone. Such informative prior distributions express our specific knowledge of the examined parameters that may be taken from other experiments or from theoretical considerations, for example. On the other hand, when no such input information is available, the use of uninformative prior distributions (small values of the prior parameters, e.g. $0 < s \leq 1$ and $\gamma \to 1$) typically yields results which are not too different from the results of conventional statistical analysis.

The posterior distributions of the means $\mu_{\rm on}$ and $\mu_{\rm b} = \alpha\mu_{\rm off}$, given the on–off data, $n_{\rm on}$ and $n_{\rm off}$, then take again the form of the Gamma distribution, see Appendix A. In particular, we have $\mu_{\rm on} \sim {\rm Ga}(p, \gamma_p)$ for the posterior distribution of $\mu_{\rm on}$ and $\mu_{\rm b} \sim {\rm Ga}\left(q, \frac{\gamma_q}{\alpha}\right)$ for the posterior distribution of $\mu_{\rm b}$. Here, the shape parameters, $p = n_{\rm on} + s_p$ and $q = n_{\rm off} + s_q$, include, except of the prior input ($s_p$ or $s_q$), also the information acquired from the on–off measurement ($n_{\rm on}$ or $n_{\rm off}$). The rate parameters of the posterior distributions of the on- and off-source means are given by the prior rates $\gamma_p$ and $\gamma_q$, respectively. The rate parameter of the posterior distribution of the background mean $\mu_{\rm b}$ is modified according to the exposures of the on- and off-source zones as expressed by the on–off parameter $\alpha$.

### 2.2. Difference of on–off means

In the context of a single on–off measurement, we address a question of what we are able to learn about the relationship of the underlying Poisson processes that generate the observed on- and off-source counts. Since there is no other relevant information, we assume that the joint probability distributions of the involved on-source and background means is given by the product of their marginal posterior distributions, as it results from maximizing missing information.

With the marginal posterior distributions of the on-source and background means, $\mu_{\rm on}$ and $\mu_{\rm b} = \alpha\mu_{\rm off}$, derived from the on–off data (see Section 2.1), we arrive at the first important result of our study. Under the transformation $\delta = \mu_{\rm on} - \mu_{\rm b}$ while keeping $\mu_{\rm b}$ unchanged, with the Jacobian $J = 1$, and then marginalizing over $\mu_{\rm b}$, we obtain after the standard calculations the probability distribution of the difference of these two unknown means (to simplify the notation we denote $f_\delta(x) = f_\delta(\delta = x|n_{\rm on}, n_{\rm off}, I)$ where $I = (s_p, s_q, \gamma_p, \gamma_q, \alpha)$ stands for prior information)[2]

$$f_\delta(x) = \frac{\gamma_p^p \left(\frac{\gamma_q}{\alpha}\right)^q}{\Gamma(p)}\, e^{-\gamma_p x} x^{p+q-1}\, U(q, p+q, \eta x), \quad x \geq 0 \tag{1}$$

$$f_\delta(x) = \frac{\gamma_p^p \left(\frac{\gamma_q}{\alpha}\right)^q}{\Gamma(q)}\, e^{\frac{\gamma_q}{\alpha} x} (-x)^{p+q-1}\, U(p, p+q, -\eta x), \quad x < 0. \tag{2}$$

Here $p = n_{\rm on} + s_p$ and $q = n_{\rm off} + s_q$ where $s_p > 0$, $s_q > 0$, $\gamma_p > 1$ and $\gamma_q > 1$ are the prior parameters, $\eta = \gamma_p + \frac{\gamma_q}{\alpha}$, $\Gamma(a) = \int_0^\infty e^{-t} t^{a-1}\, {\rm d}t$ stands for the Gamma function and

$$U(a, b, z) = \frac{1}{\Gamma(a)} \int_0^\infty e^{-zt} t^{a-1} (1+t)^{b-a-1}\, {\rm d}t \tag{3}$$

is the integral representation of the Tricomi confluent hypergeometric function [16]. The probability distribution written in Eqs. (1) and (2) is our full inference about the difference of the two unknown means $\mu_{\rm on}$ and $\mu_{\rm b} = \alpha\mu_{\rm off}$ given the on–off data. This solution is maximally noncommittal with respect to unavailable information about the relationship between these means. Note that, by definition, the domain of the new random variable $\delta = \mu_{\rm on} - \mu_{\rm b}$ is not limited and this difference may take all real values.

In practical applications, the integrals in Eq. (3) can be calculated numerically. The saddle point approximation can be used with a good precision if the parameters $p > 1$ and $q > 1$. Analytic expressions can be obtained when selecting particular parameters of the prior distributions.

In some cases, it may be preferred to work with integer values of the parameters $p$ and $q$. Then, the Tricomi confluent hypergeometric function in Eq. (3) may be after some calculations expressed as a finite series ($a, b \in N$)

$$U(a, b, z) = z^{1-b}\, (b-a-1)!\, Q(a, b, z) \tag{4}$$

where

$$Q(a, b, z) = \sum_{i=0}^{b-a-1} \binom{b-i-2}{b-a-i-1} \frac{z^i}{i!}. \tag{5}$$

Straightforward calculations then give ($p, q \in N$)

$$f_\delta(x) = \frac{\gamma_p^p \left(\frac{\gamma_q}{\alpha}\right)^q}{\eta^{p+q-1}}\, e^{-\gamma_p x}\, Q(q, p+q, \eta x), \quad x \geq 0 \tag{6}$$

$$f_\delta(x) = \frac{\gamma_p^p \left(\frac{\gamma_q}{\alpha}\right)^q}{\eta^{p+q-1}}\, e^{\frac{\gamma_q}{\alpha} x}\, Q(p, p+q, -\eta x), \quad x < 0. \tag{7}$$

Special examples are $s_p = s_q = 1$ or a limiting case when $s_p = s_q \to 0$ (for $n_{\rm on} > 0$ and $n_{\rm off} > 0$).

Notice that, if $\gamma_p = \gamma_q$, any probabilistic conclusion based on the distribution of the difference is independent of the choice of the common prior rate. This property is confirmed when integrating the distribution for the difference over an arbitrary interval. Indeed, after a suitable transformation of variables it turns out that

[2] In the following, the explanatory variable $x$ in the probability distribution denotes the values which the corresponding random variable may acquire, in the sense that, for example, the probability $P(x < \delta \leq x + {\rm d}x) = f_\delta(x)\, {\rm d}x$.

the result of integration, and the cumulative distribution function in particular, depends only on the ratio $\frac{\gamma_p}{\gamma_q}$.

It is also worth mentioning the following property of the difference of the on-source and background parameters. Let us define a new random variable $\delta' = -\delta/\alpha$. Then, it is easy to show that its distribution function, $g_{\delta'}(x)$, satisfies $g_{\delta'}(x) = f_{\delta}{}'(x)$ where $f_{\delta}{}'(x)$ denotes the distribution function of the difference $\delta = \mu_{\rm on} - \alpha\mu_{\rm off}$, as given in Eqs. (1) and (2), that is obtained under the transformation $(p, q, \gamma_p, \gamma_q, \alpha) \rightarrow \left(q, p, \gamma_q, \gamma_p, \frac{1}{\alpha}\right)$. Stated differently, when the on- and off-source regions are exchanged, i.e. $(n_{\rm on}, n_{\rm off}, \alpha) \rightarrow (n_{\rm off}, n_{\rm on}, \frac{1}{\alpha})$, and, accordingly, prior information is exchanged, $(s_p, s_q, \gamma_p, \gamma_q) \rightarrow (s_q, s_p, \gamma_q, \gamma_p)$, then the resulting distribution function describes the difference $\delta' = \mu_{\rm off} - \mu_{\rm on}/\alpha$. Thus, any imbalance between the involved regions leads to the same statistical conclusion irrespective what is the reference region. Any excess of counts in one of these zones that suggests the source presence therein is equivalently described as an unknown process that reduces the number of events in the complementary region.

From this point of view, it is worth bearing in mind that other classical test statistics possess the same property. For example, the asymptotic Li–Ma significance, see Eq. (17) in Ref. [2], is in this sense invariant under the transformation $(n_{\rm on}, n_{\rm off}, \alpha) \rightarrow (n_{\rm off}, n_{\rm on}, \frac{1}{\alpha})$. In the binomial treatment [3], the binomial $p$-value for a deficit of counts in the new on-source region is equal to the $p$-value for an excess of counts in the original on-source zone. Also the asymptotic binomial formula for the source detection, see e.g. Eq. (9) in Ref. [2], has the same characteristics. In a similar manner, when the on- and off-source regions are exchanged, it is easy to show that the transformed profile likelihood ratio, see e.g. Ref. [7], provides asymptotic confidence intervals for $\delta' = -\delta/\alpha = \mu_{\rm off} - \mu_{\rm on}/\alpha$.

### 2.3. Source detection

With the posterior probability distribution of the difference we compare the involved on-source and background means. The Bayesian probability that the source is not present in the on-source region corresponds to the non-positive difference of the on-source and background means. It is obtained by integrating the probability distribution of the difference $\delta$ given in Eq. (2) for $\mu_{\rm on} \le \mu_{\rm b}$, i.e. $\delta \le 0$. After straightforward calculations we get the second important result of this study. The Bayesian probability of the absence of a source in the on-source region takes a simple form

$$P^- = P(\delta \le 0) = \int_{-\infty}^{0} f_\delta(x)\, {\rm d}x = I_{\frac{\rho}{1+\rho}}(p, q) \tag{8}$$

where $\rho = \frac{\alpha\gamma_p}{\gamma_q}$, $I_x(a, b)$ denotes the regularized incomplete Beta function that is determined by $B(a, b) I_x(a, b) = B_x(a, b)$ where $B_x(a, b) = \int_0^x t^{a-1}(1-t)^{b-1}\, {\rm d}t$ is the incomplete Beta function and $B(a, b) = B_1(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ denotes the Beta function [16]. Obviously, the source is observed in the on-source region with the Bayesian probability

$$P^+ = P(\delta > 0) = 1 - P^- = I_{\frac{1}{1+\rho}}(q, p). \tag{9}$$

For practical reasons, we also define the Bayesian significance by $S_{\rm B} = \Phi^{-1}(P^+)$ where $\Phi = \Phi(x)$ is the cumulative standard normal distribution. This significance corresponds to the number of standard deviations from a hypothesized value in a classical one-tailed test with a normal distributed variable [3]. We use notation in which a negative value of this significance indicates that the absence of a source in the on-source region is more likely than its presence therein, i.e. if $P^- > 0.5$.

The result written in Eq. (9) represents the Bayesian probability of the source hypothesis, given the on–off data and our prior knowledge of the underlying processes. It allows us to assess the extent to which the processed data is indicative of the source of events. This approach differs from the classical concept designed to measure the exceptional nature of the on–off data with respect to the background model. Our determination of the probability of the source model also differs from the Bayesian strategy based on the initial premise of the non-negative source rate ($\mu_{\rm on} \ge \mu_{\rm b}$), see our discussion on special cases in Sections 2.4.2 and 2.4.3. In a sense, by using a wider range of alternative models ($\mu_{\rm on} > 0$ and $\mu_{\rm off} = \alpha\mu_{\rm b} > 0$), our approach can yield more robust information.

The interpretation of the probability of the source presence in the on-source zone is valid only if the on–off experiment is well designed in the sense that only background counts are recorded in the off-source zone, the corresponding background applies in the on-source region where an extra source may be present. Nonetheless, if it is not the case and, for example, an unknown source is present in the off-source zone or the on-source region is shielded due to an unknown process, the resultant probabilities apply as well, but they should be assigned different meanings. Naturally, the above mentioned options cannot be distinguished in a statistical evaluation.

The Bayesian probabilities of the source absence or presence in the on-source region do not depend on the prior rate parameters if $\gamma_p = \gamma_q$ implying $\rho = \alpha$. In such a case, if the parameters $p > 0$ and $q > 0$ acquire integer values, the result in Eq. (8) can be rephrased using the representation of the binomial distribution. Since the probability $P(N \le q-1) = I_{\frac{\alpha}{1+\alpha}}(p, q)$ [16] where $N$ is a binomial random variable with parameters $p+q-1$ and $\frac{1}{1+\alpha}$, i.e. $N \sim {\rm Bi}\left(p+q-1, \frac{1}{1+\alpha}\right)$, we have $(p, q \in N)$

$$P^- = \sum_{i=0}^{q-1} \binom{p+q-1}{i} \left(\frac{1}{1+\alpha}\right)^i \left(\frac{\alpha}{1+\alpha}\right)^{p+q-i-1}. \tag{10}$$

It gives the probability that less than $q$ events out of $p+q-1$ events are registered in the off-source region or, alternatively, $p$ or more events out of $p+q-1$ events are detected in the on-source region, if the null background hypothesis is true, i.e. $\mu_{\rm on} = \mu_{\rm b}$.

Alternatively, when the parameters $p > 0$ and $q > 0$ are integers and $\gamma_p = \gamma_q$, it also holds that the probability $P(N \le q-1) = I_{\frac{\alpha}{1+\alpha}}(p, q)$ where $N$ is a negative binomial random variable with parameters $p$ and $\frac{1}{1+\alpha}$, i.e. $N \sim N{\rm Bi}\left(p, \frac{1}{1+\alpha}\right)$. Then, one easily recovers that the probability of the absence of a source in the on-source region is $(p, q \in N)$

$$P^- = \sum_{i=0}^{q-1} \binom{p+i-1}{i} \left(\frac{1}{1+\alpha}\right)^i \left(\frac{\alpha}{1+\alpha}\right)^p. \tag{11}$$

This probability describes that less than $q$ events are registered in the off-source region before the chosen number of $p$ events is detected in the on-source region, if the null hypothesis stating that no source is present in the on-source region is true.

The above mentioned results written in Eqs. (10) and (11) hold, for example, for the uniform prior distributions of the on- and off-source means when the prior shape parameters $s_p = s_q = 1$ or for the scale invariant prior distributions when $s_p = s_q \rightarrow 0$ (for $n_{\rm on} > 0, n_{\rm off} > 0$), while $\gamma_p = \gamma_q \rightarrow 1$. In both these cases, the Bayesian probability of no source in the on-source region is similar to the classical probability to reject the background hypothesis, if it is true, in favor of an excess of the on-source events (excess $p$-value). Note that this $p$-value follows from the classical test of the ratio of two unknown Poisson means [3].

Interestingly, assuming $\gamma_p = \gamma_q \to 1$, the probability of the source absence in the on-source zone derived with the uniform priors ($s_p = s_q = 1$) is higher than the corresponding probability derived with the scale invariant priors ($s_p = s_q \to 0$ for $n_{\rm on} > 0, n_{\rm off} > 0$), i.e. $P^-(s_p = s_q = 1) > P^-(s_p = s_q \to 0)$, only if $\alpha n_{\rm on} > n_{\rm off}$, and vice versa. This result is easily obtained by combining the recurrence relations for the incomplete Beta function [16].

### 2.4. Known source

An important case occurs if it is guaranteed with certainty that a source may be observed only in the on-source region. Then, the mean event rate in the on-source zone can only increase beyond what is expected from background. Such a situation is encountered when the ability of the source to produce detectable events has been confirmed in previous analyses or deduced from theoretical considerations, for example. In our concept, the additional knowledge about the source, thought of as a new piece of prior information, is easily incorporated into the Bayesian inference by conditioning on the source rate. This modification allows us to describe the properties of the predefined source, thus also providing us with information related to its detection.

