SCHOOL OF DATA ANALYSIS

SHiP Tracks Recognition

Mikhail Hushchyn, Alenkin Oleg, Andrey Ustyuzhanin

SHiP Tracks Recognition

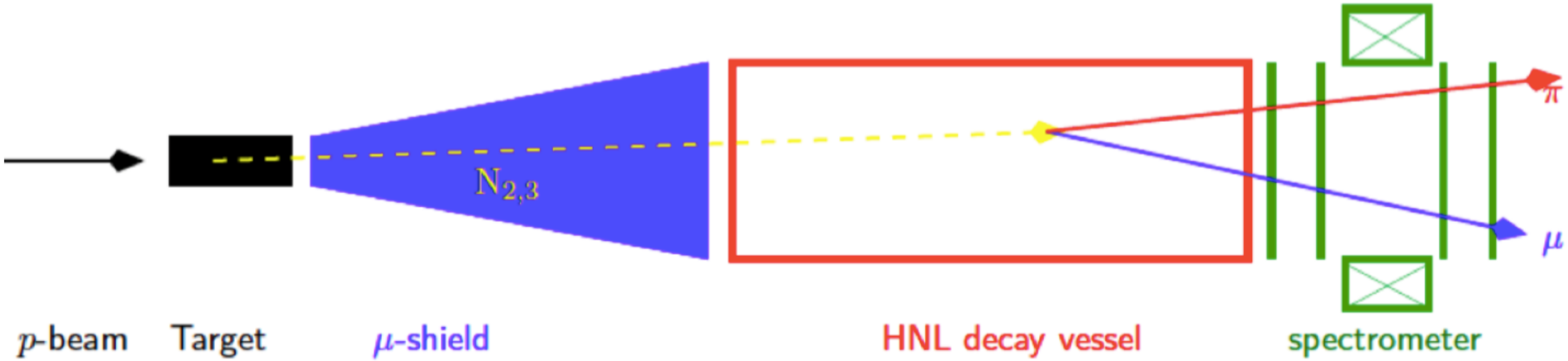
SHiP Experiment

SHiP Experiment

Search for Hidden Particles (SHiP) experiment is aimed at searching for very weakly interacting long lived particles including Heavy Neutral Leptons (HNL).

These particles were predicted by a very large number of recently elaborated models of Hidden Sectors which are capable of accommodating dark matter, neutrino oscillations, and the origin of the full baryon asymmetry in the Universe.

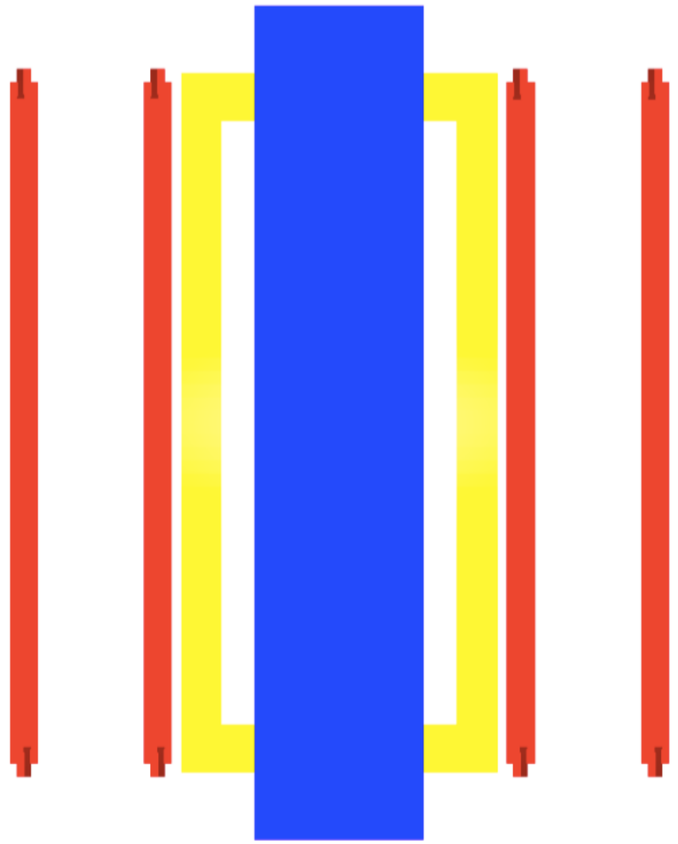
Detector Design



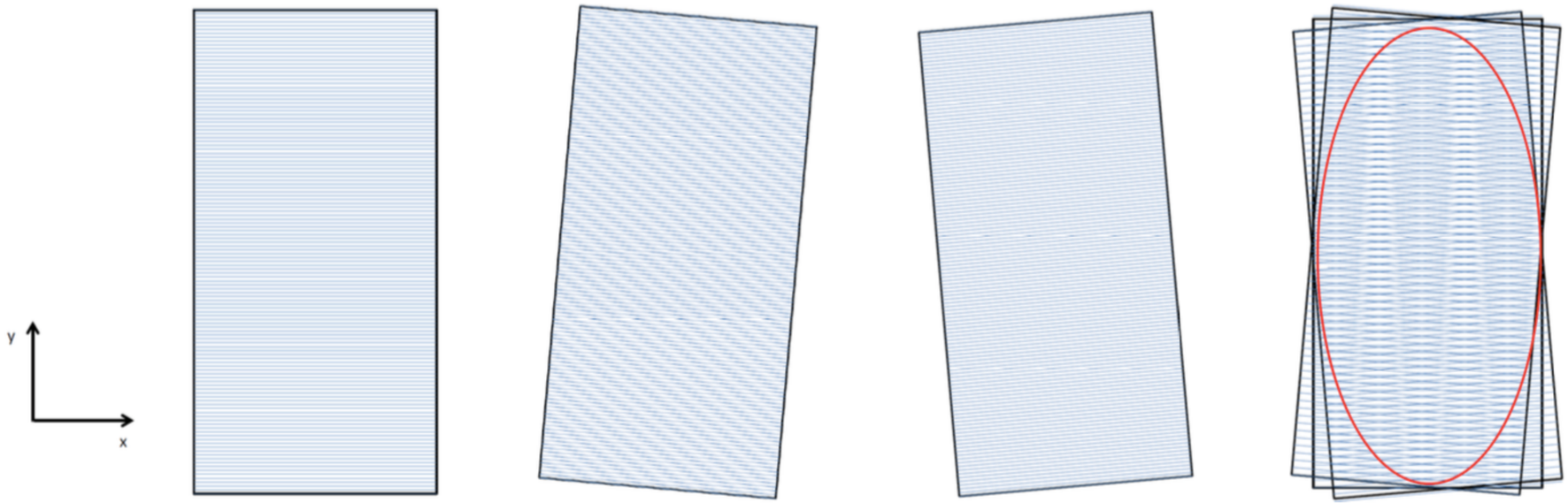
The idea of the experiment is to detect HNL by its decay products, for an example, pion and muon. The products are charged particles. Thus, they can be detected by the SHiP Spectrometer Tracker.

SHiP Spectrometer Tracker

The Spectrometer Tracker consists of the magnet (blue) and 4 stations (red).



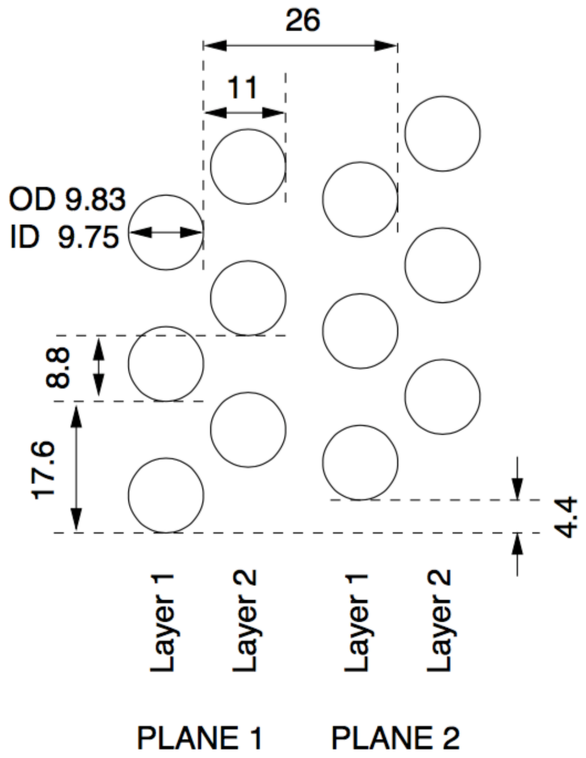
SHiP Spectrometer Tracker



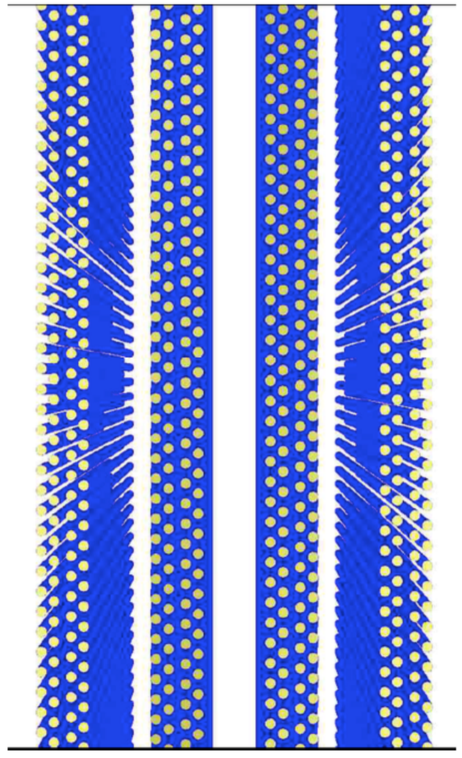
(a) Y view (b) U view (c) V view (d) all views

Each station consists of 4 layers (views): 2 x Y-views, 1 x U-view and 1 x V-view. The U-view is rotated by 5 degree with respect to the y-axis. The V-view is rotated by -5 degree with respect to the y-axis. The views sequence: YUVY.

SHiP Spectrometer Tracker



(a) Views, planes and layers



(b) Straws in the simulation (4 views)

Each view consists of the 4 layers of straw tubes.

SHiP Tracks Recognition

Problem Formulation

Data We Have

- (X, Y, Z) coordinates of a wire's ends for each straw tube
- Which straw tubes are active i.e. which straw tubes detect a particle
- Distance between a particle trajectory and a draw tube's wire

We Need

In this study we are reconstructing tracks of two particles: pion and muon which are products of a HNL decay.

Each event has two target tracks.

Goals:

〉 Reconstruct the particles' trajectories as good as possible.

〉 Determine the tracks momentum based on deflection in the magnet field.

SHiP Tracks Recognition

Reference Set

Selections

- Tracks origin must be inside the decay volume.
- Tracks should not decay inside the tracking station area.
- Tracks must have a hit in the Timing Detector.
- Tracks must have at least one hit in each Tracking Station, inside the acceptance “ellipse“ defined by ((x/245)2 + (y/495)2) >= 1.
- Tracks must be from an HNL decay product.
- The event must have one pion-track and one muon-track.
- Events with more than 500 hits are rejected

Reference Set

- We have 5 000 event.
- 670 (13.4 %) events pass through the selections.

SHiP Tracks Recognition

Baseline

General Scheme

- Tracks reconstruction before the magnet
- Tracks reconstruction after the magnet
- Combination of the tracks before and after the magnet
- Estimation of the particles' momentum and charge based on their deflection in the magnet field

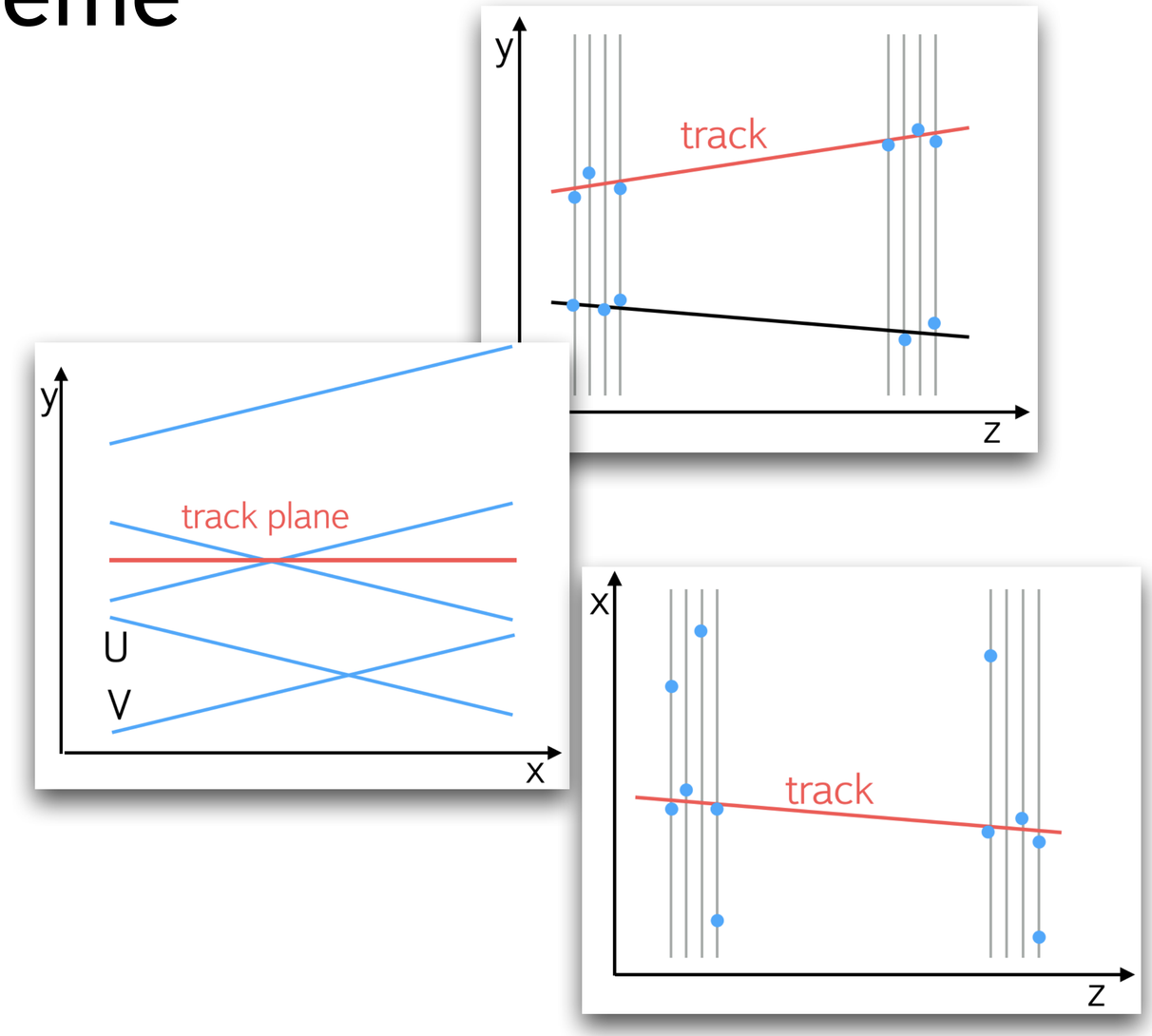
Tracks Reconstruction Scheme

Two 2D projections are used for the 3D tracks reconstruction:

1. Looking for tracks in the 16 layers with horizontal straw tubes (Y-views, y-z plane).

2. For each track found in the Y-views find intersection of the plane defined be the track in y-z plane and wires of the active straw tubes in U,V-views. Find (z, x) coordinated of the intersections.

3. Looking for tracks in 16 layers of U,V-views in x-z plane using (z, x) coordinates of the intersections.



Tracks Reconstruction in Plane

Algorithm:

1. Select start and end hits which don't belong to any recognized tracks before.

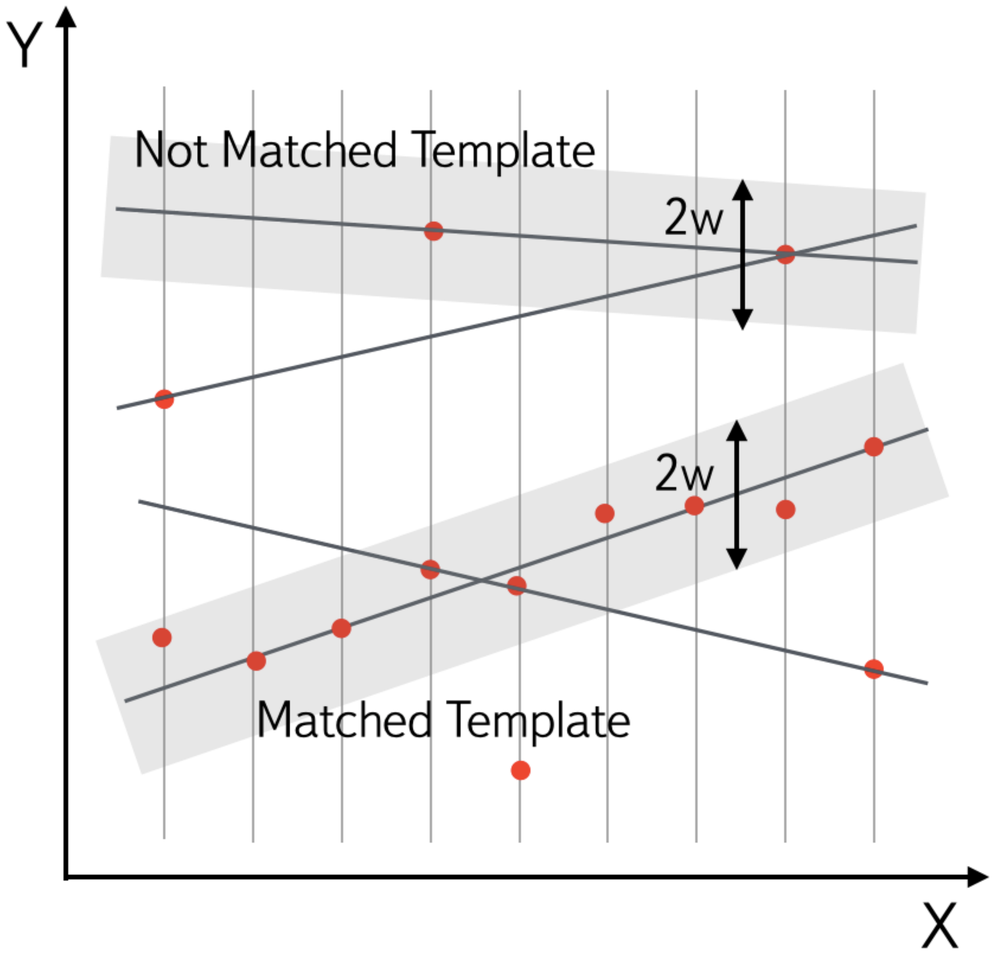
2. Build a line through these hits.

3. Count hits within a window width (w) from the line.

4. If number of the hits is larger than defined number N, mark these hits as recognized track.

5. Repeat 1-4 steps until all pairs of hits are not selected.

6. Repeat 1-5 steps for N = 16, ..., N_min

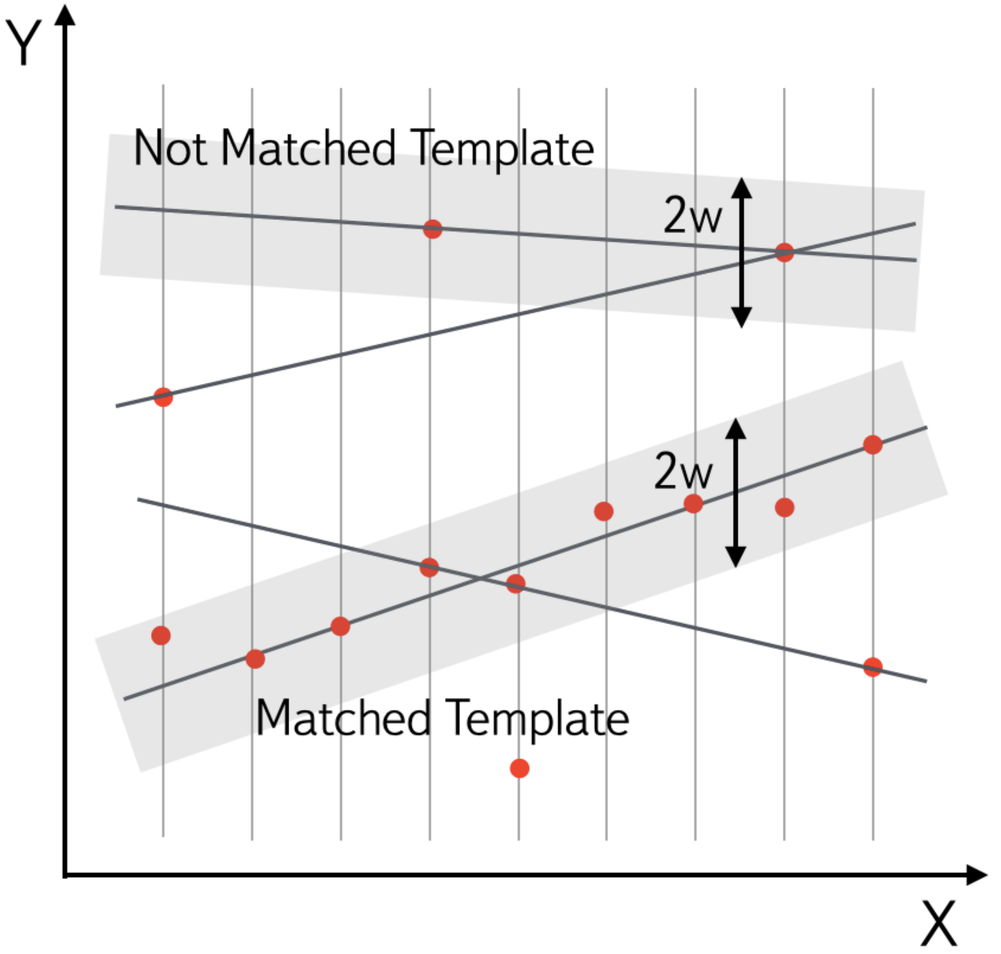


Tracks Reconstruction in Plane

The algorithm's parameters:

〉 w = 0.75, N_min = 7 for Y-views;

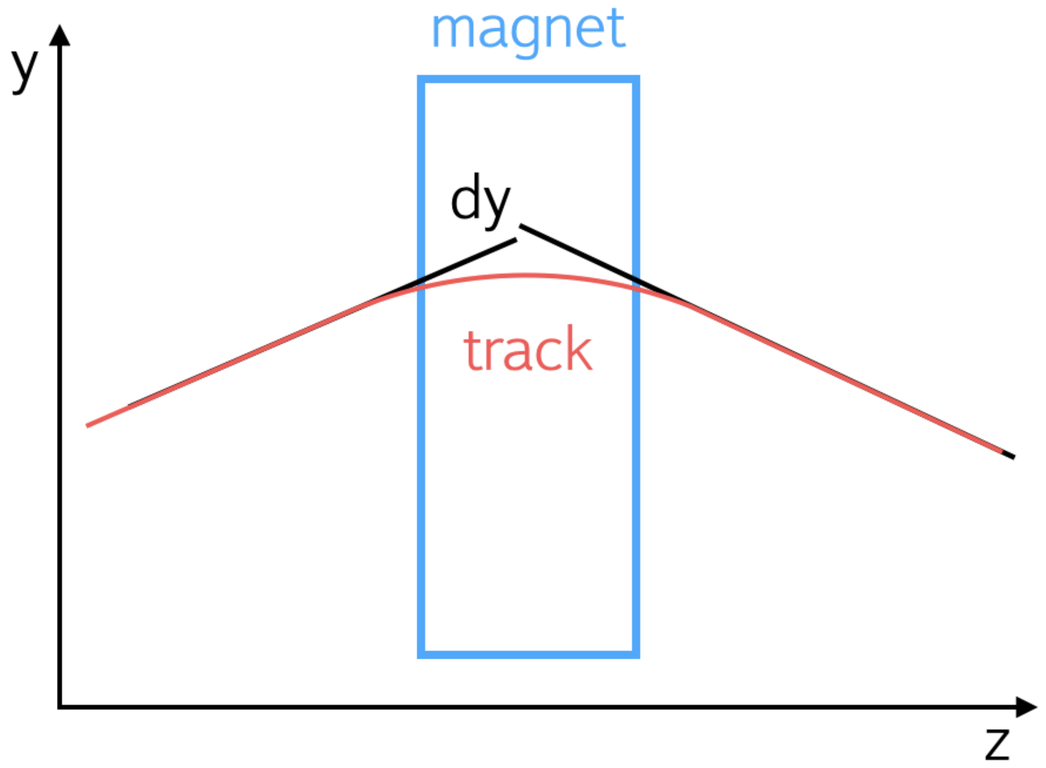
〉 w = 15.0, N_min = 6 for U,V-views;



Tracks Combination

Straight tracks before and after the magnet are extrapolated to the center of the magnet.

If distances dy and dx between the extrapolated tracks are less than 2 cm and 20 cm respectively the tracks denoted as combined.



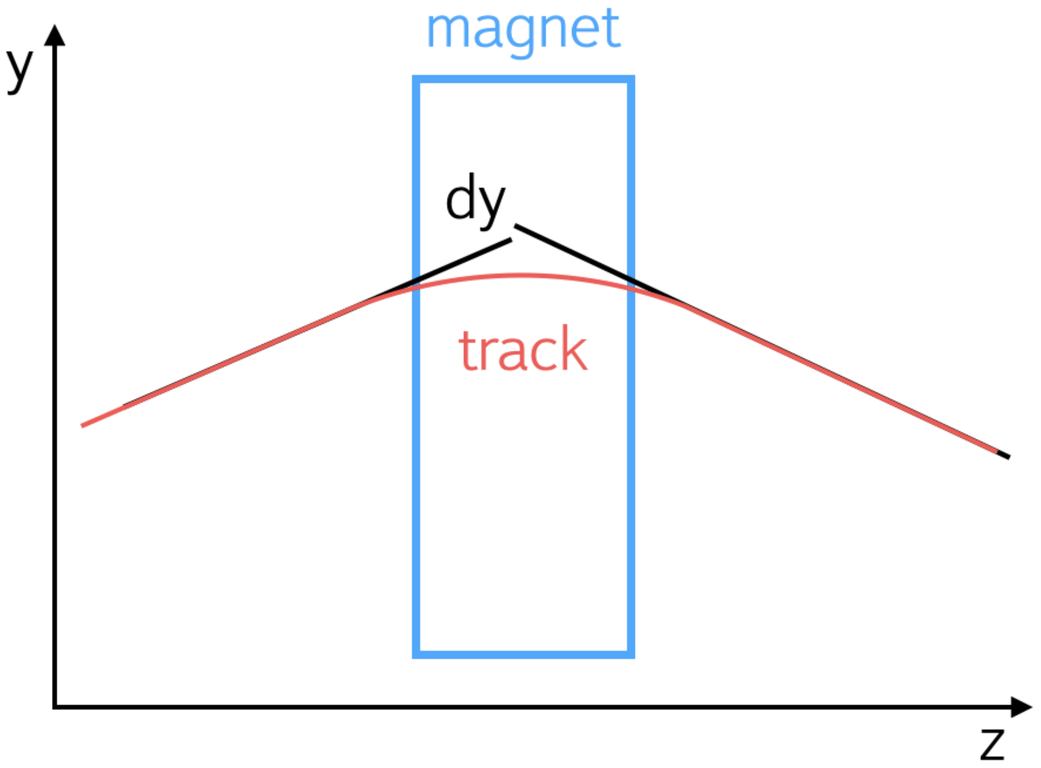
Momentum and Charge Estimation

A trajectory deflection in the magnetic field helps to estimate a particle's momentum:

$$\frac{1}{p} = \frac{Sin(\Delta k_{yz})}{0.3B}$$

where p is momentum value, B is magnetic field inductance, k is the track slope in y-z plane.

Sign of the slope changes helps to estimate charge of the particle.



Quality Metrics

Two following metrics are used to measure the tracks recognition quality:

- Track Efficiency
- Reconstruction Efficiency
- Ghost Rate
- Clone Rate

Quality Metric. Track Efficiency. Hit Matching.

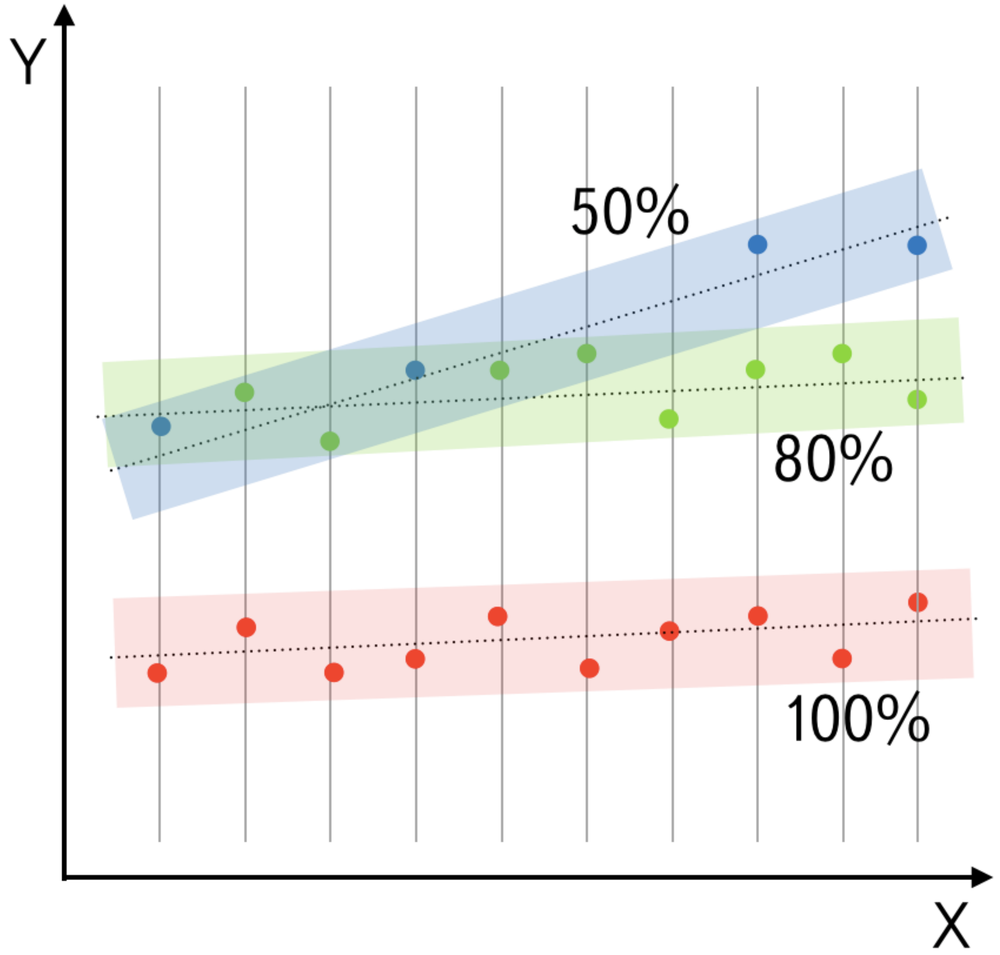
The track finding efficiency is defined as:

$$\epsilon_{track} = \frac{N_{reco\_true\_hits}}{N_{reco\_hits}} * 100\%$$

where N denotes the number of recognized the track's true hits and number of recognized hits respectively.

The track is considered to be reconstructed if its efficiency is higher than, for example, 70%.

This method is stable in the limit of very high track densities.



Quality Metric. Reconstruction Efficiency.