Assuming that the on-source mean is not less than the background one, we are now dealing the case when the processes generating observed counts in both zones are not independent. For this purpose, we consider the joint prior of both means that is written in a separable form and supplemented with the condition $\mu_{\rm on} \ge \mu_{\rm b} = \alpha\mu_{\rm off}$, i.e. $\delta \ge 0$. Under this condition, the posterior probability distribution of the non-negative difference is easily determined by using the results written in Eqs. (1) and (2). The resultant distribution allows us to deduce a credible interval or an upper bound of the source intensity, while there are no problems with negative limits. It is worth emphasizing that this fairly simple construction of limits is equivalent to the analysis scheme in which the model with non-negative source intensity ($\mu_{\rm on} \ge \alpha\mu_{\rm off}$) is examined. Therefore, using special kinds of the prior distributions, we arrive to the posterior distributions of the non-negative difference which agree with the corresponding posterior distributions of the source intensity obtained in other Bayesian approaches, see Sections 2.4.1, 2.4.2, 2.4.3 and 2.4.4.

When one is concerned with the non-negative source rate, the corresponding probability distribution is derived under the condition of non-negative values of the difference of the on-source and background means, i.e. $\mu_{\rm on} \ge \mu_{\rm b} = \alpha\mu_{\rm off}$ implying $\delta \ge 0$. The distribution of the non-negative difference then follows from Eq. (1). Another important result of our analysis that includes several previously derived results [8–11,13,14] is

$$f_\delta^+(x) = \frac{\gamma_p^p \left(\frac{\gamma_q}{\alpha}\right)^q \Gamma(q)}{\Gamma(p+q) B_{\frac{1}{1+\rho}}(q,p)} e^{-\gamma_p x} x^{p+q-1}\ U(q, p+q, \eta x), \quad x \ge 0 \qquad (12)$$

where we introduced the conditional distribution $f_\delta^+(x) = \frac{f_\delta(x)}{P^+}$ for $x \ge 0$, $P^+ = 1 - P^-$ is the Bayesian probability that the source is present in the on-source region, as given in Eq. (9), $\eta = \gamma_p + \frac{\gamma_q}{\alpha}$ and $\rho = \frac{\alpha\gamma_p}{\gamma_q}$.

In particular, if the parameters $p > 0$ and $q > 0$ acquire integer values, the probability distribution of the non-negative difference of the on-source and background means is obtained from Eq. (6) ($p, q \in N$)

$$f_\delta^+(x) = \frac{f_\delta(x)}{P^+} = e^{-\gamma_p x} \frac{Q(q, p+q, \eta x)}{Q_{\gamma_p}^+(q, p+q, \eta)}, \quad x \ge 0 \qquad (13)$$

where the function $Q(a,b,z)$ is given in Eq. (5) and $P^+$ is the probability that the source is present in the on-source region written

$$P^+ = \frac{\gamma_p^p \left(\frac{\gamma_q}{\alpha}\right)^q}{\eta^{p+q-1}} Q_{\gamma_p}^+(q, p+q, \eta) \qquad (14)$$

where ($a, b \in N$)

$$Q_{\gamma_p}^+(a,b,z) = \int_0^\infty Q(a,b,zx) e^{-\gamma_p x}\, dx = \sum_{i=0}^{b-a-1} \binom{b-i-2}{b-a-i-1} \frac{z^i}{\gamma_p^{i+1}}. \qquad (15)$$

#### 2.4.1. Scale invariant priors

The scale invariant prior for a non-negative random variable corresponds to a uniform prior of its logarithm. In our treatment, such prior distributions of the means $\mu_{\rm on}$ and $\mu_{\rm off}$ can be selected only if the numbers of detected on- and off-source events are positive. These prior distributions are classified by the rate parameters $\gamma_p = \gamma_q \to 1$, i.e. $\eta = \frac{1+\alpha}{\alpha}$ and $\rho = \alpha$, the shape parameters $s_p = s_q \to 0$, i.e. $p = n_{\rm on} > 0$ and $q = n_{\rm off} > 0$. Hence, for the posterior distributions we have $\mu_{\rm on} \sim {\rm Ga}(n_{\rm on}, 1)$ and $\mu_{\rm b} \sim {\rm Ga}(n_{\rm off}, \frac{1}{\alpha})$, see Appendix A. Then it follows from Eq. (13) that the non-negative difference of the on-source and background means, $\mu_{\rm on} \ge \mu_{\rm b}$, is

$$f_\delta^+(x) = e^{-x} \frac{\sum_{i=0}^{n_{\rm on}-1} \binom{n_{\rm on}+n_{\rm off}-i-2}{n_{\rm on}-i-1} \left(\frac{1+\alpha}{\alpha}\right)^i \frac{x^i}{i!}}{\sum_{i=0}^{n_{\rm on}-1} \binom{n_{\rm on}+n_{\rm off}-i-2}{n_{\rm on}-i-1} \left(\frac{1+\alpha}{\alpha}\right)^i}, \quad x \ge 0. \qquad (16)$$

The same result was presented in Ref. [11].

In our analysis, the Bayesian probability that a source is present in the on-source region is given explicitly by, see Eq. (14),

$$P^+ = \left(\frac{1}{1+\alpha}\right)^{n_{\rm off}} \sum_{i=0}^{n_{\rm on}-1} \binom{n_{\rm on}+n_{\rm off}-i-2}{n_{\rm on}-i-1} \left(\frac{\alpha}{1+\alpha}\right)^{n_{\rm on}-i-1}. \qquad (17)$$

#### 2.4.2. Uniform priors

Let us consider the uniform prior distributions of the means $\mu_{\rm on}$ and $\mu_{\rm off}$. In such a case, the rate parameters $\gamma_p = \gamma_q \to 1$, i.e. $\eta = \frac{1+\alpha}{\alpha}$ and $\rho = \alpha$, the shape parameters $s_p = s_q = 1$, i.e. $p = n_{\rm on} + 1$ and $q = n_{\rm off} + 1$. The posterior distributions are $\mu_{\rm on} \sim {\rm Ga}(n_{\rm on}+1, 1)$ and $\mu_{\rm b} \sim {\rm Ga}(n_{\rm off}+1, \frac{1}{\alpha})$, see Appendix A. Assuming the non-negative difference of the on-source and background means, $\mu_{\rm on} \ge \mu_{\rm b}$, we get from Eq. (13)

$$f_\delta^+(x) = e^{-x} \frac{\sum_{i=0}^{n_{\rm on}} \binom{n_{\rm on}+n_{\rm off}-i}{n_{\rm on}-i} \left(\frac{1+\alpha}{\alpha}\right)^i \frac{x^i}{i!}}{\sum_{i=0}^{n_{\rm on}} \binom{n_{\rm on}+n_{\rm off}-i}{n_{\rm on}-i} \left(\frac{1+\alpha}{\alpha}\right)^i}, \quad x \ge 0. \qquad (18)$$

The same result was obtained in Refs. [10,13].

The Bayesian probability that a source is present in the on-source region follows from Eq. (14), namely,

$$P^+ = \left(\frac{1}{1+\alpha}\right)^{n_{\rm off}+1} \sum_{i=0}^{n_{\rm on}} \binom{n_{\rm on}+n_{\rm off}-i}{n_{\rm on}-i} \left(\frac{\alpha}{1+\alpha}\right)^{n_{\rm on}-i}. \qquad (19)$$

We note that a quite different formula has been advocated in Ref. [13]. Its justification is based on the Bayes factor that accounts for a complex source model put against a simple background hypothesis. However, as pointed out in Ref. [13], the significant disadvantage is that the resultant probability strongly depends on the choice of the upper bound of the uniform prior used for the source activity.

Our Bayesian probabilities for the presence or absence of a source in the on-source region are easily obtained in the case of the null experiment, when no counts are registered in the on-source region, i.e. $n_{\rm on} = 0$ and $p = 1$, or in the experiment with zero background counts, i.e. $n_{\rm off} = 0$ and $q = 1$,

$$P^{+}_{n_{\rm on}=0} = \left(\frac{1}{1+\alpha}\right)^{n_{\rm off}+1}, \quad P^{-}_{n_{\rm off}=0} = \left(\frac{\alpha}{1+\alpha}\right)^{n_{\rm on}+1}. \tag{20}$$

Both these probabilities depend on the relationship between the on- and off-source regions, on the on–off parameter $\alpha$. Unlike other results [13], our approach provides us with well understandable solutions. For example, in the null experiment ($n_{\rm on} = 0$), the Bayesian probability of the source detection in the on-source region drops down with the increasing on-source exposure (increasing $\alpha$) as well as with the increasing number of registered off-source events. If no events are registered at all, we get $P^{+}_{n_{\rm on}=n_{\rm off}=0} = (1+\alpha)^{-1}$.

### 2.4.3. Jeffreys' priors

The key characteristic of the Jeffreys' prior distribution is that it is invariant under a transformation of parameters. Thus, it expresses the same prior belief no matter which metric is used.

In our notation scheme, Jeffreys' prior distributions of the on- and off-source means are of the form introduced in Appendix B. The posterior distributions are formally constructed if the rate and shape parameters of the Gamma distributions given in Appendix A satisfy $s_p = s_q = \frac{1}{2}$ and $\gamma_p = \gamma_q \to 1$, respectively. Then, the distribution of the difference is obtained putting $p = n_{\rm on} + \frac{1}{2}$, $q = n_{\rm off} + \frac{1}{2}$, $\eta = \frac{1+\alpha}{\alpha}$ and $\rho = \alpha$ into the relevant equations.

In particular, the distribution for the non-negative difference written in Eq. (12) implies the recent result based on Jeffreys' rule presented in Ref. [14]. Indeed, using the identities for the hypergeometric functions [16]

$$\frac{a}{x^a} B_x(a,b) = {}_2F_1(a, 1-b, a+1, x) \tag{21}$$

and

$${}_2F_1(a,b,c,x) = (1-x)^a {}_2F_1\left(a, c-b, c, \frac{x}{x-1}\right) \tag{22}$$

one has, in our notation scheme,

$$q\alpha^q B_{\frac{1}{1+\alpha}}(q,p) = \left(\frac{\alpha}{1+\alpha}\right)^q {}_2F_1\left(q, 1-p, q+1, \frac{1}{1+\alpha}\right) = {}_2F_1\left(q, p+q, q+1, -\frac{1}{\alpha}\right). \tag{23}$$

Substituting this result into Eq. (12) and decoding the values of the parameters $p$, $q$ and $\eta$, while $\gamma_p = \gamma_q \to 1$, the correspondence with the result written in Eq. (30) in Ref. [14] is evident.

This consistency is due to the above mentioned invariant property of the Jeffreys' prior. In our strategy, we started with the two independent variables $\mu_{\rm on} > 0$ and $\mu_{\rm off} > 0$ the prior distributions of which are given by Jeffreys' rule, see Eq. (B.1) in Appendix B. Choosing a new mean $\mu_{\rm s} = \mu_{\rm on} - \alpha\mu_{\rm off} \geq 0$ and keeping $\mu_{\rm off} > 0$ unchanged, the bidimensional prior distribution considered in Eq. (15) in Ref. [14] is easily obtained under this transformation.

It is worth stressing, however, that the probability of the absence of a source in the on-source zone that was derived in this study using the distribution of the difference (see Section 2.3) differs from the results of Refs. [14,15] when Jeffreys' rule for prior distributions is considered. The reason is that the other methods do not benefit from all input information or do not fully utilize the Bayesian inference.

In Ref. [14], the determination of the Bayesian probability of the background hypothesis was based on the questionable argument about how to choose the ratio of the arbitrary scale factors of the prior distributions of model parameters. This ratio was derived following the *ad hoc* assumption that if no counts are observed in both zones, the probabilities of both the signal and background model remain the same. However, one may successfully argue that such a null measurement with no background counts ($n_{\rm on} = n_{\rm off} = 0$) should update our knowledge about the signal. The point is that one has additional information since the ratio of the on- and off-source exposures is known by definition. Therefore, the result of the null experiment with no background counts is to prefer the signal alternative if $0 < \alpha < 1$ (larger off-source exposure) and vice versa. Unfortunately, the premise behind the procedure that provides the probability of the background hypothesis, as advocated in Ref. [14], does not take into account the possibility of different exposures. Interestingly, while the probability of the no-source hypothesis $H_0$ is assumed to be $P(H_0|n_{\rm on} = n_{\rm off} = 0) = \frac{1}{2}$ in Ref. [14], we obtain from Eq. (8) for the Bayesian probability of the absence of a source in the on-source region a more intuitively appealing result

$$P^{-}_{n_{\rm on}=n_{\rm off}=0} = I_{\frac{\alpha}{1+\alpha}}\left(\frac{1}{2}, \frac{1}{2}\right) = \frac{2}{\pi}\arctan(\sqrt{\alpha}). \tag{24}$$

With the increasing on-source exposure (increasing $\alpha$), the probability that a source is not in the on-source region increases if no counts ($n_{\rm on} = n_{\rm off} = 0$) are detected in both on–off zones, for the uniform priors see Eq. (20).

In Ref. [15], a predictive distribution of background counts was utilized in order to assess to what extent the source model is not supported by the on–off data. Following Jeffreys' rule, the distribution for the background mean was modeled as the Gamma distribution with parameters deduced from the off-source observation using the method of moments. The significance of the signal deviation from the background hypothesis was established based on the Poisson–Gamma mixture. In this approach, the on- and off-source zones are treated differently. The resultant $p$-values are to be interpreted as the probability of obtaining a result at least as extreme as the observed data if the null background hypothesis is true. Thus, such an approach does not fully exploit the Bayesian reasoning and, therefore, it cannot provide us with information what hypothesis is more likely, given the data.

### 2.4.4. Known background

The analysis may be adapted for the case of known background with remaining uncertainty in the on-source zone, for classical results see e.g. Ref. [5]. Let us assume that the background mean $\mu_{\rm b}$ is known, but we do not measure the counts due to the background during the experiment. Such a situation may be reviewed as the limit $q \to \infty (q = n_{\rm off} + s_q)$, $\alpha \to 0$ when $q\alpha = \mu_{\rm b}$ remains a finite constant [10]. In our scheme, the difference of the two Poisson parameters enlarged by the constant background parameter follows the Gamma distribution, i.e. $\mu_{\rm on} = (\delta + \mu_{\rm b}) \sim {\rm Ga}(p, \gamma_p)$ where $p = n_{\rm on} + s_p$ and $\gamma_p > 1$ are parameters for the prior distribution of the on-source mean. Therefore, the probability distribution of the difference $\delta$ is then given by (here we have $h_\delta(x) = h_\delta(\delta = x|n_{\rm on}, \mu_{\rm b}, I)$)

$$h_\delta(x) = \frac{\gamma_p^p}{\Gamma(p)}(x+\mu_{\rm b})^{p-1} e^{-\gamma_p(x+\mu_{\rm b})}, \quad x \geq -\mu_{\rm b}. \tag{25}$$

In addition, assuming non-negative values of the difference $\delta$, i.e. $\mu_{\rm on} \geq \mu_{\rm b}$, we have for its probability distribution

$$h^{+}_\delta(x) = \frac{h_\delta(x)}{R^{+}} = \frac{\gamma_p^p}{\Gamma(p,\mu_{\rm b})}(x+\mu_{\rm b})^{p-1} e^{-\gamma_p(x+\mu_{\rm b})}, \quad x \geq 0 \tag{26}$$

where $\Gamma(a,x) = \int_x^\infty t^{a-1} e^{-t}\,{\rm d}t$ is the upper incomplete Gamma function and

$$R^{+} = P(\delta > 0) = \int_0^\infty h_\delta(x)\,{\rm d}x = \frac{\Gamma(p,\mu_{\rm b})}{\Gamma(p)} \tag{27}$$

is the probability of the presence of a source in the on-source region if the background mean is known.

Note that for the uniform prior distribution of the on-source parameter, when $p = n_{\rm on} + 1 (s_p = 1)$ and $\gamma_p \to 1$, the result written in Eq. (26) was obtained in Ref. [9]. More general expressions with the prior parameter $\gamma_p \to 1$ were presented in Ref. [10].

### 2.5. Source intensity

With the complete information about the on–off measurement contained in the distribution of the difference $\delta$, we can estimate the source intensity. We use the shortest credible interval $\langle \delta_-, \delta_+ \rangle$ that includes the source intensity at a chosen significance level of $P$. In order to obtain these intervals, one has to solve numerically

$$P = \int_{\delta_-}^{\delta_+} f_\delta(x)\,\mathrm{d}x, \quad f_\delta(\delta_-) = f_\delta(\delta_+) \tag{28}$$

with the indicated condition on interval endpoints, if it can be fulfilled.

In those cases when the lower endpoint of a credible interval is negative, an upper bound for the source intensity is usually required. Its value, $\delta_+$, is determined numerically using the integral in Eq. (28) where we put $\delta_- \to -\infty$ and relax the constraint on interval endpoints.

When it is known that a source may be present only in the on-source zone (see Section 2.4), we obtain credible intervals for the source intensity by putting $f_\delta(x) \to f_\delta^+(x)$ with $0 \le \delta_-$ into Eq. (28). For upper bounds we set $\delta_- = 0$ while not using the constraint on interval endpoints.

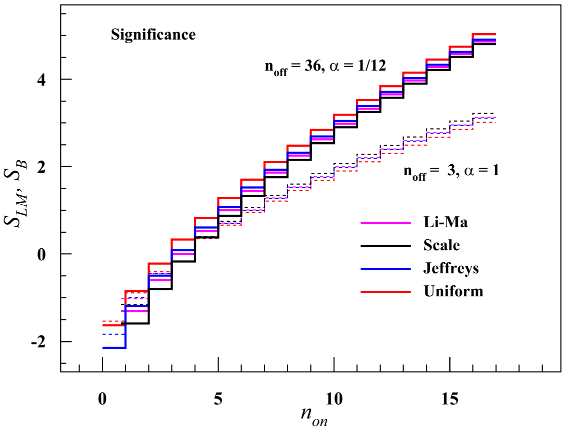


**Fig. 1.** Significances for the source detection are shown as functions of the number of events detected in the on-source zone for two different off-source measurements. In the first case (thick lines), $n_{\mathrm{off}} = 36$ and $\alpha = \frac{1}{12}$. In the second case (thin lines), $n_{\mathrm{off}} = 3$ and $\alpha = 1$. The Li–Ma significances are shown in magenta. The Bayes significances for scale invariant (black lines), Jeffreys' (blue lines) and uniform (red lines) prior distributions were derived from the probability of the source presence in the on-source region, see Eq. (9). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

## 3. Examples

### 3.1. Source detection significance

We present examples which illustrate some of the features of the method described in Section 2.3 that allows us to assign the probability of the source absence or presence in the on-source zone for a pair of on–off measurements. We focused on the cases with small numbers of events. For this purpose, we use the significance derived from the Bayesian probability of the source presence in the on-source region using the standard normal variate, see Section 2.3. In each example we calculated Bayesian significances using the scale invariant, Jeffreys' as well as uniform prior distributions of the on- and off-source means ($\gamma = \gamma_p = \gamma_q \to 1$, $s = s_p = s_q$, $s \to 0$ or $s = \frac{1}{2}, 1$). We also calculated the asymptotic Li–Ma significance [2] with which, relying on the likelihood ratio method, the no-source hypothesis is rejected if it is true. We added a sign to the Li–Ma statistic $S_{\mathrm{LM}}$ considering it as non-negative if $n_{\mathrm{on}} - \alpha n_{\mathrm{off}} \ge 0$ and negative otherwise, i.e. $S_{\mathrm{LM}} = \mathrm{sgn}(n_{\mathrm{on}} - \alpha n_{\mathrm{off}})\sqrt{S^2_{\mathrm{LM}}}$, since the original statistic [2] is equivalent to the absolute value of a standard normal variable.

In the first example, we chose the number of events detected in the off-source zone while varying the number of registered on-source counts. We dealt with two cases. In the first case, we assumed that $n_{\mathrm{off}} = 36$ counts were detected in the off-source region the exposure of which is 12-times larger than the exposure of the on-source zone, i.e. $\alpha = \frac{1}{12}$. In the second case, we chose the same exposures of the on- and off-source regions ($\alpha = 1$), and assumed that $n_{\mathrm{off}} = 3$ events were registered in the off-source zone. The numbers of on-source events were small, $n_{\mathrm{on}} \in \langle 0, 16 \rangle$. Note that for $n_{\mathrm{on}} = 0$ the scale invariant and Li–Ma significances are not determined. In Fig. 1, our results obtained within the Bayesian inference are compared with the asymptotic Li–Ma significances [2]. Obviously, better knowledge about background $(n_{\mathrm{off}} = 36, \alpha = \frac{1}{12})$ implies higher absolute values of significances. Note that in this case $(n_{\mathrm{off}} = 36, \alpha = \frac{1}{12})$, the Bayesian significances based on the uniform prior distributions are larger when compared with the scale invariant results since $\alpha n_{\mathrm{on}} < n_{\mathrm{off}} = 36$, see Section 2.3. The opposite is true in the second case ($n_{\mathrm{off}} = 3$, $\alpha = 1$) only when $n_{\mathrm{on}} > n_{\mathrm{off}} = 3$. The Bayesian significances based on Jeffreys' prior distributions always lie in between results derived assuming the uniform and scale invariant prior distributions, if the latter choice is possible ($n_{\mathrm{on}} > 0$ and $n_{\mathrm{off}} > 0$).

In the second example, using the Monte Carlo technique, we focused on the distributions of significances for the source detection. We generated $10^5$ pairs of on- and off-source counts that follow the Poisson distribution with predefined source and background means, respectively, assuming that they were registered in the regions of the same exposures ($\alpha = 1$). We determined the Bayesian probabilities of the source presence in the on-source region (see Eq. (9)) and the corresponding significances as well as the asymptotic Li–Ma significances [2] for each pair of on- and off-source counts. In Fig. 2, we present the significance distributions

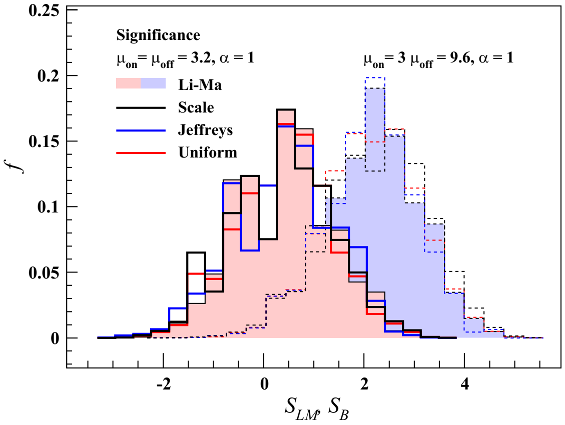


**Fig. 2.** Distributions of significances for the source detection. Histograms for the Li–Ma significances (filled areas) as well as for the Bayes significances using scale invariant (black lines), Jeffreys' (blue lines) and uniform (red lines) prior distributions are visualized. Two examples for the same exposures of the on- and off-source zones ($\alpha = 1$) are presented. In the first example (light red area and thick lines), the on- and off-source counts were generated with the mean parameters $\mu_{\mathrm{on}} = \mu_{\mathrm{off}} = 3.2$. In the second example (light blue area and thin dashed lines), the mean parameters were $\mu_{\mathrm{on}} = 3\mu_{\mathrm{off}} = 9.6$. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

for no source in the on-source zone, when the on- and off-source means $\mu_{\text{on}} = \mu_{\text{off}} = 3.2$ (light red area and thick lines). Typically, in about 10% of on–off pairs, it happened that no events in the on- or off-source zone were generated. This results in dips in scale invariant (thick black line) and Li–Ma (light red area) histograms located around zero significance since, in such cases, the relevant significances cannot be determined. In Fig. 2, we also show significance distributions that we received in the case when the source is present in the on-source region using $\mu_{\text{on}} = 3\mu_{\text{off}} = 9.6$ (light blue and thin dashed lines). In both presented cases, except the problem with zero counts, the significance distributions obtained using the Bayesian inference, with the scale invariant, Jeffreys' or uniform prior distributions, are similar to each other as well as to the corresponding outputs obtained with the help of the asymptotic Li–Ma formula [2].

### 3.2. Gamma-ray bursts

The method described in Section 2 was applied to the data sets examined in Refs. [14,15]. We used information about very high energy (VHE) photons from gamma-ray bursts (GRB) collected by the VERITAS setup [17] and by the Fermi Large Area Telescope [18]. These data sets of VHE photons detected during or shortly after 12 bursts are listed in the first four columns in Table 1. Typically, only a few VHE photons were registered in the directions of GRBs. In most cases, the number of collected events is not too different from the corresponding number of events expected from background.

We assumed the same prior distributions for the on- and off-source means with the common shape parameter, $s = s_p = s_q$, and zero rate parameters, i.e. $\gamma = \gamma_p = \gamma_q \to 1$. With these restrictions we calculated the distributions of the difference of the on-source and background means. The probability that a source is absent in the on-source region is then given by Eq. (8). We also determined credible intervals and, if appropriate, upper bounds of the source intensity at a given level of significance as described in Section 2.5.

The conditional distributions of the difference ($\mu_{\text{on}} \geq \mu_{\text{b}}$) for all data sets are depicted in Fig. 3. These results were obtained with Jeffreys' prior distributions ($s = \frac{1}{2}$ and $\gamma \to 1$). Specific properties of the GRB sources are summarized in Table 1. In this table, we present the Bayesian probabilities of the absence of a source of VHE photons in the on-source zone ($P^-$). Negative values of the corresponding Bayesian significance ($S_{\text{B}}$) indicate that the absence of a source in the on-source region is more likely than its presence therein, i.e. $P^- > 0.5$. For all data sets, we also give credible intervals for the difference of the on-source and background means at a 99% level of confidence. With the distributions of the difference conditioned on the non-negative source intensity we reproduce the upper bounds ($\delta_+^+$) obtained in Ref. [14] at the same level of confidence (see Section 2.4.3).

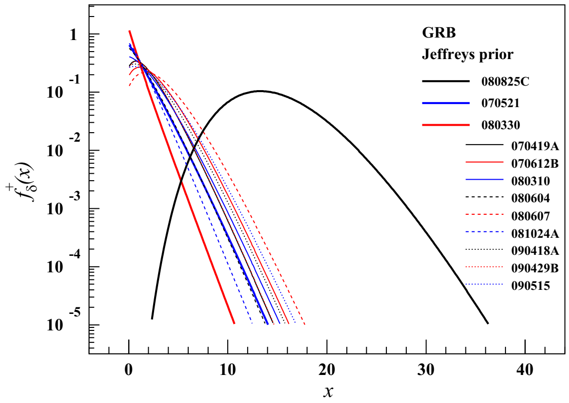


**Fig. 3.** Distributions of the difference conditioned on non-negative values of the on-source rate for 12 data sets connected with GRB observations [17,18]. These results were obtained with Jeffreys' prior distributions for the on- and off-source means ($s = \frac{1}{2}$ and $\gamma \to 1$). Full black, blue and red curves are for GRB 080825C, GRB 070521 and GRB 080330, respectively. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

In Table 1, we also present other on–off results which were obtained within the classical concept and have, therefore, a different meaning. In particular, we calculated the Li–Ma significance [2]. Also confidence intervals derived using the unbounded likelihood method [7] were determined. It is worth stressing that both these classical statistics are obtained asymptotically relying upon the likelihood ratio and the Wilks' theorem [1]. The asymptotic confidence intervals or upper bounds are constructed in such way that they cover an unknown true value of the parameter under consideration with a specified probability. The Li–Ma significance corresponds to the probability with which the null background hypothesis is rejected if it is true. We added a sign to the Li–Ma statistic $S_{\text{LM}}$ in order that $S_{\text{LM}} < 0$ if $n_{\text{on}} - \alpha n_{\text{off}} < 0$.

There is a clear evidence that VHE photons from GRB 080825C were detected by the Fermi-Lat instrument [18]. The significance of the presence of a source in the on-source zone, $S_{\text{B}} = 6.22$, provides the same conclusion as the asymptotic Li–Ma significance, as the original finding [18] and other results [14,15]. Our $1\sigma$ estimate of the source intensity, $\delta \in \langle 9.75, 17.49 \rangle$ obtained with Jeffreys'

**Table 1**
GRB data collected by VERITAS [17], GRB 080825C was observed by Fermi-Lat [18]. The same sets of data as in Refs. [14,15] are used. GRB assignments, measured counts and on–off parameters $\alpha$ are listed in the first four columns. The Bayesian probability of the absence of a source in the on-source zone ($P^-$), corresponding significance ($S_{\text{B}}$) and Li–Ma significance ($S_{\text{LM}}$) are given in the following three columns. We show credible intervals for the difference ($\langle \delta_-, \delta_+ \rangle$) and its upper bounds obtained by assuming that a source may be present only in the on-source zone ($\delta_+^+$), both at a 99% level of confidence. Confidence intervals for the difference ($\langle \delta_-, \delta_+ \rangle_{\text{L}}$) derived in the unbounded profile likelihood method [7] at the same level of confidence are given in the rightmost column. For Bayesian results, Jeffreys' prior distributions ($s = \frac{1}{2}$ and $\gamma \to 1$) were used.

| GRB | $n_{\text{on}}$ | $n_{\text{off}}$ | $\alpha$ | $P^-$ | $S_{\text{B}}$ | $S_{\text{LM}}$ | $\langle \delta_-, \delta_+ \rangle$ | $\delta_+^+$ | $\langle \delta_-, \delta_+ \rangle_{\text{L}}$ |
|---|---|---|---|---|---|---|---|---|---|
| 070419A | 2 | 14 | 0.057 | 0.110 | 1.23 | 1.08 | −1.11, 7.74 | 6.88 | −0.88, 7.34 |
| 070521 | 3 | 113 | 0.057 | 0.923 | −1.43 | −1.48 | −6.86, 2.91 | 6.12 | −6.77, 3.58 |
| 070612B | 3 | 21 | 0.066 | 0.106 | 1.25 | 1.14 | −1.50, 8.61 | 8.00 | −1.23, 8.55 |
| 080310 | 3 | 23 | 0.128 | 0.455 | 0.11 | 0.03 | −3.60, 6.92 | 7.16 | −3.37, 7.08 |
| 080330 | 0 | 15 | 0.123 | 0.932 | −1.49 |  | −3.84, 3.43 | 4.10 | −3.38, 2.40 |
| 080604 | 2 | 40 | 0.063 | 0.591 | −0.23 | −0.33 | −3.03, 5.10 | 6.12 | −2.93, 5.66 |
| 080607 | 4 | 16 | 0.112 | 0.080 | 1.41 | 1.32 | −1.82, 10.12 | 9.17 | −1.42, 9.84 |
| 080825C | 15 | 19 | 0.063 | $710^{-10}$ | 6.22 | 6.36 | 5.05, 26.60 |  | 5.86, 26.12 |
| 081024A | 1 | 7 | 0.142 | 0.441 | 0.15 | 0.01 | −2.13, 5.64 | 5.30 | −1.89, 5.19 |
| 090418A | 3 | 16 | 0.123 | 0.233 | 0.73 | 0.64 | −2.50, 8.24 | 7.64 | −2.17, 8.01 |
| 090429B | 2 | 7 | 0.106 | 0.106 | 1.25 | 1.12 | −1.04, 6.41 | 6.92 | −0.99, 7.41 |
| 090515 | 4 | 24 | 0.126 | 0.282 | 0.58 | 0.50 | −3.25, 8.63 | 8.34 | −2.94, 8.66 |

prior distributions, corresponds to previously presented estimates [14,15,18].

Other data sets of VHE photons collected in the directions of GRBs show no signature that would distinguish them from background data. The Bayesian probabilities of the absence of a source in the selected on-source regions are above 8%, see Table 1. The absolute values of the corresponding significance are below 1.50. Three data sets indicate a deficit of events in the on-source region, i.e. $P^- > 0.5$ and $S_B < 0$.