The reconstruction efficiency is defined as:

$$\epsilon_{reco} = \frac{N_{ref}^{reco}}{N_{ref}}$$

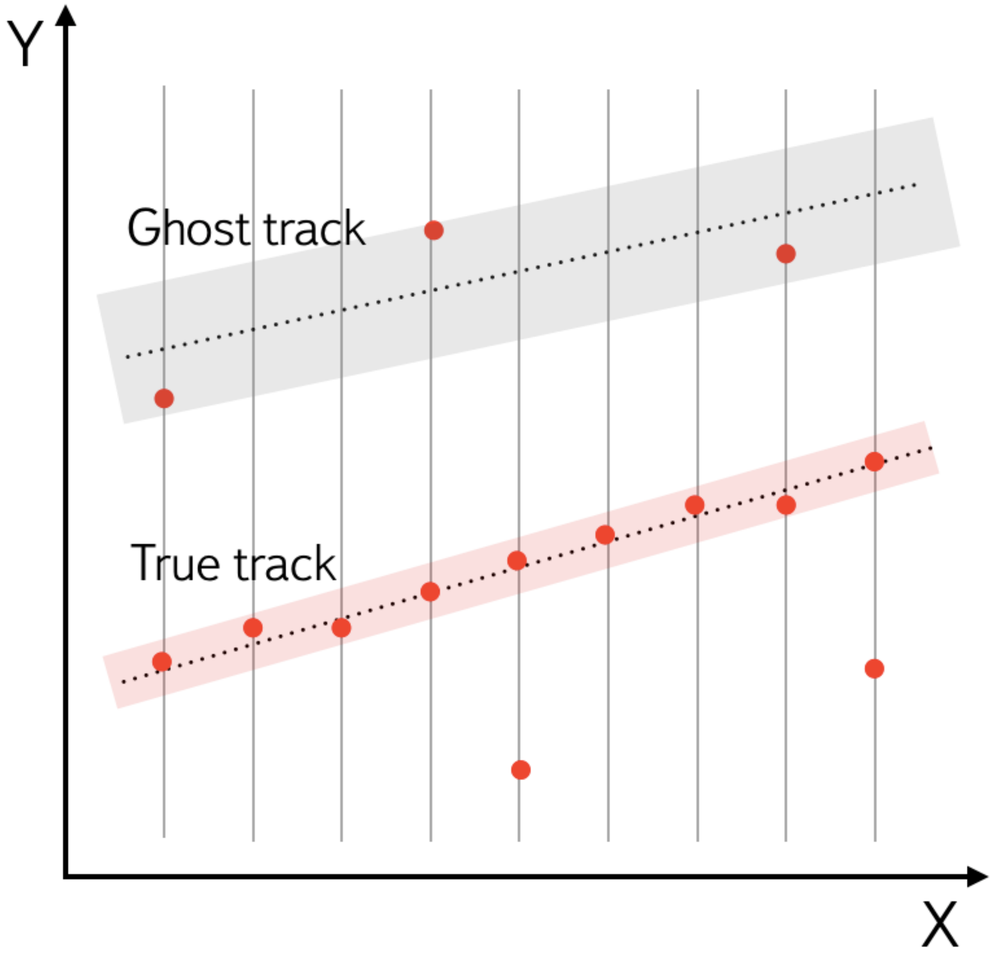
where N_{ref}^{reco} is the number of reference tracks that are reconstructed by at least one track.

Quality Metric. Ghosts.

Ghosts are tracks produced by the pattern recognition algorithm that do not reconstruct any true track within or without the reference set.

A ghost rate is defined as:

$$\epsilon_{ghost} = \frac{N^{ghost}}{N_{ref}}$$



Quality Metric. Clones.

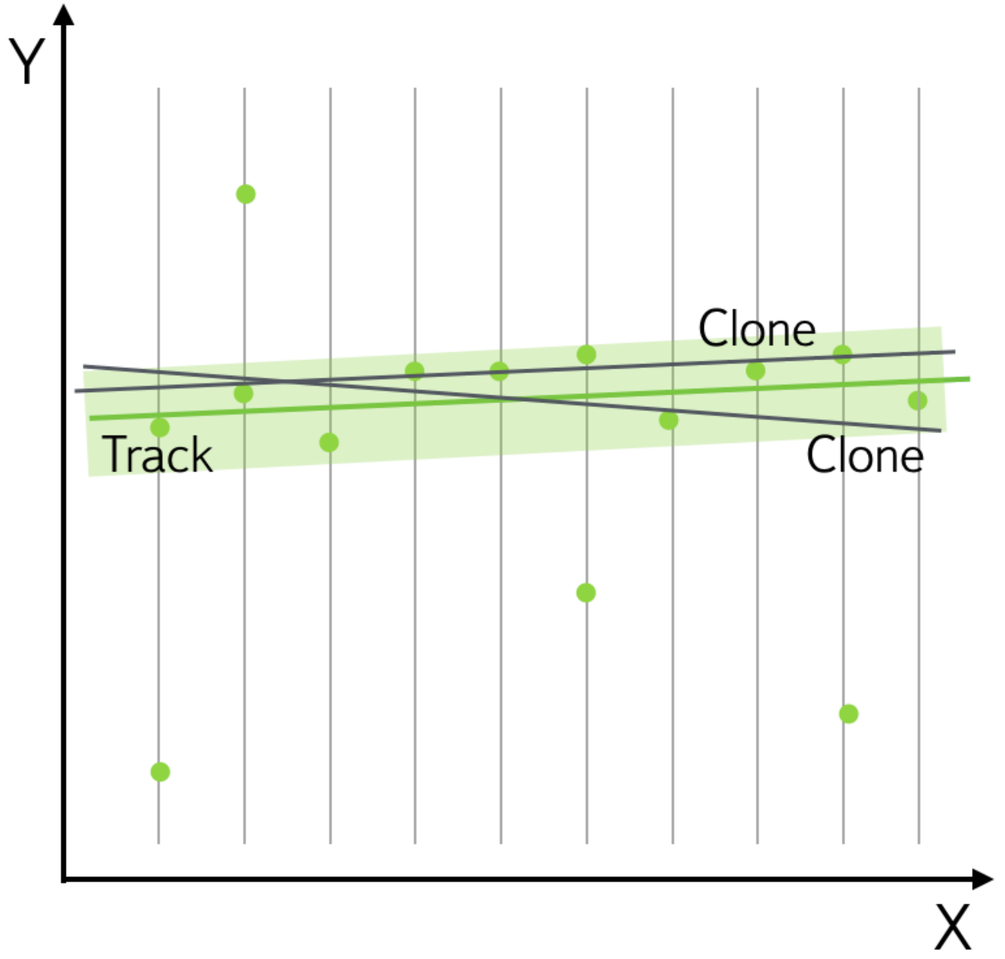
The definitions for efficiency and ghost rate are sensitive to multiple reconstructions of a track. Such redundant reconstructions are sometimes called clones.

For a given track m with N_m^{reco} tracks reconstructing it, the number of clones is

$$N_m^{clone} = \begin{cases} N_m^{reco} - 1, \text{if} N_m^{reco} > 0 \\ 0 \qquad\quad , \text{otherwise} \end{cases}$$

A clone rate then is

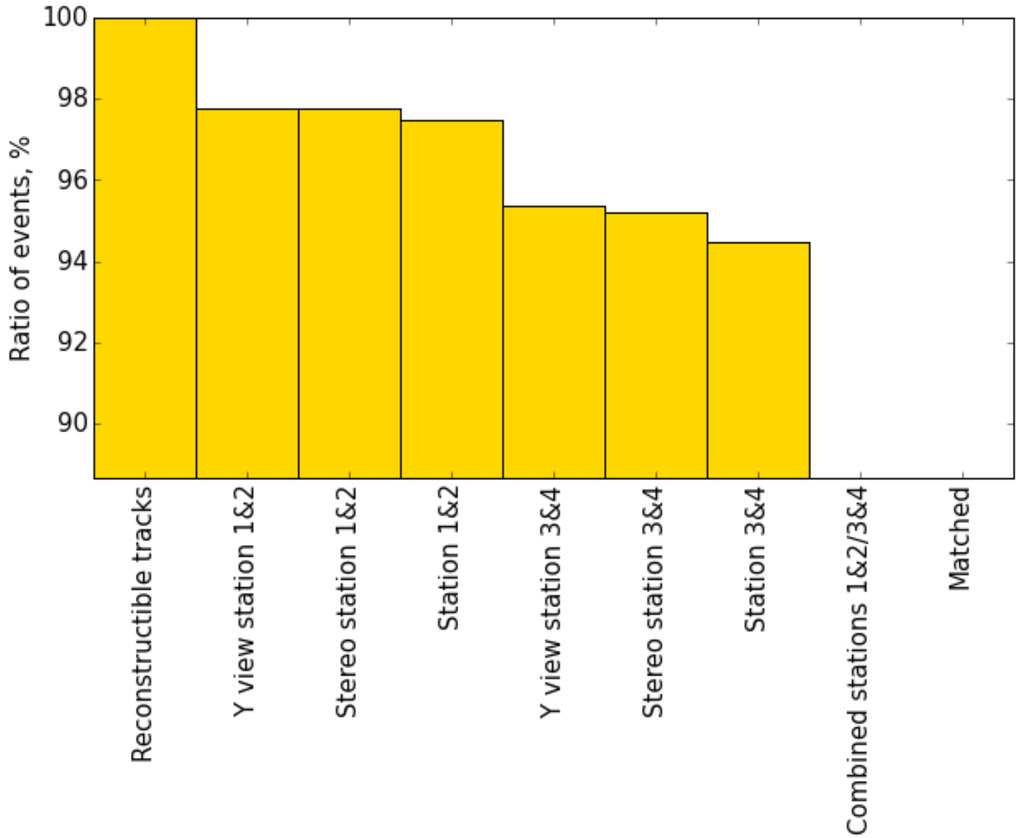
$$\epsilon_{clone} = \frac{\sum_m N_m^{clone}}{N_{ref}}$$



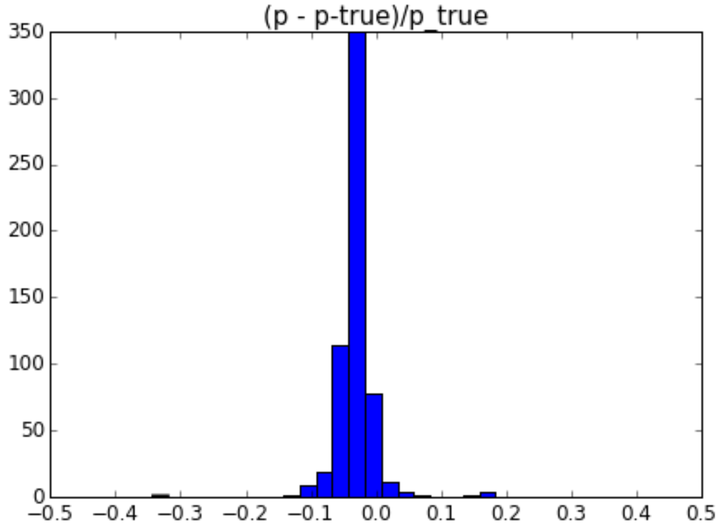
Quality Metrics

	Avg. Tracks Efficiency	Avg. Reconstruction Efficiency	Ghost Rate	Clone Rate
Y-views, stations 1&2	0,989	0,989	0,008	0,002
Stereo-views, stations 1&2	0,975	0,968	0,01	0,001
All-views, stations 1&2	0,982	0,987	0,009	0,002
Y-views, stations 3&4	0,991	0,989	0,017	0,001
Stereo-views, stations 3&4	0,982	0,972	0,022	0,001
All-views, stations 3&4	0,985	0,985	0,017	0,001
Combination	-	0,928	0,003	0,000

Quality Metrics



Efficiency is 88.7 %



Avg. accuracy is 4 %

Time

0.4 sec. / event.