Of particular interest are the results derived from the data associated with GRB 080330 since no on-source event was recorded in this observation. We recall that the data is easy to evaluate in the Bayesian approach. No special assumptions or external constraints are needed. The only exception is that the choice of the scale invariant priors (s→0) is excluded. Using Jeffreys' prior distributions, our analysis yields the Bayesian probability of the absence of a source in the on-source zone of about 93%, see Table 1. Our upper bound for the source intensity is somewhat higher than the estimate which was obtained by extrapolation within the unbounded likelihood method [7].

With the aim to demonstrate the impact of different prior distributions, we choose the data set of GRB 070521. This data yields the lowest positive ratio of the number of on-source events with respect to the background counts expected in the on-source region. In Fig. 4, various distribution functions of the difference for GRB 070521 are shown. Three types of prior distributions were examined. Namely, we present results based on the scale invariant (s→0, in black), Jeffreys' ($s=\frac{1}{2}$, blue) and uniform ($s=1$, red) prior distributions. The depicted distributions were obtained without (dashed curves) or with (full curves) assuming that a source may be present only in the on-source zone. The former distributions are used in order to determine the probability of the source presence in the on-source zone. The latter conditional distributions are then used for estimating credible intervals or upper bounds of the source intensity.

In this case, and also for other data sets listed in Table 1, we mostly obtained very similar results for the three types of prior distributions used for the on- and off-source means. This situation is documented in Fig. 5 where the Bayesian probabilities for the source absence in the on-source zone are shown as functions of the common shape parameter of the prior distributions. The no-source probabilities mostly slowly decrease with the increasing value of the prior shape parameter, from the scale invariant option (s→0) down to the uniform choice ($s=1$). The largest decline is found for the data associated with the observation of GRB 081024A when the lowest total number of events was detected.

Finally, the 99% upper bounds of the source intensity, and the $1\sigma$ credible interval in case of 080825C, are depicted in Fig. 6 as functions of the common shape parameter of the prior distributions. These characteristics were obtained by assuming that a source may be present only in the on-source region. We learned that the limits of the source intensity are weakly dependent on the prior choice of the common shape parameter for $0 < s \leq 1$.

## 4. Conclusions

In this study we dealt the issue of detection of a source the activity of which is immersed in the surrounding background. For this purpose, we adopted the Bayesian concept that provides, on one side, a unified and intuitively appealing approach to the problem of drawing inferences from observations and, on the other side, it offers a powerful and sufficiently general framework for determining optimal behavior in the face of uncertainty. As often reported, the Bayesian inference also allows us to alleviate some of the issues that affect conventional statistical approaches.

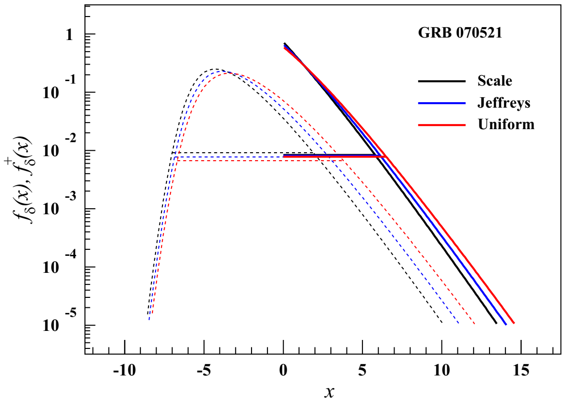


**Fig. 4.** Distributions of the difference using the GRB 070521 data. The distributions shown in black, blue and red were obtained with scale invariant (s→0), Jeffreys' ($s=\frac{1}{2}$) and uniform (s=1) priors, respectively. The thick full curves are for the distributions determined by assuming that a source may be present only in the on-source zone. The dashed lines indicate solutions without this information. Horizontal lines visualize credible intervals at a 99% level of confidence. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

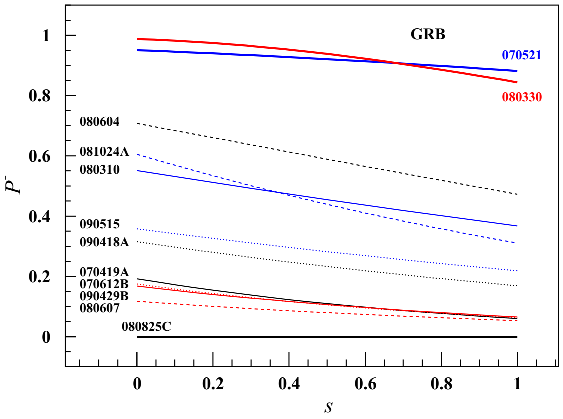


**Fig. 5.** Probabilities of the source absence in the on-source zone are shown as functions of the common shape parameter of prior distributions. The thick full black, blue and red curves are for GRB 080825C, GRB 070521 and GRB 080330, respectively. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

We have proposed a consistent description of the on–off measurement. We focused on cases of small numbers of registered events that obey a Poisson distribution. For the on- and off-source means, we used an adequately large class of conjugate prior distributions for the Poisson likelihood function. It consists of Gamma distributions, each of which is parametrized by two parameters, by the rate and shape parameter. The Gamma family includes several interesting and widely used options, i.e. scale invariant, uniform or Jeffreys' prior distributions.

We examined the distribution of the difference between the on-source and background means. This distribution is maximally noncommittal with regard to their dependence, but it carries all the information available from the on–off experiment. Using it, the probability of the presence of a source in the on-source zone and other source properties are consistently derived within the Bayesian concept and, therefore, have well defined meaning. We stress that our interpretation of the on–off data is different from

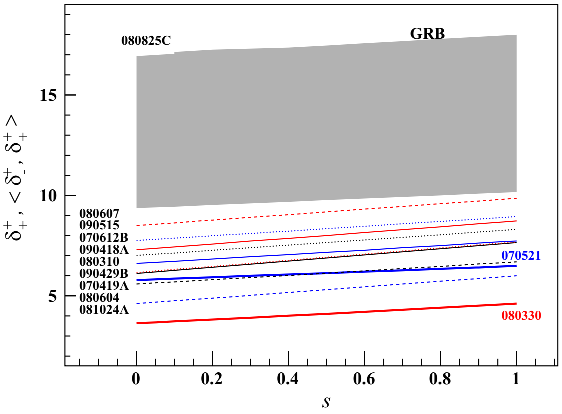


**Fig. 6.** Upper bounds for the difference at a 99% level of confidence are shown as functions of the common shape parameter of prior distributions. They were obtained by assuming that a source may be present only in the on-source zone. The thick full blue and red curves are for GRB 070521 and GRB 080330, respectively. A grey band represents the $1\sigma$ credible interval for the activity associated with GRB 080825C. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

reasoning behind hypothesis testing, regardless of whether the test is conducted in a classical or Bayesian framework.

The distribution of the difference is well suited for weak sources whose observations may reveal a signal either in the on-source or off-source zone, due to experimental limitations, for example. Except one case, the proposed Bayesian solutions can be used for any number of on–off counts, including the null experiment or the experiment with no background. To our knowledge, such results of the Bayesian inference have not yet been discussed in the literature. By conditioning on the values of the difference we obtained a probability distribution that allows us to describe the on–off problem with a preassigned source in the on-source region the activity of which is to be examined. In this case, the resultant conditional distribution includes several results that have previously been obtained in other Bayesian approaches. Using this distribution, well reasoned limits of the source activity are easily determined.

We also presented several numerical examples that may serve as guides for practical applications. In most cases, when little is known about investigated phenomena, it turned out that the scale invariant, if applicable, or uniform prior distributions are good choices. The corresponding formulae reduce to simple algebraic sums, as described in Sections 2.4.1 and 2.4.2 provided that a source may be present only in the on-source zone. The Bayesian inference using Jeffreys' prior distributions should be a better compromise. However, this option, as well as the choice of informative priors, requires more complicated calculations based on integral expressions.

## Acknowledgment

We would like to thank the two unknown reviewers for their constructive criticism and many useful suggestions that helped us with the presentation of our study. This work was supported by the Czech Science Foundation under project 14-17501S.

## Appendix A. Bayesian inference with Poisson likelihood

We consider that the number of events registered in a counting experiment, a random variable $n$, obeys the Poisson distribution with a mean $\mu > 0$, i.e. $n \sim \mathrm{Po}(\mu)$. The probability to observe $n$ events ($n = 0, 1, \ldots$) is

$$P_{\mathrm{Po}}(n|\mu) = \frac{\mu^n}{n!} e^{-\mu}. \tag{A.1}$$

Our aim is to deduce some information about the Poisson mean $\mu$ from a measurement in which $n$ counts were registered. For this purpose, we adopt the Bayesian reasoning. The probability distribution of the Poisson mean to have the value $\mu > 0$ is found by means of Bayes' theorem

$$f(\mu|n) \propto L(\mu|n) p(\mu) \tag{A.2}$$

where $L(\mu|n) = P_{\mathrm{Po}}(n|\mu)$ is the likelihood function and $p(\mu)$ denotes the prior distribution of the mean $\mu$.

The problem is solved once we specify the form of the prior distribution. To this end, we use Gamma distributions that provide a family of conjugate prior distributions for the Poisson likelihood function

$$p(\mu) = f_{\mathrm{Ga}}(\mu|s, \lambda) = \frac{\lambda^s}{\Gamma(s)} \mu^{s-1} e^{-\lambda\mu} \tag{A.3}$$

where $s > 0$ is the shape parameter, $\lambda > 0$ denotes the rate parameter and $\Gamma(s)$ is the Gamma function. Notice that the mean and variance of a random variable obeying the Gamma distribution are $\mathrm{E}(\mu) = \lambda^{-1} s$ and $\mathrm{Var}(\mu) = \lambda^{-2} s$, respectively. Hence, with the increasing value of the shape parameter $s$, the prior distribution is peaked at larger values around a mode $\lambda^{-1}(s-1)$. With the increasing value of the rate parameter $\lambda$, that shifts the position of the mean towards lower values, the prior distribution becomes narrower.

The Gamma family of prior distributions is sufficiently large. The two prior parameters $s$ and $\lambda$ may be chosen to contain our degree of belief about the problem before the experiment is conducted. Notice that traditionally accepted prior assumptions about the studied parameter are included among these possibilities. For example, in a limiting case, if $\lambda \to 0$, the choice $s = 1$ represents the uniform prior, $s = \frac{1}{2}$ is for the Jeffreys' prior (see Appendix B) and, if $n > 0$, then the scale invariant prior distribution with s→0 may be selected.

The posterior distribution of the Poisson mean $\mu$ then depends on the prior choice and experimental data. If $n$ events were collected, one easily finds that $\mu \sim \mathrm{Ga}(p, \gamma)$, where $p = n + s > 0$ and $\gamma = \lambda + 1 > 1$, follow from Eq. (A.2) for the prior distributions chosen from the Gamma family defined in Eq. (A.3). Hence, the posterior distribution function is

$$f(\mu|n) = f_{\mathrm{Ga}}(\mu|p, \gamma) = \frac{\gamma^p}{\Gamma(p)} \mu^{p-1} e^{-\gamma\mu}. \tag{A.4}$$

Let us finally note that for the random variable $\mu' = k\mu$, where $k > 0$ is a constant, one obtains $\mu' \sim \mathrm{Ga}(p, \frac{\gamma}{k})$.

## Appendix B. Jeffreys' prior

By definition, the Jeffreys' prior is proportional to the square root of the determinant of the Fisher information. In the case of a single-valued Poisson mean $\mu > 0$, it is written

$$p(\mu) \propto \sqrt{-\mathrm{E}\left[\frac{\partial^2 \ln L(\mu|n)}{\partial^2 \mu}\right]} = \frac{1}{\sqrt{\mu}} \tag{B.1}$$

where $L(\mu|n) = P_{\mathrm{Po}}(n|\mu)$ is the likelihood function given in Eq. (A.1) and E denotes the mean value with respect to the Poisson model under study.

## References

[1] S.S. Wilks, Ann. Math. Stat. 9 (1938) 60.
[2] T.P. Li, Y.Q. Ma, Astrophys. J. 272 (1983) 317.
[3] R.D. Cousins, J.T. Linnemann, J. Tucker, Nucl. Instrum. Methods A 595 (2008) 480.
[4] G. Cowan, K. Cranmer, E. Gross, O. Vitells, Eur. Phys. J. C71 (2011) 1554; G. Cowan, K. Cranmer, E. Gross, O. Vitells, Eur. Phys. J. C73 (2013) 2501.
[5] G.J. Feldman, R.D. Cousin, Phys. Rev. D57 (1998) 3873.
[6] R.D. Cousins, Nucl. Instrum. Methods A 417 (1998) 391.
[7] W.A. Rolke, A.M. López, J. Conrad, Nucl. Instrum. Methods A 551 (2005) 493.
[8] O. Helene, Nucl. Instrum. Methods 212 (1983) 319.
[9] O. Helene, Nucl. Instrum. Methods 228 (1984) 120.
[10] H.B. Prosper, Nucl. Instrum. Methods A 241 (1985) 236.
[11] H.B. Prosper, Phys. Rev. 37 (1988) 1153.
[12] S. Gillesen, H.L. Harney, Astron. Astrophys. 430 (2005) 355.
[13] P. Gregory, Bayesian Logical Data Analysis for the Physical Sciences, Cambridge University Press, Cambridge, 2005, (Chapter 14).
[14] M.L. Knoetig, Astrophys. J. 790 (2014) 106.
[15] D. Casadei, Astrophys. J. 798 (2015) 5.
[16] F.W.J. Olver, D.M. Lozier, R.F. Boisvert, C.W. Clark (Eds.), NIST Handbook of Mathematical Functions, Cambridge University Press, Cambridge, 2010 (Chapters 8, 13 and 15).
[17] V.A. Acciari, et al., Astrophys. J. 743 (2011) 62.
[18] A.A. Abdo, et al., Astrophys. J. 707 (2009) 580.

## E A Bayesian on-off analysis of cosmic ray data

D.Nosek, J.Noskova, *A Bayesian on-off analysis of cosmic ray data*, Nuclear Instruments and Methods in Physics Research A 867, 222-230, 2017 [arXiv:1707.03155].
DOI: 10.1016/j.nima.2017.06.034

Nuclear Inst. and Methods in Physics Research, A 867 (2017) 222–230

Contents lists available at ScienceDirect

Nuclear Inst. and Methods in Physics Research, A

journal homepage: www.elsevier.com/locate/nima

# A Bayesian on–off analysis of cosmic ray data

Dalibor Nosek [a,*], Jana Nosková [b]

[a] Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic
[b] Czech Technical University, Faculty of Civil Engineering, Prague, Czech Republic

ARTICLE INFO

*Keywords:*
Bayesian inference
On–off problem
Source detection
Cosmic rays

ABSTRACT

We deal with the analysis of on–off measurements designed for the confirmation of a weak source of events whose presence is hypothesized, based on former observations. The problem of a small number of source events that are masked by an imprecisely known background is addressed from a Bayesian point of view. We examine three closely related variables, the posterior distributions of which carry relevant information about various aspects of the investigated phenomena. This information is utilized for predictions of further observations, given actual data. Backed by details of detection, we propose how to quantify disparities between different measurements. The usefulness of the Bayesian inference is demonstrated on examples taken from cosmic ray physics.

## 1. Introduction

The search for new phenomena often yields data that consists of a set of discrete events distributed in time, space, energy or some other observables. In most cases, source events associated with a new effect are hidden by background events, while these two classes of events cannot be distinguished in principle. Such a search can be accomplished with an on–off measurement by checking whether the same process of a constant but unknown intensity may be responsible for observed counts in the on-source region, where a new phenomenon is searched for, and in the reference off-source region, where only background events contribute. Any inconsistency between the numbers of events collected in these zones, when they are properly normalized, then indicates the predominance of a source producing more events in one explored region over the other.

In this study, we focus on the problems which are often encountered when searching for cosmic ray sources while detecting rare events. Characteristics of possible sources are usually proposed based on analysis of a test set of observed data. Then, further observations are to be conducted in order to examine the presence of a source or to improve conditions for its verifications. But, due to unknown phenomena, the outcome is always uncertain which calls, first, for as less as possible initial assumptions about underlying processes and, second, for the quantification of disparities between observations with the option to correct for experimental imperfections.

In order to satisfy the first condition, we follow our previous analysis of on–off measurements formulated within the Bayesian setting [1]. Unlike other Bayesian approaches [2–9], we handle the source and background processes on an equal footing. This option provides us with solutions that are minimally affected by external presumptions. In order to track the behavior of a signal registered in a selected on-source region, we utilize variables with the capability to assess the consistency between on–off measurements. Specifically, giving the net effect, the difference variable [1] is well suited for estimating source fluxes if exposures are known. In case of stable or at least predictable background rates, we eliminate the effect of exposures by using fractional variables which reveal relatively the manifestation of a source. For example, the time evolution of a given source, if still observed in the same way, is easily examined by the ratio of the on-source rate to the total rate. In a more general case, we employ the on-source rate expressed in terms of the rate deduced from the background. In summary, we receive posterior distributions of different variables that include what is available from measurements, while providing us with all kinds of estimates, as traditionally communicated, and allowing us to make various observation-based predictions.

Related to the on–off issue, the Bayesian inference provides solutions in the case of small numbers, including the null experiment or the experiment with no background, when classical methods based on the asymptotic properties of the likelihood ratio statistic [10–13] are not easily applicable. Also, there are no difficulties with the regularity conditions of Wilks' theorem, with unphysical likelihood estimates or with the discreteness of counting experiments, in general, see e.g. Refs. [14–17]. On the other hand, the subjective nature of Bayesian reasoning, often

[*] Corresponding author.
*E-mail address:* nosek@ipnp.troja.mff.cuni.cz (D. Nosek).

http://dx.doi.org/10.1016/j.nima.2017.06.034
Received 24 August 2016; Received in revised form 1 May 2017; Accepted 21 June 2017
Available online 27 June 2017

mentioned as its disadvantage, may be at least partially eliminated by using a family of uninformative prior options.

The proposed method is suitable for experiments searching for rare events in which the observational conditions may not be adjusted optimally, with little opportunity for repeating measurements conducted under exactly the same conditions. Besides searches for possible sources of the highest energy cosmic rays, see e.g. Refs. [18–22], examples include observations of peculiar sources which exhibit surprising temporal or spectral behavior. Another class of observations comprises searches for events accompanying radiation from transient sources that have been identified in different energy ranges. The identification of the properties of very-high-energy $\gamma$-rays associated with observed gamma-ray bursts belongs to this class of problems [1–3].

The structure of this paper is as follows. Our formulation of the Bayesian approach to the on–off problem is described in Section 2, complemented by five Appendices. Further details about our approach can be found in Ref. [1]. In Section 2.1 we summarize how to store experimental information by using appropriate on–off variables. Two ways to examine possible inconsistencies in independent observations are proposed in Sections 2.2 and 2.3. Several realistic examples taken from cosmic ray physics are presented and discussed in Section 3. The paper is concluded in Section 4.

## 2. Bayesian inferences from on–off experiment

In the on–off experiment, two kinds of measurements are collected in order to validate a source signal immersed in background. The number of on-source events, $n_{\rm on}$, is recorded in a signal on-source region, while the number of off-source events, $n_{\rm off}$, detected in a background off-source zone serves as a reference measurement. The on- and off-source counts are modeled as discrete random variables generated in two independent Poisson processes with unknown on- and off-source means, $\mu_{\rm on}$ and $\mu_{\rm off}$, i.e. $n_{\rm on} \sim {\rm Po}(\mu_{\rm on})$ and $n_{\rm off} \sim {\rm Po}(\mu_{\rm off})$. The relationship between the on- and off-source zone is ensured by the ratio of on- and off-source exposures $\alpha > 0$.

In the Bayesian approach, for on- and off-source means we adopted a family of prior distributions conjugate to the Poisson sampling process [1]. This family consists of Gamma distributions, i.e.

$$\mu_{\rm on} \sim {\rm Ga}(s_p, \gamma_p - 1), \qquad \mu_{\rm off} \sim {\rm Ga}(s_q, \gamma_q - 1), \tag{1}$$

where $s_p > 0$ and $s_q > 0$ are prior shape parameters, and the prior rate parameters $\gamma_p > 1$ and $\gamma_q > 1$. It includes several frequently discussed options, i.e. scale invariant, uniform, as well as Jeffreys' prior distributions. After the on–off measurement has been conducted, when $n_{\rm on}$ and $n_{\rm off}$ counts were registered independently in the on- and off-source regions, using Eq. (1) we obtain independent posterior distributions

$$(\mu_{\rm on} \,|\, n_{\rm on}) \sim {\rm Ga}(p, \gamma_p), \qquad (\mu_{\rm b} \,|\, n_{\rm off}) \sim {\rm Ga}\left(q, \frac{\gamma_q}{\alpha}\right), \tag{2}$$

where $\mu_{\rm b} = \alpha\mu_{\rm off}$ denotes the expected background rate in the on-source zone and $p = n_{\rm on} + s_p$ and $q = n_{\rm off} + s_q$. For more details see Ref. [1].

We recall that our next steps diverge from the traditional treatment. In order to assess what is observed, we define suitable on–off variables by combining the on- and off-source means, assuming that the underlying processes are independent. From the Bayesian perspective, this choice is motivated by the fact that, according to Jeffreys' rule, the joint prior distribution is separable in the on- and off-source means [1,2]. Furthermore, as in classical statistical approaches [10–16], the proposed option allows us to obtain adequate results regardless of in which of the two zones the source effects are revealed [1,7].

### 2.1. On–off variables

In our previous work [1], we focused on the properties of the difference between the on-source and background means, $\delta = \mu_{\rm on} - \mu_{\rm b}$, using maximally uninformative joint distributions, as dictated by the principle of maximum entropy. In this section, we briefly recapitulate our previous result and introduce other on–off variables that equally well describe the on–off problem.

Under the transformation $\delta = \mu_{\rm on} - \mu_{\rm b}$, with a real valued domain, while keeping $\mu_{\rm b} = \alpha\mu_{\rm off}$ unchanged and marginalizing over $\mu_{\rm b}$, the probability density function of the difference is (for details of our notation see Ref. [1])

$$f_\delta(x) = \frac{\gamma_p^p \left(\frac{\gamma_q}{\alpha}\right)^q}{\Gamma(p)} \, e^{-\gamma_p x} x^{p+q-1} \, U(q, p+q, \eta x), \quad x \ge 0, \tag{3}$$

$$f_\delta(x) = \frac{\gamma_p^p \left(\frac{\gamma_q}{\alpha}\right)^q}{\Gamma(q)} \, e^{\frac{\gamma_q}{\alpha} x} (-x)^{p+q-1} \, U(p, p+q, -\eta x), \quad x < 0, \tag{4}$$

where $p = n_{\rm on} + s_p$, $q = n_{\rm off} + s_q$, $\eta = \gamma_p + \frac{\gamma_q}{\alpha}$, $\Gamma(a)$ stands for the Gamma function and $U(a, b, z)$ is the Tricomi confluent hypergeometric function [23]. Exhaustive discussion concerning this distribution can be found in Ref. [1], where also some special cases ($\gamma_p = \gamma_q \to 1$) based on uninformative prior distributions, scale invariant ($s_p = s_q \to 0$), Jeffreys' ($s_p = s_q = \frac{1}{2}$) and uniform ($s_p = s_q = 1$) options, are described.

The difference $\delta$ yields information about the source flux. The posterior distribution of the source flux is obtained by a scale transformation, i.e. $j = \delta/a$ where $a = \frac{\alpha}{1+\alpha} A$ is the exposure of the on-source zone and $A$ denotes the integrated exposure of the on–off experiment, both considered as constants.

A similar picture is obtained with the ratio of the on-source and background means ($\mu_{\rm b} = \alpha\mu_{\rm off}$)

$$\beta = \frac{\mu_{\rm on}}{\mu_{\rm b}}, \quad \beta \ge 0. \tag{5}$$

This variable represents the intensity registered in the on-source region expressed in terms of the background intensity, i.e. $\beta \le 1$ when no source is present in the on-source zone. The ratio $\beta$ obeys the generalized Beta distribution of the second kind [24], $\beta \sim {\rm B}_{\rm g2}(p, q, \rho)$ where $p = n_{\rm on} + s_p$, $q = n_{\rm off} + s_q$ and $\rho = \alpha\gamma_p/\gamma_q$, with the probability density function

$$f_\beta(x) = \frac{\rho^p}{B(p,q)} \frac{x^{p-1}}{(1+\rho x)^{p+q}}, \quad x \ge 0, \tag{6}$$

where $B(a, b)$ is the Beta function [23]. This posterior distribution was obtained after the transformation $\beta = \mu_{\rm on}/\mu_{\rm b}$ while treating $\mu_{\rm on}$ and $\mu_{\rm b}$ as independent variables (see Eq. (2)) and keeping $\mu_{\rm b}$ unchanged, with the Jacobian $J = \mu_{\rm b}$, and marginalizing over $\mu_{\rm b}$.

In a special case, using the uniform prior distributions for the on- and off-source means, i.e. $\gamma_p = \gamma_q \to 1$ and $s_p = s_q = 1$, and assuming that the on–off data were registered in the regions of the same exposure, when $\rho = \alpha = 1$, the posterior distribution for the ratio $\beta$ written in Eq. (6) reduces to the result given originally in Ref. [5]. Assuming $\gamma_p = \gamma_q \to 1$ and $\alpha = 1$, i.e. $\rho = 1$, the result presented in Eq. (13) in Ref. [6] is obtained.

In some cases, it may be appropriate to use a variable

$$\omega = \frac{\mu_{\rm on}}{\mu_{\rm on} + \mu_{\rm off}}, \quad \omega \in \langle 0, 1 \rangle, \tag{7}$$

that represents the fraction of the total intensity registered in the on-source zone. Considering that $\omega = \alpha\beta/(1 + \alpha\beta)$, we recover from Eq. (6) that the probability density function of the proportion $\omega$ is

$$f_\omega(x) = \frac{\kappa^p}{B(p,q)} \frac{x^{p-1}(1-x)^{q-1}}{[1 + (\kappa - 1)x]^{p+q}}, \quad x \in \langle 0, 1 \rangle, \tag{8}$$

where $p = n_{\rm on} + s_p$, $q = n_{\rm off} + s_q$ and $\kappa = \gamma_p/\gamma_q$ is the ratio of the prior rate parameters. In this case, equally intensive on- and off-source processes ($\mu_{\rm on} = \mu_{\rm b}$) are described by a balance value of $\omega = \frac{\alpha}{1+\alpha}$.

Note that any Bayesian statement based on the probabilities inferred from the above derived distributions is independent of the prior rate parameters when $\gamma_p = \gamma_q$ and thus $\rho = \alpha$. For $\gamma_p = \gamma_q$, we even have that the proportion $\omega$ obeys the Beta distribution, i.e. $\omega \sim \mathrm{B}(p, q)$. This widely used option also follows from using the prior Beta distributions conjugate to the binomial sampling process, i.e. prior $\omega \sim \mathrm{B}(s_p, s_q)$. In the context of on–off measurements, the classical analysis of the binomial proportion is discussed in Refs. [12,15], for example. Point estimates of the proportion $\omega$ are traditionally used in the analysis of directional data in cosmic ray physics, see e.g. Refs. [18–20,22,25,26].

The proposed Bayesian solutions to the on–off problem have other interesting features. Unlike traditional approaches [2–9], we treat the on- and off-source processes as independent. Hence, our posterior distributions are maximally noncommittal about missing information on the relationship between these processes. Moreover, receiving information separately from the on- and off-source observations, the on–off problem is examined without a predetermined assumption in which zone the source is to be searched for [1]. Thus, any detected imbalance will lead to the same conclusion notwithstanding the region where more activity is expected [1]. Note that most classical test statistics relevant to the on–off problem possess the same property [11,15,16].

Other technical details are summarized in Appendices. In Appendix A we show that all three on–off variables provide the same probability of the source absence in the on-source zone. Note, however, that the fractional variables $\beta$ or $\omega$, which are easier to handle, do not substitute for the difference $\delta$.

A way how to determine the shortest credible intervals for the on–off variables is described in Appendix B. In Appendix C we show how to modify Bayesian solutions, when a source is known to be present in the on-source zone. Similar solutions are also obtained in often adopted schemes, whereby source and background parameters are treated as independent variables [2,3,6–9]. In Appendix D we present Bayesian solutions for cases when background rates are known with sufficient precision.

### 2.2. *Waiting for next events*

Current experiments collecting rare events raise interest for predictions based on previous observations. Typically, we want to know how many events must be registered in a subsequent experiment in order to identify a given number of events in a selected on-source zone, while relying on previous data collected under the same conditions with the same instrument. This issue is solved by constructing a relevant predictive distribution.

According to previous considerations, we assume that the numbers of on- and off-source events registered in a new experiment up to and including time $t$ are generated in two independent Poisson processes $\{N_{\rm on}(t); t \geq 0\}$ and $\{N_{\rm off}(t); t \geq 0\}$ with respective rates $\mu_{\rm on}$ and $\mu_{\rm off}$, i.e. among others, $N_{\rm on}(t) \sim \mathrm{Po}(\mu_{\rm on} t)$ and $N_{\rm off}(t) \sim \mathrm{Po}(\mu_{\rm off} t)$. Hence, we know that events of the merged Poisson process $\{N(t) = N_{\rm on}(t) + N_{\rm off}(t); t \geq 0\}$, $N(t) \sim \mathrm{Po}(\mu t)$ where $\mu = \mu_{\rm on} + \mu_{\rm off}$, arrive into the on-source zone with the probability $\omega = \mu_{\rm on}/\mu$ independently of each other and independently of their arrival times, see e.g. Ref. [27]. Consequently, if the total number of events $n > 0$ is collected up to time $t$, the corresponding number of on-source events, $Y_{\rm on} = (N_{\rm on}(t) \mid N(t) = n)$, has a binomial distribution with parameters $n$ and $\omega$, i.e. $Y_{\rm on} \sim \mathrm{Bi}(n, \omega)$. We also know that the total number of events recorded until a predefined number $k > 0$ of events arrive into the on-source zone, $Y = (N(t) \mid N_{\rm on}(t) = k$, the on-source event is the last one), has a shifted negative binomial distribution (waiting time distribution) with parameters $k$ and $\omega$, i.e. $Y \sim \mathrm{NBi}(k, \omega)$ with support $n = k, k+1, \ldots$, see e.g. Ref. [28].

Further, we ask for the probability $p_{n,k}(\omega)$ that more than $n$ events in total are collected before the $k$th on-source event is registered if, as justified above, events are switched independently between on- and off-source zones with the probability $\omega$. We obtain ($k > 0$ and $n = k, k+1, \ldots$)

$$p_{n,k}(\omega) = P(Y > n \,|\, \omega) = P(Y_{\rm on} < k \,|\, \omega) = \sum_{i=0}^{k-1} \binom{n}{i} \omega^i (1-\omega)^{n-i}, \tag{9}$$

where we use the relation between the negative binomial variable $Y$ and the binomial variable $Y_{\rm on}$, see e.g. Eq. (5.31) in Ref. [28]. This way, Eq. (9) gives the probability of the waiting time for the $k$th on-source event when the time is measured in terms of the total number of collected events $n$.

In order to determine the chances of identifying on-source events in a new series of observations, we need to be informed about the binomial parameter $\omega$. We use the fact that, in the Bayesian concept, the information on future measurements is contained in the posterior predictive distribution of unobserved observations, conditional on the already observed data. This distribution is obtained by marginalizing the distribution of the new data, given parameters, over the posterior distribution of parameters, given the previous data, accounting thus for uncertainty about involved parameters.