SHiP Tracks Recognition

New Tracks Combination

Idea

Train a classifier to predict that tracks before and after the magnet belong to the same particle.

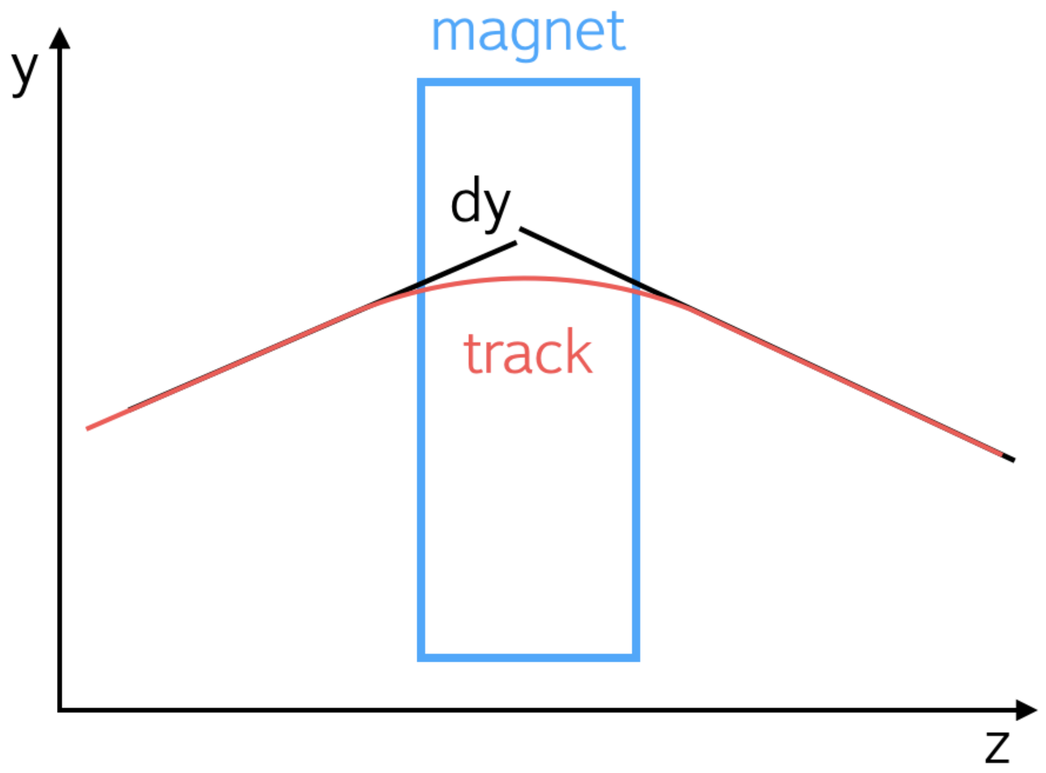
Features

For each pair of tracks the following features were used:

〉 Tracks parameters in y-z and x-z planes before and after the magnet

〉 Differences of the tracks parameters before and after the magnet

〉 Distances dx, dy and dr between track extrapolated to the center of the magnet.



Train Classifier

Gradient Boosting on Trees Classifier (XGBoost) was used.

Easy to train.

ROC AUC is 0.995.

Result

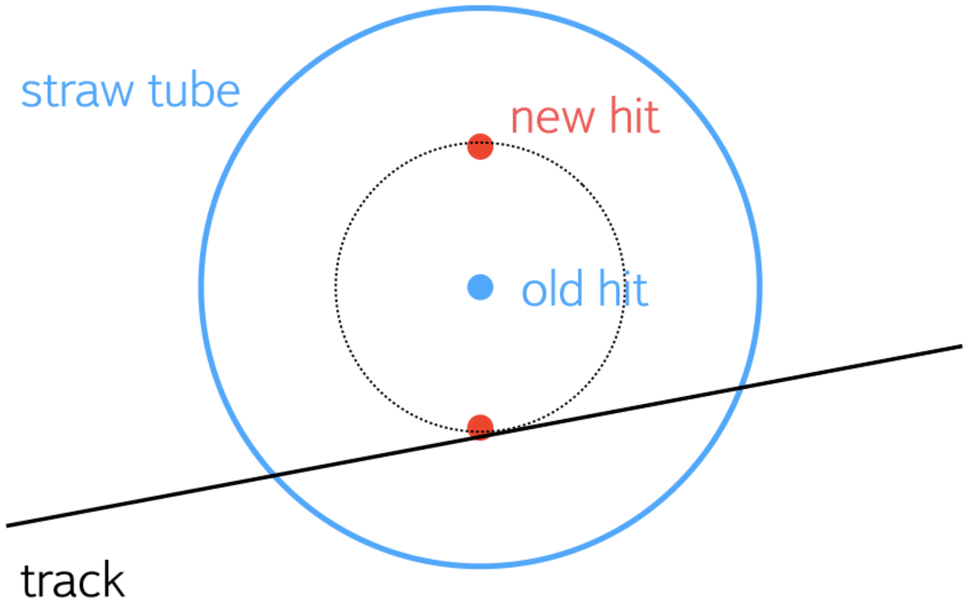
The efficiency per event for the Baseline + New Tracks Combination is 91.8% (+ 3.1%).

SHiP Tracks Recognition

Baseline + Double Hits

Double Hits

Each hit was doubled by using distance from a track to a straw tube’s wire.



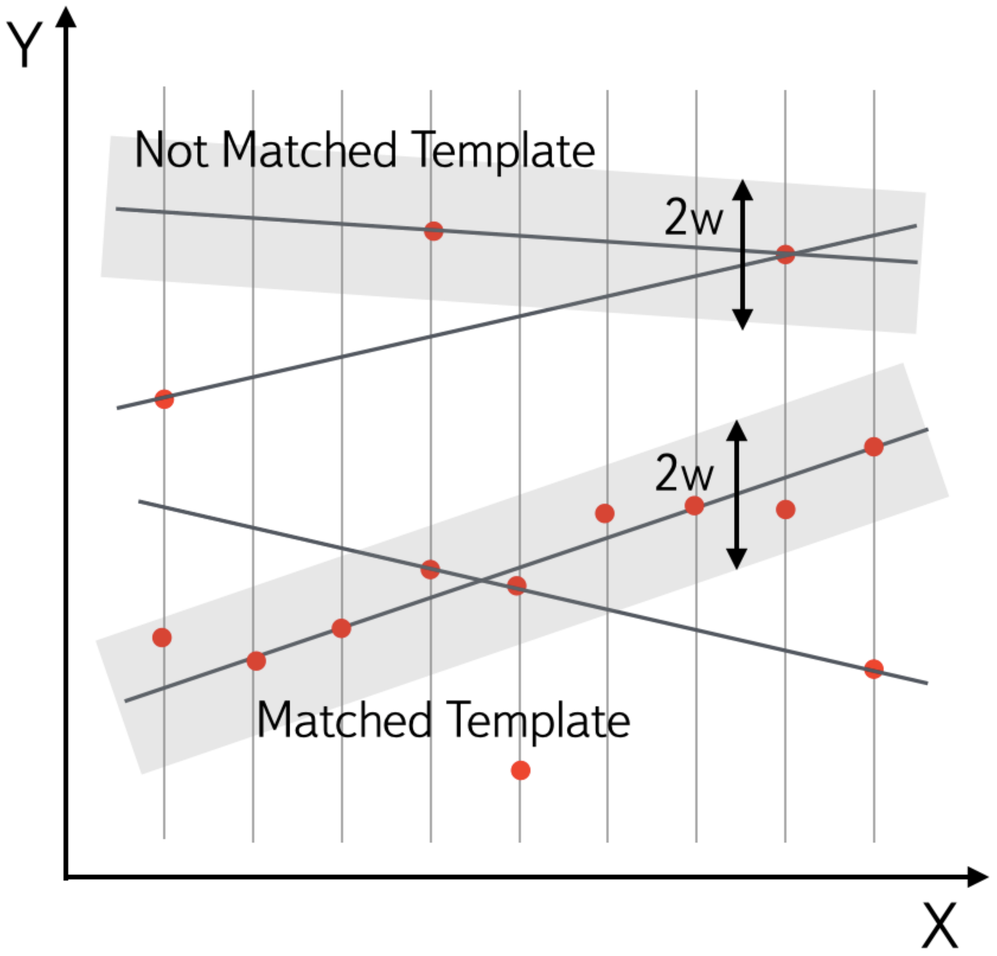
Tracks Reconstruction in Plane

The algorithm's parameters:

〉 w = 0.2, N_min = 7 for Y-views;

〉 w = 5.0, N_min = 6 for U,V-views;

〉 The New Tracks Combination was used;