Since the Poisson processes guarantee that the new and old observations in disjoint time intervals are independent, when conditioned on parameters $\mu_{\rm on}$ and $\mu_{\rm off}$, or, equivalently, on $\mu = \mu_{\rm on} + \mu_{\rm off}$ and $\omega = \mu_{\rm on}/\mu$, and since the waiting time probability given in Eq. (9) is independent of $\mu$, we can write

$$P(Y > n, \mu, \omega \,|\, D) = P(Y > n \,|\, \omega) p(\mu, \omega \,|\, D), \tag{10}$$

where $D = (n_{\rm on}, n_{\rm off})$ denotes the old on–off data and $p(\mu, \omega \mid D)$ is the joint posterior distribution of $\mu$ and $\omega$ which is obtained via Bayes' rule using the prior distributions for $\mu_{\rm on}$ and $\mu_{\rm off}$ in Eq. (1). Hence, by marginalizing over $\mu$ and $\omega$, we obtain from Eqs. (9) and (10) that, in the new data set, the waiting time for the $k$th on-source event exceeds $n$ with the probability

$$\begin{aligned} P_{n,k} &= \int_0^1 \left[ \int_0^\infty P(Y > n, \mu = y, \omega = x \,|\, D) \mathrm{d}y \right] \mathrm{d}x \\ &= \int_0^1 p_{n,k}(x) f_\omega(x) \mathrm{d}x, \end{aligned} \tag{11}$$

where $f_\omega(x) = p(\omega = x \mid D) = \int_0^\infty p(\mu = y, \omega = x \mid D)\mathrm{d}y$ is the posterior distribution of the proportion $\omega$ given in Eq. (8). In particular, assuming that $\omega \sim B(p, q)$ for $\gamma_p = \gamma_q$ ($\kappa = 1$) where $p = n_{\rm on} + s_p$ and $q = n_{\rm off} + s_q$ are known from the previous measurement, it follows that

$$P_{n,k} = \int_0^1 p_{n,k}(x) f_\omega(x) \mathrm{d}x = \sum_{i=0}^{k-1} \binom{n}{i} \frac{B(p+i, q+n-i)}{B(p,q)}. \tag{12}$$

Here, the Beta functions are replaced by the incomplete Beta functions, $B(a, b) \to B_{\frac{1}{1+\alpha}}(a, b)$, if a source is considered to be present in the on-source zone, see Appendix C.

The application of this result to the new data allows us to assess the consistency between subsequent observations. Consider that $n$ new events in total are registered until the $k$th new event arrives into the on-source zone, while the previous data has been processed. We know that the probability of the new observation is $P_{n,k}$ provided the new and old data are generated in the counting model described above. In the classical sense, it means that our initial assumptions are not valid at a level of confidence $\mathrm{CL} < 1 - P_{n,k}$. Hence, at this level of confidence, our data-driven model fails to describe what has been measured and we conclude that, besides other possibilities, the new data may indicate a smaller on-source signal or a larger background rate than would correspond to the previous measurement.

### 2.3. *Comparison of on–off measurements*

In this section we address the question of how to compare two independent on–off measurements. Our goal is to quantify statistically which of the measurements indicate a more intense emitter, while relying on information about observations contained in the posterior distributions of on–off variables. Besides sequential measurements performed under similar conditions, we also admit experiments conducted with different equipments, for example, when different sources in different spatial, time or energy ranges are observed.

We assume that two independent on–off observations, marked by indices 1 and 2, were collected and processed by the method described in Section 2.1. Depending on what we want to examine, we choose one type of the on–off variable. The relationship between the two Bayesian outputs is quantified by the probability $P(\tau_1 < A\,\tau_2 \mid D_1, D_2)$ where $\tau_1(\tau_2)$ is a suitable on–off variable ($\tau = \delta, \beta$ or $\omega$) for the first (second) measurement and $D_1 = (n_{\mathrm{on}_1}, n_{\mathrm{off}_1})$ ($D_2 = (n_{\mathrm{on}_2}, n_{\mathrm{off}_2})$) denotes the corresponding on–off data. This probability is determined by integrating the joint probability distribution of $\tau_1$ and $\tau_2$ over a relevant two-dimensional domain. Here, a constant $A$ is used to account, at least to first order, for different observational conditions or experimental imperfections (see below).

From a practical perspective, the best way is to compare source fluxes. For this, we utilize the unconditional distributions of the differences $\delta_1$ and $\delta_2$, respectively, see Eqs. (3)–(4). The probability that the flux $j_1 = \delta_1/a_1$ observed in the first observation is less than the flux $j_2 = \delta_2/a_2$ deduced from the second one, both fluxes treated as random variables, is

$$P(j_1 < j_2) = P\left(\delta_1 < \frac{a_1}{a_2}\delta_2\right) = \int_{-\infty}^{\infty} f_{\delta_1}(x_1)\left[\int_{\frac{a_2}{a_1}x_1}^{\infty} f_{\delta_2}(x_2)\mathrm{d}x_2\right]\mathrm{d}x_1. \quad (13)$$

Here, the assessment of stability of source fluxes requires the knowledge of the on-source exposures, $a_1$ and $a_2$. However, they may be affected by various imperfections associated with details of detection and data processing, especially when different sources are examined by different techniques.

The discrepancy between two independent observations can also be described by comparing the ratio variables while canceling out the effect of exposures. If we adopt the unconditional distributions for the ratio variables $\beta_1$ and $\beta_2$ given in Eq. (6), the inconsistency between two sets of on–off data can be quantified by the probability

$$P(\beta_1 < \xi\beta_2) = \int_{0}^{\infty} f_{\beta_1}(x_1)\left[\int_{\xi^{-1}x_1}^{\infty} f_{\beta_2}(x_2)\mathrm{d}x_2\right]\mathrm{d}x_1. \quad (14)$$

Here, for further possible applications, we introduced a parameter $\xi > 0$, allowing us to compare multiples of the ratio variables. In a first order approach, this parameter can be employed to eliminate imperfections attributable to detection and data evaluation.

When two measurements collected in the same on- and off-source zones are studied ($\alpha_1 = \alpha_2$), the proportion $\omega$ is advantageously used after a straightforward modification of Eq. (14). Note also that the proposed probabilities are easily modified if sources are assumed to be present in their on-source zones, see Appendix C. Specifically, when non-negative source rates are guaranteed due to external arguments, the probabilities of inconsistency are obtained by putting the conditional distributions into the relevant equations while changing the integration limits accordingly.

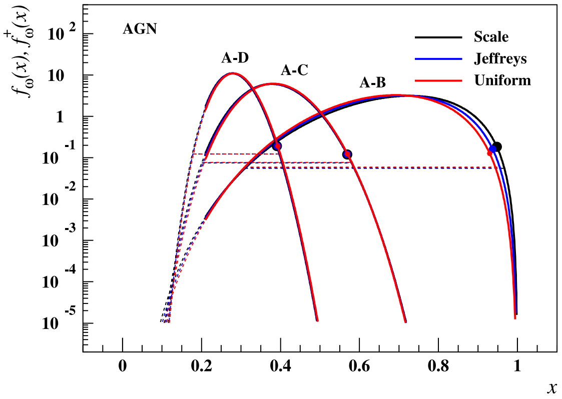


**Fig. 1.** Distributions of proportion $\omega$ for AGN data [22]. The same uninformative priors for on- and off-source means ($s = s_p = s_q$ and $\gamma_p = \gamma_q \to 1$) are used. Results for scale invariant ($s \to 0$), Jeffreys' ($s = \frac{1}{2}$) and uniform ($s = 1$) priors are shown in black, blue and red, respectively. Distributions for the proportion, $f_\omega(x)$, and distributions $f_\omega^+(x)$, when conditioned on a non-negative source rate ($\omega \geq \frac{\alpha}{1+\alpha}$), are depicted as dashed and thick full curves, respectively. Horizontal dashed lines visualize credible intervals for the proportion ($\langle\omega_-, \omega_+\rangle$) at a $3\sigma$ level of confidence. Upper limits at the same confidence level for the proportion assumed to be non-negative ($\omega_+^+$ for $\omega \geq \frac{\alpha}{1+\alpha}$) are shown by colored points. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The integration in Eqs. (13) and (14) is to be performed numerically over the indicated two-dimensional sets. In some counting experiments, background rates can be estimated with sufficient accuracy from auxiliary measurements or modeled numerically. With this simplification, we obtained explicit formulae for the probabilities of inconsistency summarized in Appendix E.

The probabilities of inconsistency given in Eqs. (13) and (14) have somewhat different meanings. The difference $\delta$ allows us to quantify disparities between source fluxes, when on-source exposures are known. The probabilities based on the fractional variables $\beta$ and $\omega$ describe discrepancies between on-source observations when expressed with respect to the background or total measurements, respectively. Thus, in more complicated cases, additional information about details of detection and data processing is needed for their correct interpretation (e.g. background rates, energy ranges, data quality limits etc.).

The probabilities written in Eqs. (13) and (14) do not substitute for the probabilities of the source presence in the on-source zone, see Appendix A. Indeed, it can be more likely that a larger flux is observed from a source which is found to be less significant than the other, i.e. $P(j_1 < j_2) > 0.50$ while $P_1^+ > P_2^+$ and vice versa. Note also that quantified disparities between source fluxes, $P(j_1 < j_2)$, when compared to ratio results, $P(\beta_1 < \beta_2)$, for a given pair of observations, may reveal hitherto unnoticed features that could affect measurements, were not considered during data processing or disrupted homogeneity of the underlying Poisson processes.

**Table 1**
AGN and Cen A data measured by the Auger surface detector [20–22] and the HS data detected by the Telescope Array [29]. Source assignment, period, exposure $A$ in km$^2$ sr y, measured on- and off-source counts and the on–off parameter $\alpha$ are listed in the first sixth columns. The endpoints of examined periods are denoted by $A$ = (May 27, 2006), $B$ = (Aug 31, 2007), $C$ = (Dec 31, 2009), or $C$ = (Jan 1, 2010) for Cen A, and $D$ = (Mar 31, 2014), respectively. For the HS we used the two-year data collected from $E$ = (May 5, 2013) to $F$ = (May 11, 2015), see Table 1 in Ref. [29]. The Bayesian probabilities of no source ($P^-$), corresponding significances ($S_{\mathrm{B}}$) and Li-Ma significances ($S_{\mathrm{LM}}$) are given in the next three columns. For Bayesian results, Jeffreys' prior distributions were adopted, i.e. $s = \frac{1}{2}$ and $\gamma \to 1$.

| Data | Period | A | $n_{\mathrm{on}}$ | $n_{\mathrm{off}}$ | $\alpha$ | $P^-$ | $S_{\mathrm{B}}$ | $S_{\mathrm{LM}}$ |
|---|---|---|---|---|---|---|---|---|
| AGN | $A$–$B$ | 4 500 | 9 | 4 | 0.266 | $8.2\ 10^{-5}$ | 3.77 | 3.73 |
| AGN | $A$–$C$ | 15 980 | 21 | 34 | 0.266 | $1.7\ 10^{-3}$ | 2.93 | 2.90 |
| AGN | $A$–$D$ | 47 363 | 41 | 105 | 0.266 | $2.0\ 10^{-2}$ | 2.05 | 2.03 |
| Cen A | $C$–$D$ | 31 383 | 3 | 76 | 0.047 | 0.59 | −0.22 | −0.31 |
| HS | $E$–$F$ | | 5 | 32 | 0.075 | $7.0\ 10^{-2}$ | 1.48 | 1.40 |

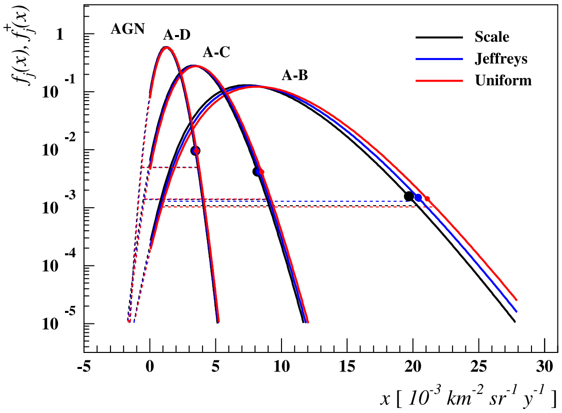


**Fig. 2.** Distributions of source flux $j = \delta/a$ ($a = \frac{\alpha}{1+\alpha}A$) for AGN data [22]. Both types of distributions are shown, $f_j(x)$ (dashed curves) and $f_j^+(x)$ for $j \geq 0$ (thick full curves). For further details see caption to Fig. 1.

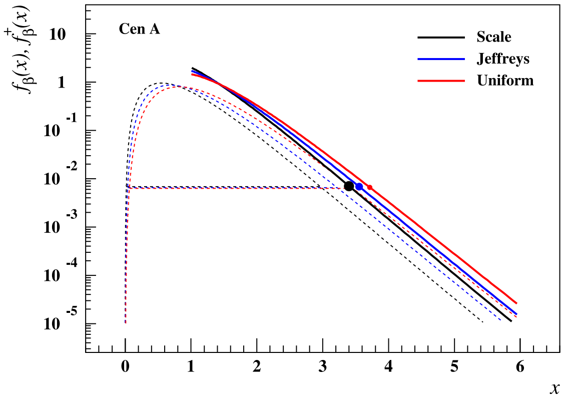


**Fig. 3.** Distributions of ratio $\beta$ for Cen A data [22]. Both types of distributions are shown, $f_\beta(x)$ (dashed curves) and $f_\beta^+(x)$ for $\beta \geq 1$ (thick full curves). For further details see caption to Fig. 1.

## 3. Examples

The usefulness of the method described in Section 2 is demonstrated using arrival directions of the highest energy cosmic rays measured by the Pierre Auger Observatory [20–22]. Considering a predefined set of positions of nearby active galactic nuclei (AGN), we provide information to what extent is this set of possible sources related to directional data after this association has been suggested [18,19]. In a similar way, we also examine a signal that has been initially associated with the region around Centaurus A (Cen A) [20,21]. We emphasize that earlier conclusions [18–22] are in line with our analysis. Our aim is not to reassess previous studies, we only point out how the previous findings may be viewed from different perspectives.

Regardless of the results of further analysis [20,22], we assumed that the signals from AGNs [18,19] and Cen A [20,21] have not yet been confirmed. Given the data that were observed in the preselected on–off regions, we calculated the posterior distributions of the difference and fractional variables. We assumed the same prior distributions for the on- and off-source means with common shape parameters and zero rate parameters, i.e. $s = s_p = s_q$ and $\gamma = \gamma_p = \gamma_q \to 1$. Furthermore, we derived the posterior distributions of the source flux $j$ using $j = \delta/a$ where $a = \frac{\alpha}{1+\alpha}A$ and $A$ denotes the integrated exposure of the period of data taking.