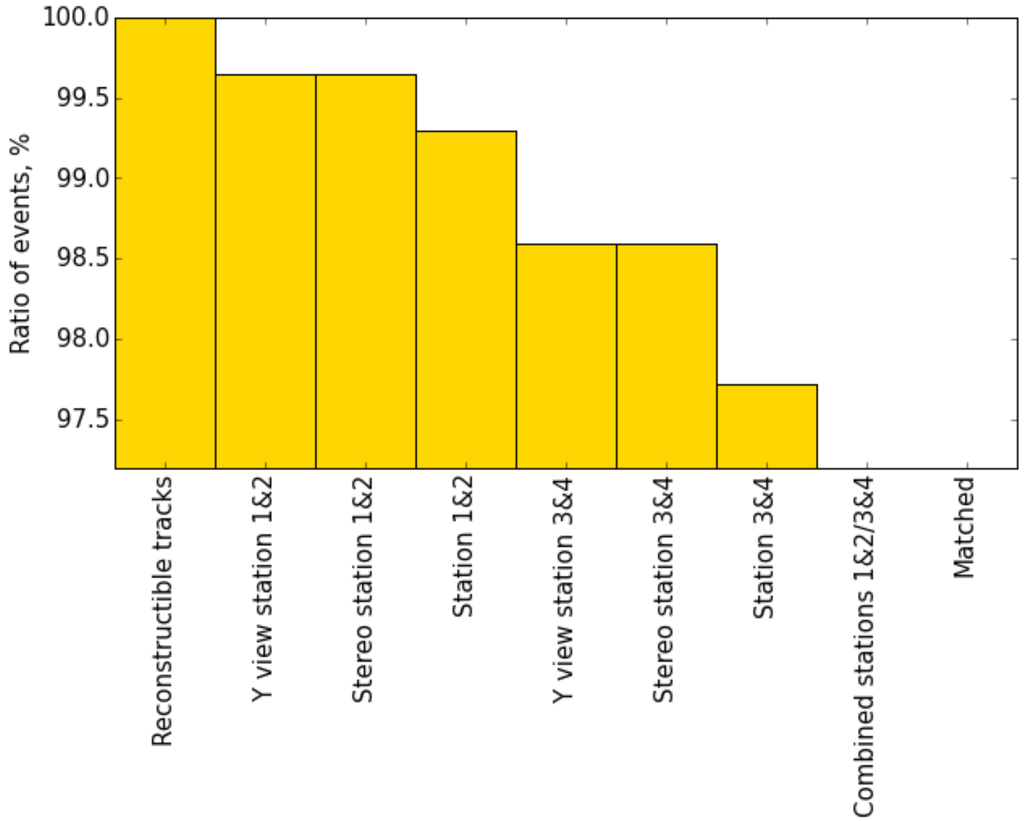


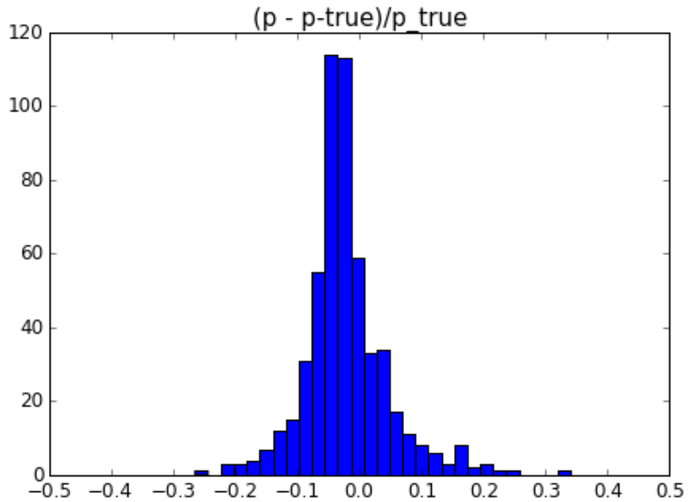
Quality Metrics

	Avg. Tracks Efficiency	Avg. Reconstruction Efficiency	Ghost Rate	Clone Rate
Y-views, stations 1&2	0,995	0,998	0,015	0,002
Stereo-views, stations 1&2	0,990	0,993	0,012	0,004
All-views, stations 1&2	0,990	0,996	0,017	0,001
Y-views, stations 3&4	0,996	0,996	0,050	0,002
Stereo-views, stations 3&4	0,994	0,990	0,044	0,001
All-views, stations 3&4	0,995	0,991	0,050	0,002
Combination	-	0,984	0,014	0,001

Baseline + Double Hits

Quality Metrics





Efficiency is 97.2 %

Avg. accuracy is 6 %

Baseline + Double Hits

Time

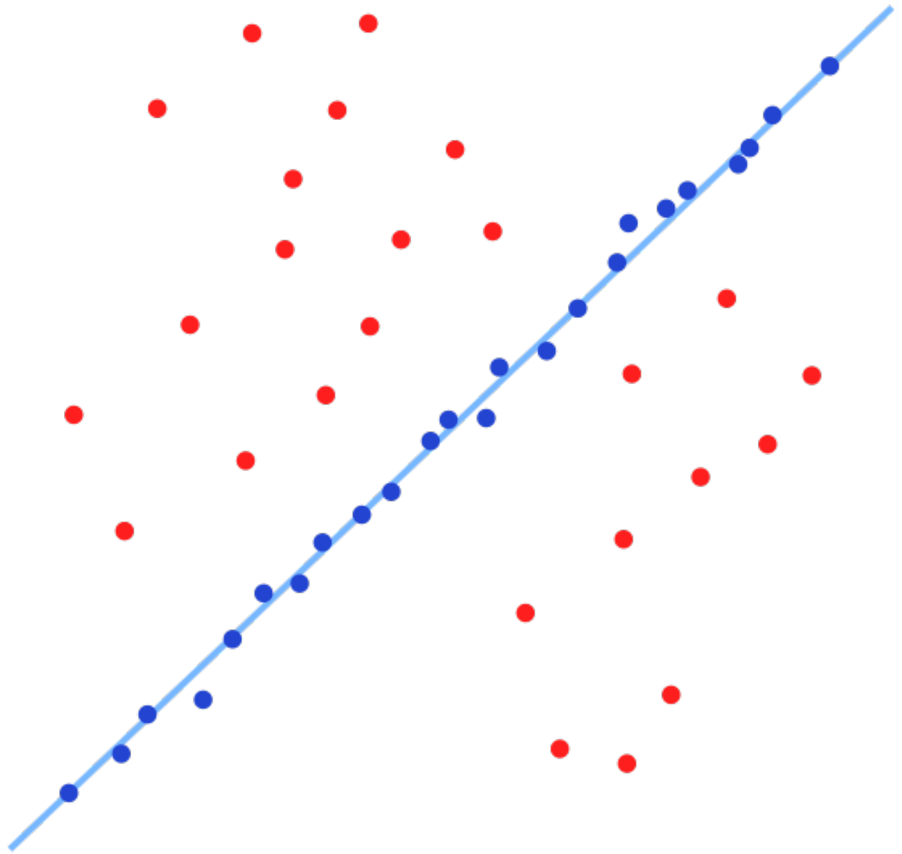
0.9 sec. / event.

SHiP Tracks Recognition

RANSAC

Definition of the RANSAC

RANSAC (RANdom SAmple Consensus) is an iterative method for regression problem with samples contaminated with outliers.



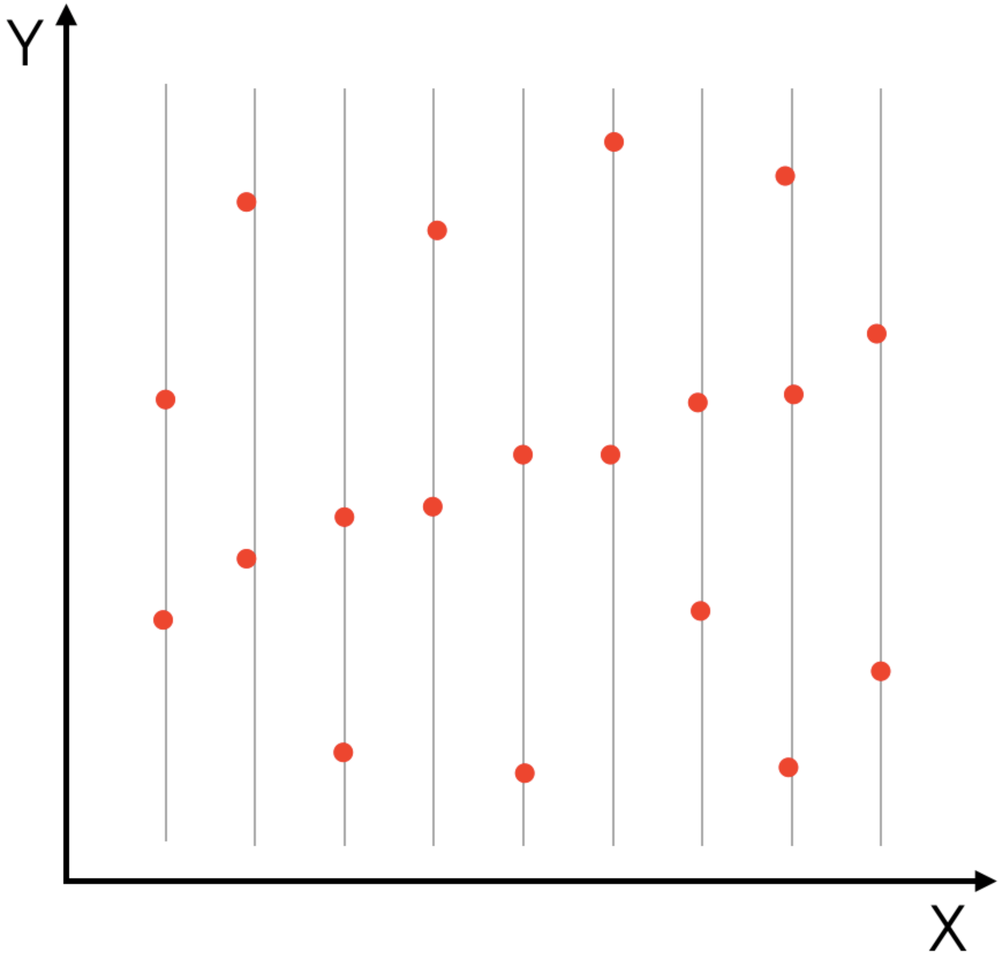
The RANSAC iterates two steps:

1. Fitting a model using a random sample subset.

2. Verification of the model using the entire sample.

RANSAC

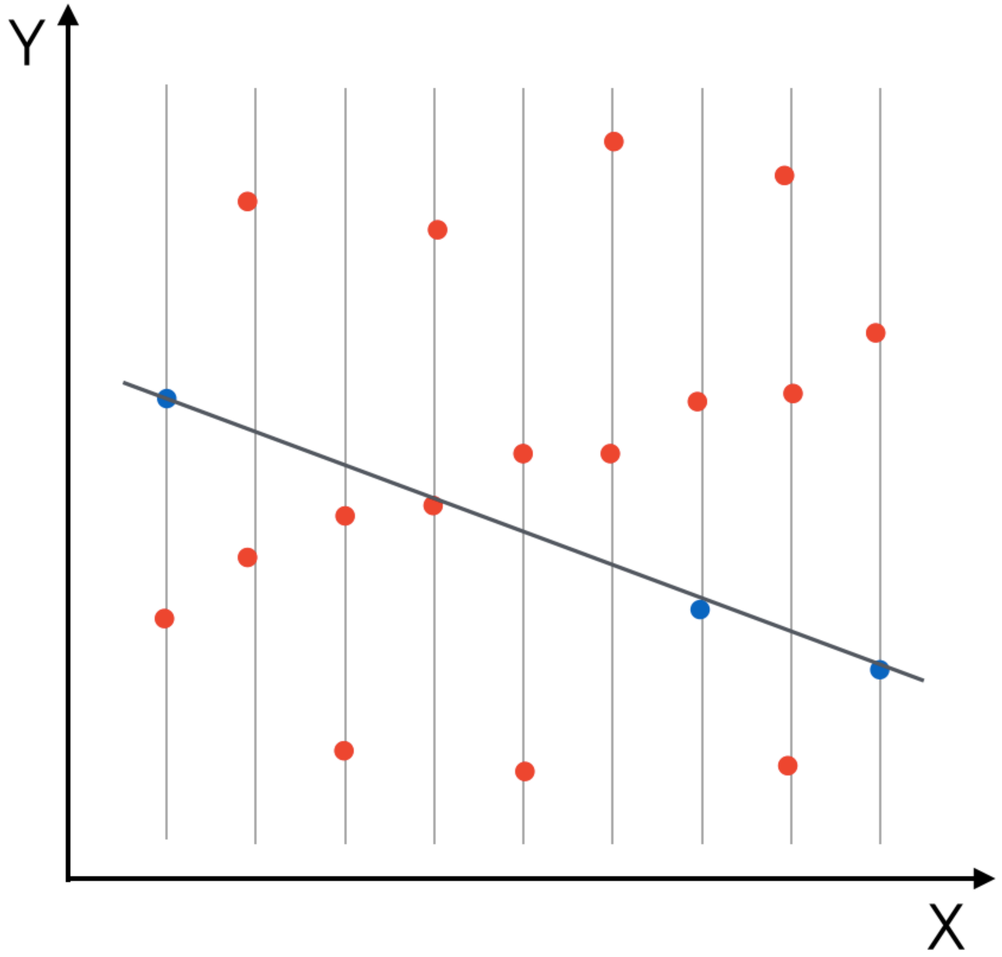
Consider a problem of fitting a linear model with observed data which contains outliers.