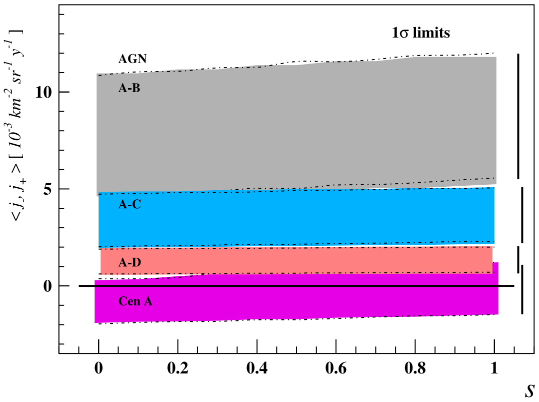


**Fig. 4.** Credible intervals for source flux $j = \delta/a$ ($a = \frac{\alpha}{1+\alpha}A$) at a 1$\sigma$ level of confidence are shown as functions of the common shape parameter of prior distributions $s = s_p = s_q$ ($\gamma_p = \gamma_q \to 1$). Results for AGN (gray, blue and red bands) and Cen A (magenta) data [22] are depicted. Dashed-dot lines indicate limits estimated using the approach based on known background (see Appendix D). Black vertical lines show classical limits deduced within the unbounded profile likelihood analysis [16]. The horizontal black line represents the background expectation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

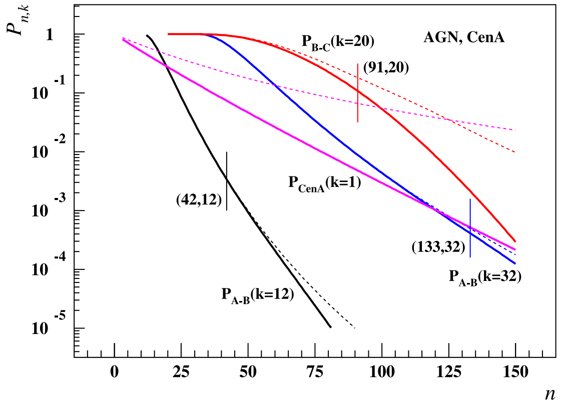


**Fig. 5.** Waiting time predictions. The probabilities $P_{n,k}$ that less than $k$ on-source events are observed are shown as functions of the total number of registered events $n$. Predictions based on the AGN signals observed in $A$–$B$ period for the next 12 (32) AGN events are shown in black (blue). $B$–$C$ predictions for the next 20 AGN counts are in red. Magenta lines are for predictions of one next Cen A event, based on the Cen A data from $C$–$D$ period. Dashed (full) lines show unconditional (conditional) results based on Jeffreys' prior distributions. Colored vertical lines indicate observations of $(n, k)$ events collected in the subsequent AGN periods. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In the following, we show how the three on–off variables can be used when examining the previously suggested associations. Based on the results of Section 2.2, we provide examples related to the issue of waiting for the next on-source events. We also present examples of how to compare various independent measurements, see Section 2.3. In the latter case, we include the latest hot spot (HS) data obtained by the Telescope Array surface detector [29].

### 3.1. Active galactic nuclei

Among other important results [22], one of the topics of discussion regarding the distributions of arrival directions of the highest energy

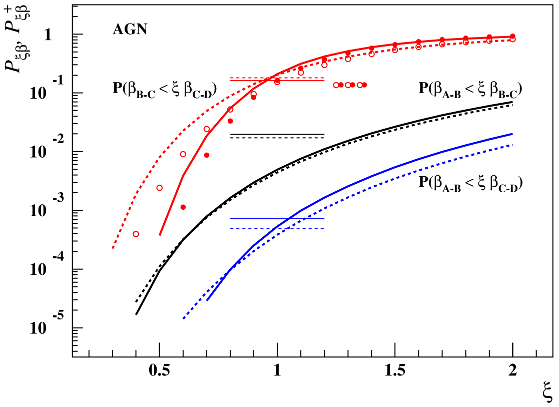


**Fig. 6.** Probabilities of inconsistency for the ratio $\beta$, $P_{\xi\beta} = P(\beta_1 < \xi\beta_2)$, deduced from the AGN data are shown as functions of the parameter $\xi$ (see Section 2.3). Black, blue and red lines are for the comparison of three separated AGN periods. Dashed and full lines show unconditional ($P_{\xi\beta}$) and conditional ($P^{+}_{\xi\beta}$) results, respectively, based on Jeffreys' priors. Red empty (full) points show unconditional (conditional) results for $B$–$C$ and $C$–$D$ periods assuming known background rates (see Appendix D) and uniform priors for on-source means. Thin horizontal lines indicate the probabilities of inconsistency, $P(j_1 < j_2)$, between AGN fluxes. Horizontal chains of three red points are for source fluxes provided that background rates are known (see Appendix E). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

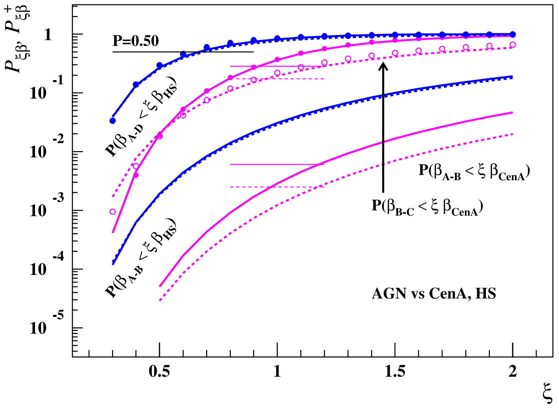


**Fig. 7.** Probabilities of inconsistency for the ratio $\beta$ are shown as functions of the parameter $\xi$ (see Section 2.3). The AGN data collected in periods $A$–$B$ and $B$–$C$ are compared to the Cen A signal in period $C$–$D$, see magenta lines. The probabilities that quantify inconsistency between the HS signal and the AGN data registered in periods $A$–$B$ and $A$–$C$ are shown in blue. The black horizontal line indicates a probability of 0.50. For further details see caption to Fig. 6. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

cosmic rays has focused on their association with a set of positions of nearby objects from the 12th edition of quasars and AGNs [30]. An initially revealed signal [18,19] has been reinvestigated in subsequent studies using the newly registered data [20,22].

In order to document the uses and advantages of the Bayesian reasoning, we examined data registered by the surface detector of the Pierre Auger Observatory since May 27, 2007 up to March 31, 2014 (see Table A1 in Ref. [22]), after the AGN signal was recognized [18,19]. Specifically, we used events with energies in excess of 53 EeV and with zenith angles not exceeding 60°. For the association of the selected events with the nearby AGNs we accepted a set of parameters as defined in Refs. [18,19] and then slightly modified [20,22]. A complex AGN source consists of a unification of circular zones with angular radii 3.1° around the positions of AGNs within 75 Mpc (redshifts $z \leq 0.018$) [30].

We examined three sets of data collected successively, as reported in Refs. [20,22]. Namely, we analyzed arrival directions of events registered since May 27, 2006 up to August 31, 2007 (here denoted as period $A$–$B$, II in Ref. [20]), up to December 31, 2009 (here $A$–$C$, II + III in Ref. [20]) and, finally, up to March 31, 2014 (here $A$–$D$, see also Ref. [22]). The integrated exposures of the Auger surface detector, measured counts of on- and off-source events and on–off parameters $\alpha$, all taken from Refs. [20,22], are summarized in the first six columns in the upper three lines in Table 1. In this table, we also show some statistical characteristics based on the Jeffreys' priors ($s = \frac{1}{2}$, $\gamma \to 1$) and asymptotic Li-Ma significances [11].

The posterior distributions for the proportion $\omega$ are depicted in Fig. 1. In this figure, we show results with three kinds of uninformative prior distributions, namely, for scale invariant ($s \to 0$, in black), Jeffreys' ($s = \frac{1}{2}$, blue) and uniform ($s = 1$, red) prior distributions. Two families of posterior distributions are depicted, unconditional distributions (dashed curves) as well as distributions conditioned on a non-negative source rate in the on-source region (thick full lines), i.e. assuming $\omega \geq \frac{\alpha}{1+\alpha}$, see Appendix C. In Fig. 1, also credible intervals and upper limits for the proportion $\omega$ at a $3\sigma$ level of confidence are visualized (see Appendix B).

As an alternative, in Fig. 2 we show posterior distributions for the AGN flux $j = \delta/a$, given the on–off data in three examined period, and again using the three uninformative prior options. Relevant credible intervals at a $1\sigma$ level of significance are depicted in Fig. 4 as functions of the common shape parameter. The classical estimates [16] and the results with known background rates (see Appendix D) are also shown in Fig. 4.

The posterior distributions shown in Figs. 1 and 2 clearly illustrate that the Bayesian inferences are only slightly dependent on the choice of uninformative prior distributions ($s \in \langle 0, 1 \rangle$, $\gamma \to 1$) if the AGN source exhibits a sufficiently high activity, see also Fig. 4. In such cases, due to large probabilities of the source presence in the AGN region, all conditional distributions approximately follow in their domains relevant unconditional distributions. Furthermore, we learned how accessible information about the AGN source evolves with an increasing number of events recorded in the three successive sets of on–off data. Our Bayesian estimates agree with the reported fractions of events associated with the AGN region and their downward trend [20,22].

A decreasing AGN signal is also reflected in the predictions of the waiting time for the next on-source events when compared with future observations, see Section 2.2. In Fig. 5, we show the probability that less than a given number of AGN events were detected in a number of subsequent measurements, while relying on previous observations. For example, the Auger data collected in $A$–$B$ period predicts that a total of 42 events should be registered prior to the next 12 AGNs events with a probability below 4 $10^{-3}$ (black lines). Hence, when confronted with the Auger data from $B$–$C$ period, in which these numbers were observed, such a waiting time is very unlikely. This result allows us to conclude that the $B$–$C$ data is inconsistent with the $A$–$B$ observation at about a $3\sigma$ level of confidence.

Independent AGN observations are compared in Fig. 6, see Section 2.3. In this case, the source fluxes as well as the ratios $\beta$ ($\xi \approx 1$) are well suited since still the same on- and off-source zones are observed with the same instrument. The parameter $\xi$ is employed to show the probability that one ratio is $\xi$-times smaller than the other or it can correct for imperfections, if known (e.g. different background rates, energy ranges, seasonal effects etc.). Our results depicted in Fig. 6 agree with the findings drawn from the waiting time analysis. Namely, it is very unlikely that the AGN ratio from $A$–$B$ period is less than the ratios derived from the two subsequent periods, and the same holds for the fluxes (black and blue results). But the probability of inconsistency between $B$–$C$ and $C$–$D$ periods are much larger (red results). Note also that the discrepancy between the probabilities calculated for the

AGN fluxes and ratios, when relating $A$–$B$ and $B$–$C$ periods for $\xi \approx 1$ (in black), may indicate inhomogeneities of the underlying Poisson processes.

We also examined two possible signals deduced from different on–off measurements that were collected by different experiments. In Fig. 7, the two-years HS data collected on the northern hemisphere by the Telescope Array surface detector (see Table 1 in Ref. [29]) is compared to the AGN signal measured by the Auger surface array on the southern hemisphere [20,22]. Using two sets of the AGN data, $A$–$B$ and $A$–$D$ periods, the probabilities of inconsistency for the ratio $\beta$ are shown as functions of the parameter $\xi$ (blue lines). Interestingly, since $P(\beta_{\text{A–D}} < \xi\beta_{\text{HS}}) > 0.70$ for $\xi \approx 1$, it is more likely that the less visible HS source ($S_{\text{B}} = 1.48$, see Table 1) manifests itself more markedly, when confronted with background, than the latest signal from AGN emitters ($A$–$D$ period) which are more easily identified ($S_{\text{B}} = 2.05$, see Table 1). In this example, the parameter $\xi$ may be utilized to correct for different energy scales ($E \geq 55$ EeV for the HS [29] while $E > 53$ EeV for the AGNs [22] plus systematic uncertainties) and for different background fluxes (at these energies, the overall flux on the northern hemisphere was measured to be at least twice as large as the southern flux, see e.g. Ref. [31]). If the northern background is truly larger than the southern one and, consequently, the observation of the HS signal is more difficult, one can correct for this effect by using $\xi > 1$, enlarging even more the probability that the AGNs are weaker emitters.

### 3.2. Centaurus A

Centaurus A (NGC 5128), located at a distance less than 4 Mpc, is known as a promising candidate source of the highest energy cosmic rays. Moreover, the nearby Centaurus cluster with large concentration of galaxies lies in approximately the same direction, at a distance of about 50 Mpc. The excess of the highest energy events found in the vicinity of Cen A and the properties of observed signal have been originally reported in Ref. [20]. However, this observation was not confirmed in successive measurements [22].

In this example, we show how the disappearance of a previously specified signal [20] can be justified by using subsequently collected data within the Bayesian analysis. We adopted the data registered by the surface detector of the Pierre Auger Observatory since January 1, 2010 up to March 31, 2014 [22] (here period $C$–$D$), after the original Cen A signal was identified [20]. The arrival directions of events with energies above 53 EeV and zenith angles up to 60° were taken from Table A1 in Ref. [22]. Based on the previous findings [20], we assumed a circular region with an angular radius of 18°, located around the position of Cen A ($\alpha = 201.4°, \delta = -43.0°$). The basic characteristics of the Cen A region and the numbers of events collected in the examined period are summarized in the last but one row in Table 1.

In Fig. 3, we give an example of most unbiased information on the highest energy cosmic rays associated with the preselected Cen A zone, which can be derived from the latest data [22]. In this figure, the posterior distributions for the ratio $\beta$ and corresponding credible intervals at a $3\sigma$ level of confidence are shown for three kinds of uninformative prior distributions. We distinguish for unconditional distributions ($\beta \geq 0$) and distributions conditioned on a non-negative source rate in the on-source region ($\beta \geq 1$). Credible intervals for the source fluxes $j$ at a $1\sigma$ level of confidence are depicted in Fig. 4 as functions of the common shape parameter $s \in \langle 0, 1 \rangle$ ($\gamma \to 1$).

The Bayesian inference indicates that the presence of the source in the originally selected Cen A region is less likely than its absence therein when observations since 2010 are considered, i.e. $P^- \geq 0.50$ ($S_{\text{B}} \leq 0$) for almost all prior options, for the Cen A flux see Fig. 4. This conclusion agrees with the classical results based on asymptotic techniques, see Table 1. Hence, the conditional distributions for the ratio $\beta$, $f_\beta^+(x)$ shown in Fig. 3, poorly reflect reality.

The absence of the signal registered in the Cen A region in the latest observation can be quantified using the waiting time for one next Cen A event, see Section 2.2. It is found in a marked difference between unconditional and conditional predictions that disqualifies the latter option, see Fig. 5. Based on this data, over fifty events should be needed in order that the new one was identified in the Cen A region at a 90% level of confidence. Using the method of Section 2.3, the same is documented in Fig. 7. Namely, it is more likely that the four-years Cen A signal is weaker than the AGN activity measured in two preceding periods (magenta results). Here, the parameter $\xi$ can account for different background zones of Cen A and AGNs emitters and different shapes of their energy spectra, for example.

In this regard, it is worth recalling that the Auger collaboration has lately pointed out that the significance of the excess of events in the angular windows and energy range, as examined in this study, is less than its originally observed value [22]. This was obtained by using a broader set of data collected between January 1, 2004 and March 31, 2014, including events with zenith angles up to 80°, when the hypothesis of isotropy was tested. The most significant departure from isotropy in the available set of data was reported for events with energies beyond 58 EeV and with arrival directions within a circle of an angular radius of 15° centered on Cen A [22].

## 4. Conclusions

We focused on the search for new phenomena, when all relevant characteristics of a source which is suspected of causing observed effects cannot be set in an optimal way. The issue was dealt with in the context of on–off measurements assuming registration of small numbers of events that obey Poisson distributions. For this purpose, the Bayesian way of reasoning was utilized. This approach is not only statistically well justified and intuitively easily interpretable, but also provides readily computable results.

We examined three appropriately chosen on–off variables that store information available from the on–off experiment. In addition to traditionally presented results, we proposed how to utilize observation-based information for predictions and comparisons, focusing on quantification of signal stability.

By using successive measurements, increasing sets of the highest energy events collected at the Pierre Auger Observatory were examined. For comparison, also directional data reported by the Telescope Array was considered. Using the recent Auger observations, we summarized the outputs accessible in the proposed approach. We discussed the extent to which the comparison of on–off measurements may help when searching for cosmic ray sources.

## Acknowledgments

We would like to acknowledge and thank our colleagues from the Pierre Auger Collaboration for many helpful discussions and, especially, for the tremendous work with data, its analysis and interpretation from which we benefit. We thank an anonymous reviewer for valuable comments and suggestions that helped us improve the presentation of this paper. This work was supported by the Czech Science Foundation grant 14-17501S. The research of J.N. was partly supported by the Czech Science Foundation under project GACR P103/12/G084.

## Appendix A. Source detection

Using the posterior distribution for the difference, see Eqs. (3) and (4), the probability for the absence of a source in the on-source region is [1]

$$P^- = I_{\frac{\rho}{1+\rho}}(p, q), \tag{A.1}$$

where $p = n_{\text{on}} + s_p$, $q = n_{\text{off}} + s_q$, $\rho = \alpha\gamma_p/\gamma_q$ and $I_x(a, b)$ denotes the regularized incomplete Beta function [23]. Using other on–off variables, we obtain after straightforward calculations

$$P^- = P(\tau \leq \lambda_\tau) = \int_{-\infty}^{0} f_\delta(x)\text{d}x = \int_0^1 f_\beta(x)\text{d}x = \int_0^{\frac{\alpha}{1+\alpha}} f_\omega(x)\text{d}x, \tag{A.2}$$

where $\tau = \delta, \beta, \omega$, while $\lambda_\tau = 0, 1, \frac{\alpha}{1+\alpha}$ denotes the balance value for the difference, ratio and proportion, respectively. The probability of the presence of a source in the on-source zone is $P^+ = 1 - P^-$. When viewed in terms of a normal variate with zero mean and unit variance, the probability $P^-$ is converted to a Bayesian significance $S_\mathrm{B}$.

## Appendix B. Credible intervals

For the shortest credible interval, $\langle \tau_-, \tau_+ \rangle$, that contains the on–off variable with a probability $P$, one has to solve numerically ($\tau = \delta, \beta, \omega$)

$$P = \int_{\tau_-}^{\tau_+} f_\tau(x)\mathrm{d}x, \qquad f_\tau(\tau_-) = f_\tau(\tau_+), \tag{B.1}$$

under the indicated condition on endpoints of the corresponding probability density function $f_\tau(x)$. An upper bound for the source intensity, $\tau_+$, is determined numerically using the integral in Eq. (B.1) where we put $\tau_- \to -\infty, 0$ or 0 for $\tau = \delta, \beta$ or $\omega$, respectively, and relax the indicated limitation on endpoints of $f_\tau(x)$. For a non-negative source intensity ($\mu_\mathrm{on} \geq \mu_\mathrm{b}$), see Appendix C, credible intervals are derived by putting $f_\tau(x) \to f_\tau^+(x)$ into Eq. (B.1) while we set 0,1 or $\frac{\alpha}{1+\alpha} \leq \tau_- < \tau_+ < +\infty, +\infty$ or 1, respectively. For upper bounds for a know source we set directly $\tau_- = 0, 1$ or $\frac{\alpha}{1+\alpha}$ and relax the limitation on endpoints of $f_\tau^+(x)$.

## Appendix C. Known source

In a variety of problems we know with certainty that an active source is present in the on-source region or at least we have a good indication that it may be assumed. This issue is encountered when searching for accompanying radiation from already identified emitters, for example. When the mean event rate in the on-source zone can only increase beyond what is expected from background, the corresponding probability distributions are derived conditioning on the non-negative values of the difference of the on-source and background means, i.e. $\mu_\mathrm{on} \geq \mu_\mathrm{b} = \alpha\mu_\mathrm{off}$ or, alternatively, $\tau \geq \lambda_\tau$ where $\tau = \delta, \beta, \omega$ and $\lambda_\tau = 0, 1, \frac{\alpha}{1+\alpha}$ for the difference, ratio and proportion, respectively. For the conditional distributions we have

$$f_\tau^+(x) = f_\tau(x | x \geq \lambda_\tau) = \frac{f_\tau(x)}{P^+}, \quad x \geq \lambda_\tau, \tag{C.1}$$

where the Bayesian probability of the presence of a source in the on-source zone, $P^+ = 1 - P^- = I_{\frac{1}{1+\rho}}(q, p)$, follows from Eq. (A.1). Note that if the probability $P^+$ approaches one, when it is exceedingly likely that the source contributes to the intensity detected in the on-source zone, the on–off problem is well described in the unconditional regime, since $f_\tau^+(x)$ tends to $f_\tau(x)$ in the domain where $x \geq \lambda_\tau$.

We recall that, by this construction, we obtain results which were derived in another way [2,3,6–9], assuming that the source and background rates are non-negative, i.e. $\mu_\mathrm{s} = \mu_\mathrm{on} - \mu_\mathrm{b} \geq 0$ and $\mu_\mathrm{b} = \alpha\mu_\mathrm{off} \geq 0$, for more information see Ref. [1]. Specifically, in the context of the on–off problem, the use of the proportion $\omega$ with the Jeffreys' prior distributions was advocated in Ref. [8]. In our scheme, substituting the corresponding parameters ($p = n_\mathrm{on} + \frac{1}{2}$, $q = n_\mathrm{off} + \frac{1}{2}$ and $\gamma_p = \gamma_q \to 1$) into Eqs. (8) and (C.1), the posterior $\omega$-distribution written in Eq. (27) in Ref. [8] is recovered.

## Appendix D. Known background

The probability distributions of examined variables are further simplified in the case of a known background. Such a simplification may be used, for example, when searching for sources of cosmic rays in a small on-source region ($0 < \alpha \ll 1$) complemented by a much larger off-source zone which is comprised of the remaining part of the sky within the field of view of the experiment, where $n_\mathrm{off} \gg 1$. Then, the number of background events observed in the on-source zone follow approximately the Poisson distribution with an estimated mean parameter $\mu_\mathrm{b} = \alpha\mu_\mathrm{off} \approx \alpha n_\mathrm{off}$, since its estimated variance is negligible, $\sigma^2(\alpha n_\mathrm{off}) \approx \alpha^2 n_\mathrm{off} \ll \mu_\mathrm{b}^2$. Another example is the analysis of a counting experiment that utilizes a constant background rate estimated based on modeling considerations.

In such a case, we easily obtain $\mu_\mathrm{on} = (\delta + \mu_\mathrm{b}) \sim \mathrm{Ga}(p, \gamma_p)$ [1] and the ratio $\beta = (\mu_\mathrm{on}/\mu_\mathrm{b}) \sim \mathrm{Ga}(p, \gamma_p\mu_\mathrm{b})$, where $p = n_\mathrm{on} + s_p$. The proportion is given by the transformation $\omega = (\alpha\beta)/(1 + \alpha\beta)$. In summary, the probability density functions of all on–off variables are, respectively,

$$h_\delta(x) = \frac{\gamma_p^p}{\Gamma(p)}(x + \mu_\mathrm{b})^{p-1} e^{-\gamma_p(x+\mu_\mathrm{b})}, \quad x \geq -\mu_\mathrm{b}, \tag{D.1}$$

$$h_\beta(x) = \frac{(\gamma_p\mu_\mathrm{b})^p}{\Gamma(p)} x^{p-1} e^{-\gamma_p\mu_\mathrm{b}x}, \quad x \geq 0, \tag{D.2}$$

and

$$h_\omega(x) = \frac{(\gamma_p\mu_\mathrm{off})^p}{\Gamma(p)} \frac{x^{p-1}}{(1-x)^{p+1}} e^{-\frac{\gamma_p\mu_\mathrm{off}x}{1-x}}, \quad x \in \langle 0, 1 \rangle. \tag{D.3}$$

In addition, assuming non-negative source rate in the on-source region, $\mu_\mathrm{on} \geq \mu_\mathrm{b}$ (i.e. $\delta \geq 0$, $\beta \geq 1$ or $\frac{\alpha}{1+\alpha} \leq \omega \leq 1$), we have for the corresponding probability density functions

$$h_\tau^+(x) = \frac{h_\tau(x)}{R^+}, \quad x \geq \lambda_\tau, \tag{D.4}$$

where $\tau = \delta, \beta, \omega$, while $\lambda_\tau = 0, 1, \frac{\alpha}{1+\alpha}$, and $R^+$ is the probability of the presence of a source in the on-source region provided a constant background mean is used,[1] i.e.

$$R^+ = \int_0^\infty h_\delta(x)\mathrm{d}x = \int_1^\infty h_\beta(x)\mathrm{d}x = \int_{\frac{\alpha}{1+\alpha}}^1 h_\omega(x)\mathrm{d}x = \frac{\Gamma(p, \gamma_p\mu_\mathrm{b})}{\Gamma(p)}, \tag{D.5}$$

where $\Gamma(a, x) = \int_x^\infty t^{a-1}e^{-t}\mathrm{d}t$ is the upper incomplete Gamma function. It is useful to know that $R^+(p, x) = \frac{\Gamma(p,x)}{\Gamma(p)} = e^{-x}\sum_{k=0}^{p-1}\frac{x^k}{k!}$ for integer values of $p$.

Notice that for $\gamma_p \to 1$, $R^- = 1 - R^+$ is the $p$-value obtained in the classical framework, when the background hypothesis (i.e. $\mu_\mathrm{on} \leq \mu_\mathrm{b}$) is tested against the alternative of a source presence in the on-source zone ($\mu_\mathrm{on} > \mu_\mathrm{b}$) for the Poisson sampling process [15].

## Appendix E. Comparison with known backgrounds

When fluctuations in the background are completely disregarded, see Appendix D, the probabilities of inconsistency introduced in Section 2.3 can be expressed explicitly. We assume two independent observations, marked by indices 1 and 2. If only non negative integer values of relevant shape parameters ($s_{p_1}$ and $s_{p_2}$) are considered, the integration in Eq. (14) is easily performed using the posterior distributions given in Eq. (D.1). Then, the probability of inconsistency between source fluxes when the on-source exposures ($a_1$ and $a_2$) are known, see Eq. (13), can be written in a compact formula ($p_1 = n_{\mathrm{on}_1} + s_{p_1}, p_2 = n_{\mathrm{on}_2} + s_{p_2}$, $p_1, p_2 \in N$)

$$\begin{aligned} P(j_1 < j_2) = {} & \frac{e^{-u}\ v^{p_2}}{(1+v)^{p_1+p_2}} \sum_{k=1}^{p_2}\sum_{i=k}^{p_2} \binom{p_1 + p_2 - i - 1}{p_2 - i} \\ & \times \frac{u^{i-k}}{(i-k)!}\left(\frac{1+v}{v}\right)^i R_i(v). \end{aligned} \tag{E.1}$$

Here, $u = \gamma_{p_2}\mu_{\mathrm{b}_2} - v\,\gamma_{p_1}\mu_{\mathrm{b}_1} \geq 0$ depends on the known background rates, $\mu_{\mathrm{b}_1}$ and $\mu_{\mathrm{b}_2}$, $v = (\gamma_{p_2}a_2)/(\gamma_{p_1}a_1)$ depends on the ratio of two on-source exposures, $\gamma_{p_1}$ and $\gamma_{p_2}$ denote the prior rates of the on-source means and $R_i(v) = 1$ for the unconditional $\delta$-distributions given in Eq. (D.1), while

$$R_i(v) = \frac{R^+(p_1 + p_2 - i, (1+v)\gamma_{p_1}\mu_{\mathrm{b}_1})}{R^+(p_1, \gamma_{p_1}\mu_{\mathrm{b}_1})\ R^+(p_2, \gamma_{p_2}\mu_{\mathrm{b}_2})}, \tag{E.2}$$

for the conditional $\delta$-distributions, see Eqs. (D.4) and (D.5). In Eq. (E.1) we compare source fluxes provided $u \geq 0$. If $u < 0$, we simply exchange measurements, using $P(j_1 < j_2) = 1 - P(j_2 < j_1)$.

[1] Note that there are typographical errors in Eqs. (26) and (27) in Ref. [1]. There should be $\Gamma(p, \gamma_p\mu_\mathrm{b})$ instead of $\Gamma(p, \mu_\mathrm{b})$.

In a similar way and under the same conditions, we can compare two independent on–off measurements through the ratios $\beta_1$ and $\beta_2$ when background uncertainties are not considered. Using the parameter $\xi$, the probability of inconsistency between two ratios (see Eq. (14)) is written $(p_1, p_2 \in N)$

$$P(\beta_1 < \xi\beta_2) = \frac{w^{p_2}}{(1+w)^{p_1+p_2}} \sum_{k=1}^{p_2} \binom{p_1+p_2-k-1}{p_2-k} \left(\frac{1+w}{w}\right)^k R_k(w), \quad \text{(E.3)}$$

where $w = \xi^{-1}(\gamma_{p_2}\mu_{b_2})/(\gamma_{p_1}\mu_{b_1})$ and $R_k(w) = 1$ for the unconditional $\beta$-distributions (Eq. (D.2)) and for the conditional ones (Eqs. (D.4) and (D.5)) it is written in Eq. (E.2). The formula in Eq. (E.3) holds for $\xi \leq 1$. If $\xi > 1$, we use $P(\beta_1 < \xi\beta_2) = 1 - P(\beta_2 < \xi^{-1}\beta_1)$.

## References

[1] D. Nosek, J. Nosková, Nucl. Instrum. Methods A 820 (2016) 23.
[2] M.L. Knoetig, Astrophys. J. 790 (2014) 106.
[3] D. Casadei, Astrophys. J. 798 (2015) 5.
[4] O. Helene, Nucl. Instrum. Methods 212 (1983) 319.
[5] O. Helene, Nucl. Instrum. Methods 228 (1984) 120.
[6] H.B. Prosper, Nucl. Instrum. Methods A 241 (1985) 236.
[7] H.B. Prosper, Phys. Rev. 37 (1988) 1153.
[8] S. Gillesen, H.L. Harney, Astron. Astrophys. 430 (2005) 355.
[9] P. Gregory, Bayesian Logical Data Analysis for the Physical Sciences, Cambridge University Press, Cambridge, 2005 (Chapter 14).
[10] S.S. Wilks, Ann. Math. Stat. 9 (1938) 60.
[11] T.P. Li, Y.Q. Ma, Astrophys. J. 272 (1983) 317.
[12] R.D. Cousins, Nucl. Instrum. Methods A 417 (1998) 391.
[13] G. Cowan, K. Cranmer, E. Gross, O. Vitells, Eur. Phys. J. C 71 (2011) 1554; G. Cowan, K. Cranmer, E. Gross, O. Vitells, Eur. Phys. J. C 73 (2013) 2501.
[14] G.J. Feldman, R.D. Cousin, Phys. Rev. D 57 (1998) 3873.
[15] R.D. Cousins, J.T. Linnemann, J. Tucker, Nucl. Instrum. Methods A 595 (2008) 480.
[16] W.A. Rolke, A.M. López, J. Conrad, Nucl. Instrum. Methods A 551 (2005) 493.
[17] S. Algeri, J. Conrad, D.A. van Dyk, Mon. Not. R. Astron. Soc. 458 (2016) L84.
[18] J. Abraham, et al. (The Pierre Auger Collaboration), Science 318 (2007) 928.
[19] J. Abraham, et al. (The Pierre Auger Collaboration), Astropart. Phys. 29 (2008) 188.
[20] P. Abreu, et al. (The Pierre Auger Collaboration), Astropart. Phys. 34 (2010) 314.
[21] P. Abreu, et al., J. Cosmol. Astropart. Phys. 06 (2011) 022.
[22] A. Aab, et al. (The Pierre Auger Collaboration), Astrophys. J. 804 (2015) 15.
[23] F.W.J. Olver, D.M. Lozier, R.F. Boisvert, Clark C.W. (Eds.), NIST Handbook of Mathematical Functions, Cambridge University Press, Cambridge, 2010 (Chapters 8 and 13).
[24] B. McDonald, Econometrica 52 (1984) 647.
[25] T. Abu-Zayyad, et al., Astrophys. J. 757 (2012) 26.
[26] T. Abu-Zayyad, et al., Astrophys. J. 777 (2013) 88.
[27] H.C. Tijms, A First Course in Stochastic Models, John Wiley & Sons Ltd., Chichester, 2003 (Chapters 1).
[28] N.L. Johnson, A.W. Kemp, S. Kotz, Univariate Discrete Distributions, John Wiley & Sons, Inc., Hoboken, 2005 (Chapters 5).
[29] K. Kawata, et al. (The telescope array collaboration) in: The 34th International Cosmic Ray Conference, 30 July–6 August 2015, the Hague, The Netherlands.
[30] M.-P. Véron-Cetty, P. Véron, Astron. Astrophys. 455 (2006) 773.
[31] A. Aab, et al. (The Pierre Auger Collaboration), J. Cosmol. Astropart. Phys. 08 (2015) 049.