RANSAC

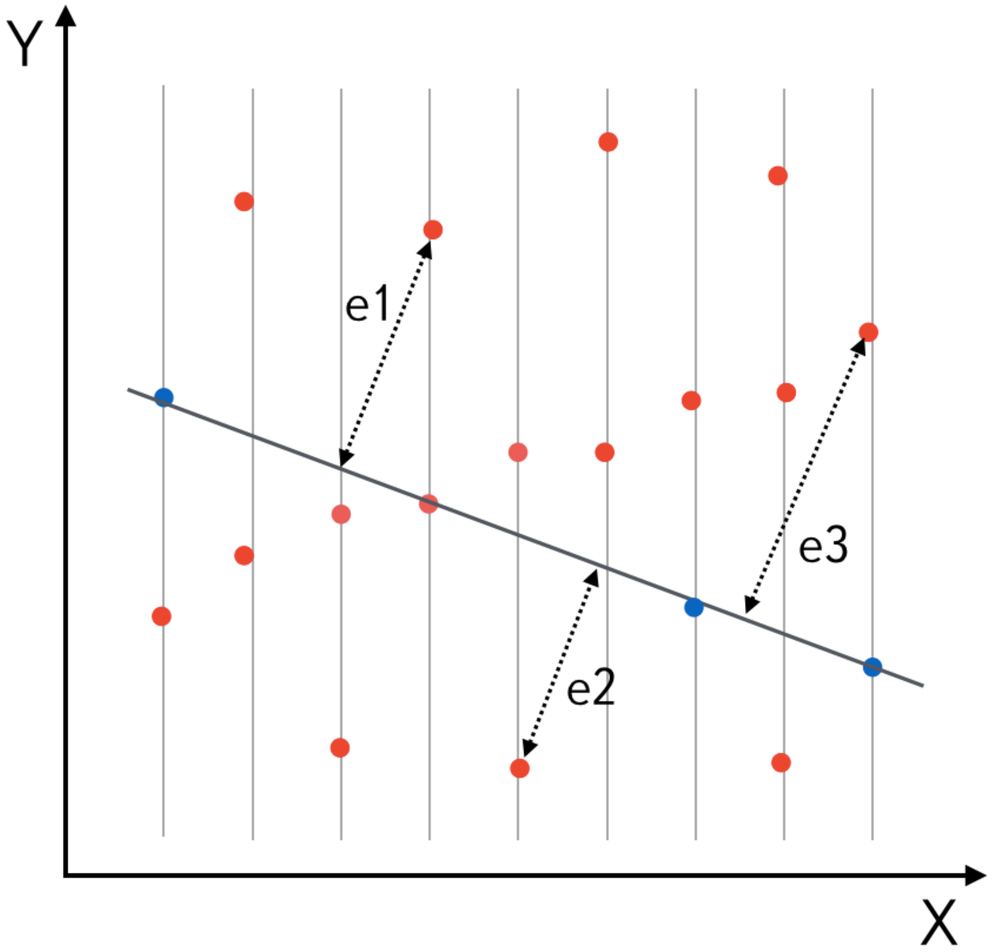
1. The RANSAC selects a random subset of the hits. For example, 3 hits.

2. Then, the linear model is fitted using these subset.



RANSAC

3. Calculate Error of the data with respect to the fitted model.



RANSAC

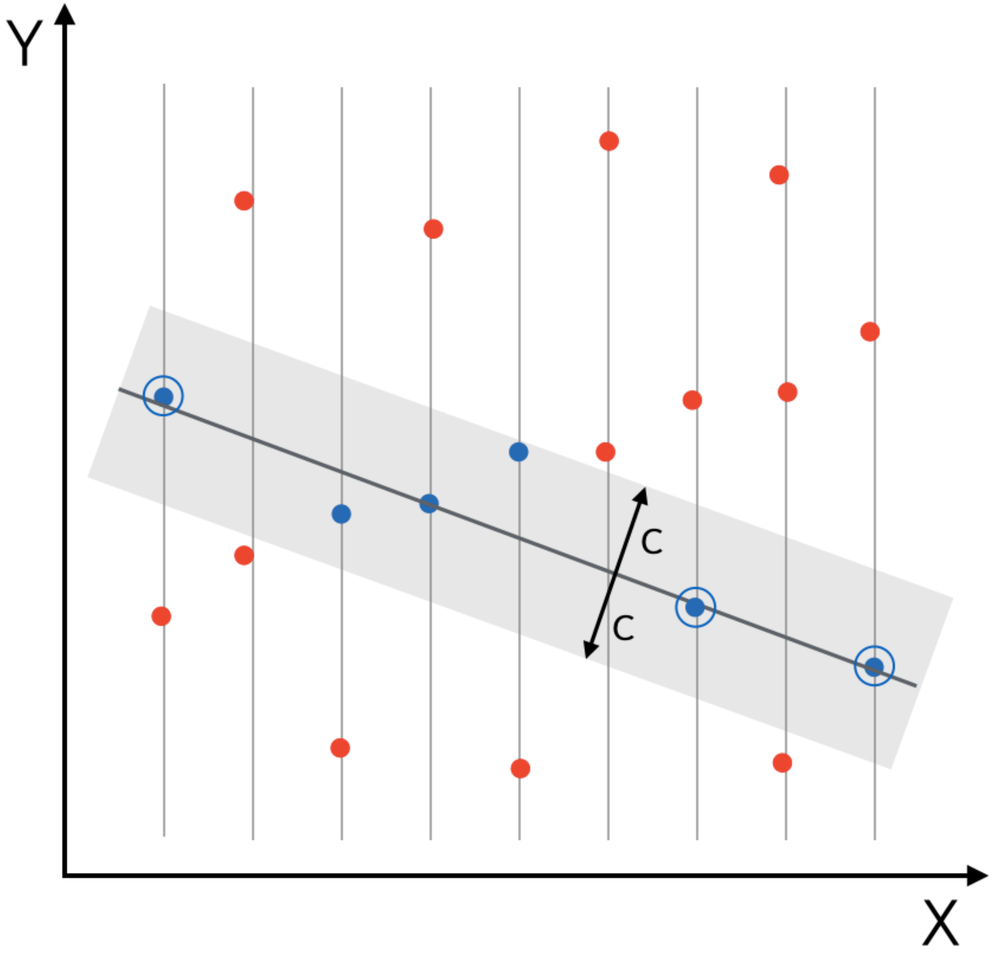
4. Number of inlier candidates is calculated. A hit is marked as inlier when:

$$|e| < c$$

where e is the error of the hit with respect to the fitted model, c is the error threshold.

5. The 1-4 steps are repeated until the maximum number of iteration is not exceeded.

6. A model with maximum number of inliers is returned.



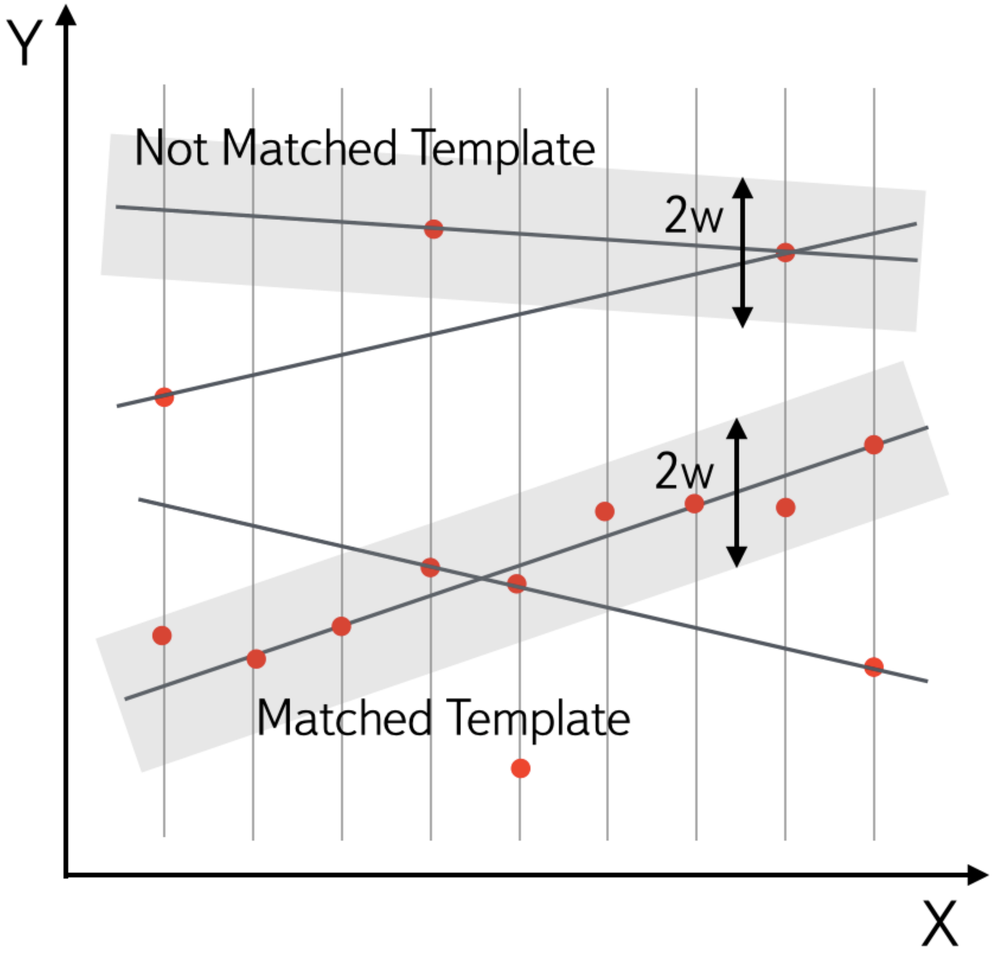
Tracks Reconstruction in Plane with RANSAC

Algorithm :

1. Run a RANSAC on all unmarked hits.

2. Mark inliers as a track candidate.

5. Repeat 1-2 steps until desired number of tracks are not found or until all tracks candidates have number of hits larger than a minima one.



Tracks Reconstruction in Plane

RANSAC parameters for Y_views:

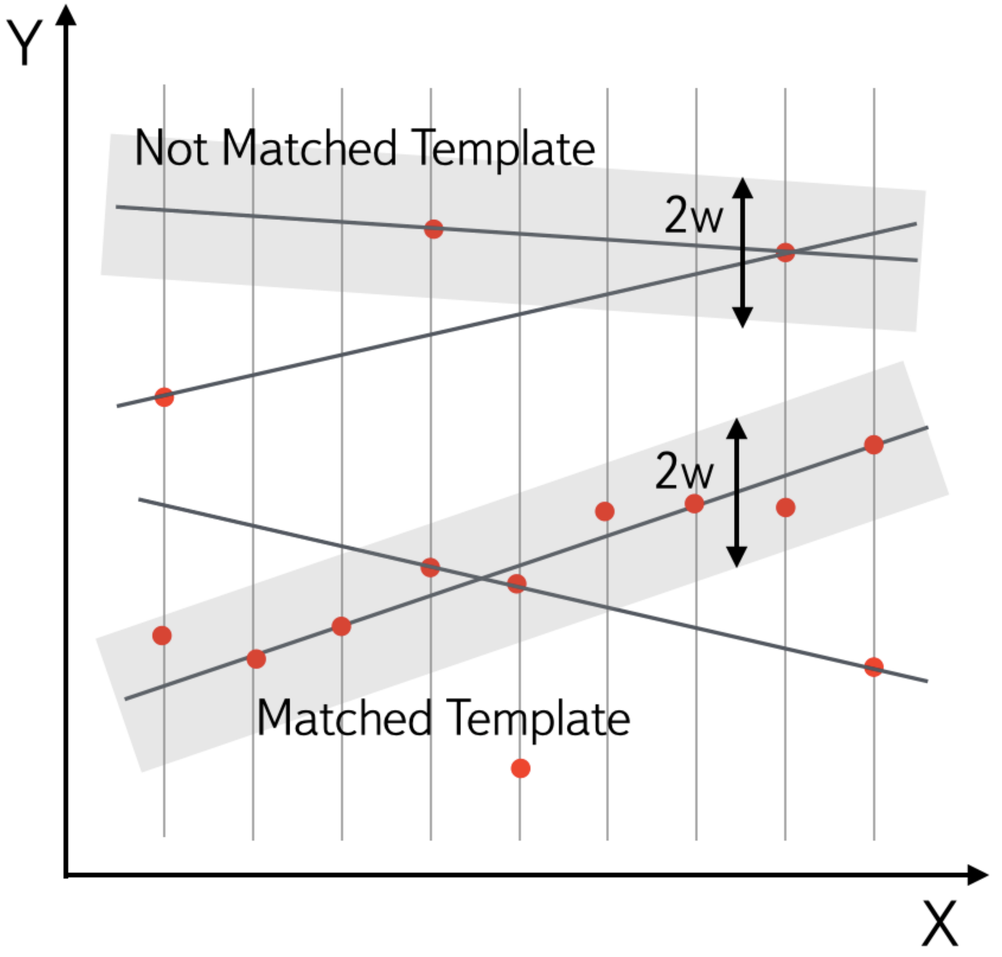
- n_tracks = 2;
- min_hits = 2;
- min_samples = 2;
- residual_threshold = 0.2;

RANSAC parameters for Stereo_views:

- n_tracks = 1;
- min_hits = 2;
- min_samples = 2;
- residual_threshold = 5;

The New Tracks Combination was used.

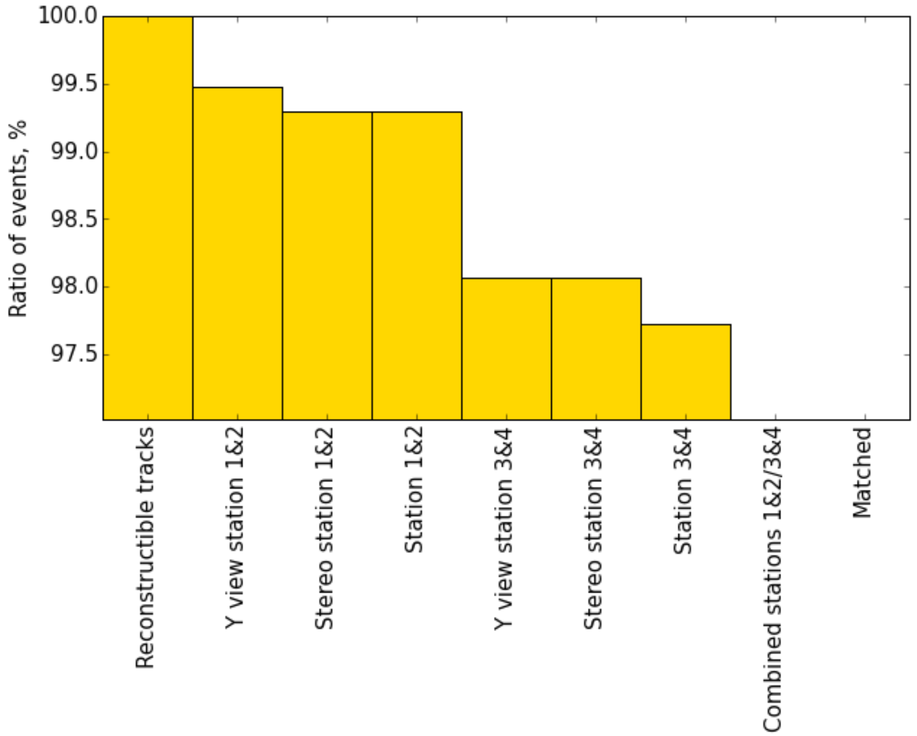
The Double Hits trick was used.

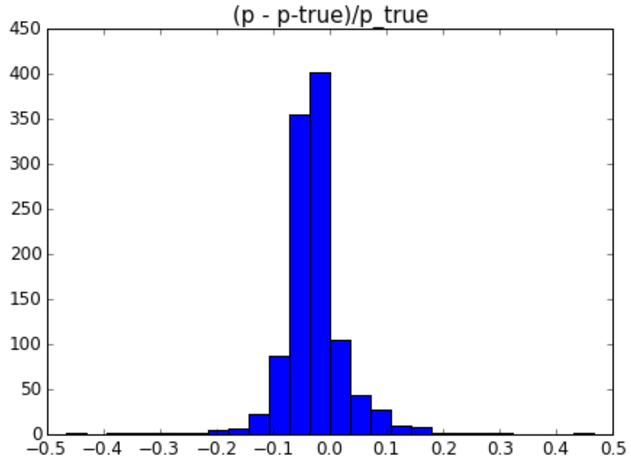


Quality Metrics

	Avg. Tracks Efficiency	Avg. Reconstruction Efficiency	Ghost Rate	Clone Rate
Y-views, stations 1&2	0,992	0,997	0,002	0,003
Stereo-views, stations 1&2	0,989	0,992	0,004	0,004
All-views, stations 1&2	0,986	0,997	0,003	0,000
Y-views, stations 3&4	0,997	0,993	0,007	0,011
Stereo-views, stations 3&4	0,999	0,986	0,008	0,004
All-views, stations 3&4	0,991	0,991	0,007	0,002
Combination	-	0,984	0,011	0,002

Quality Metrics





Efficiency is 97.0 %

Avg. accuracy is 6 %

Time

0.8 sec. / event.

SHiP Tracks Recognition

Hough Transform

Definition of the Hough Transform

The Hough transform is a method of finding instances within a certain class by voting procedure in a parameters space of the class.

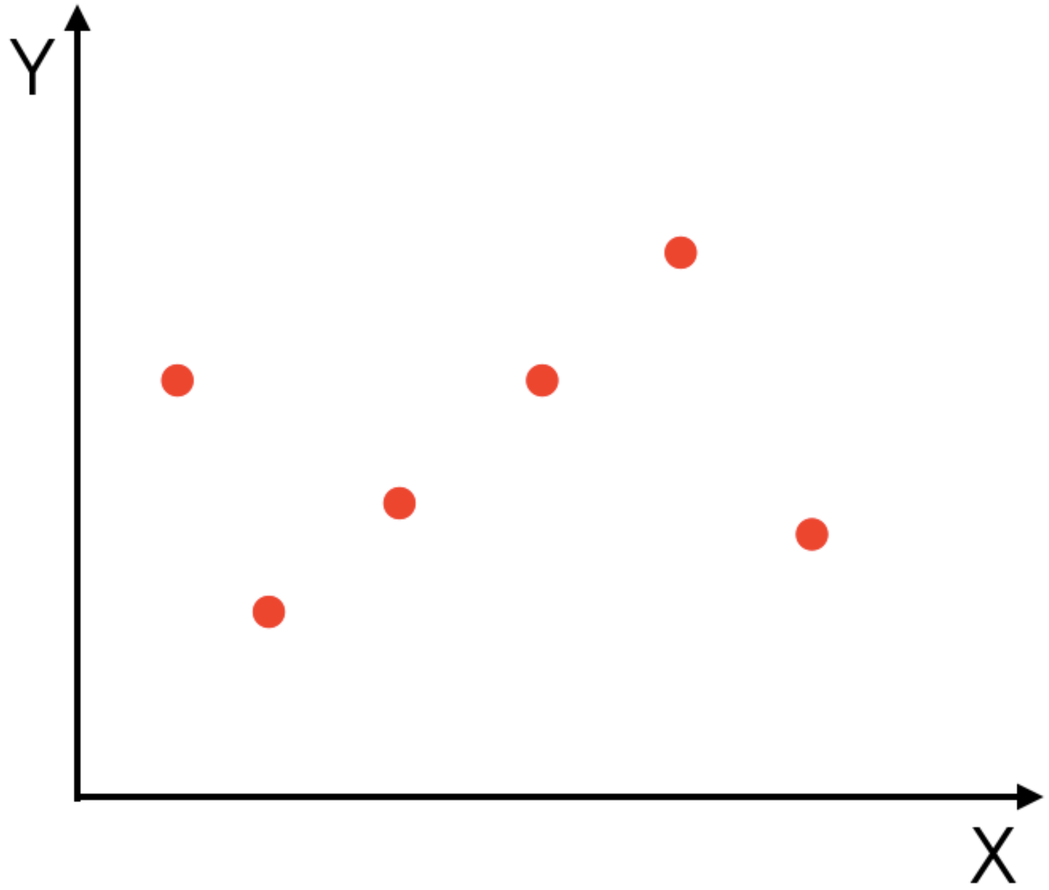
The Hough Transform

Consider a set of hits. Let search lines among the hits.

A line can be parametrized as

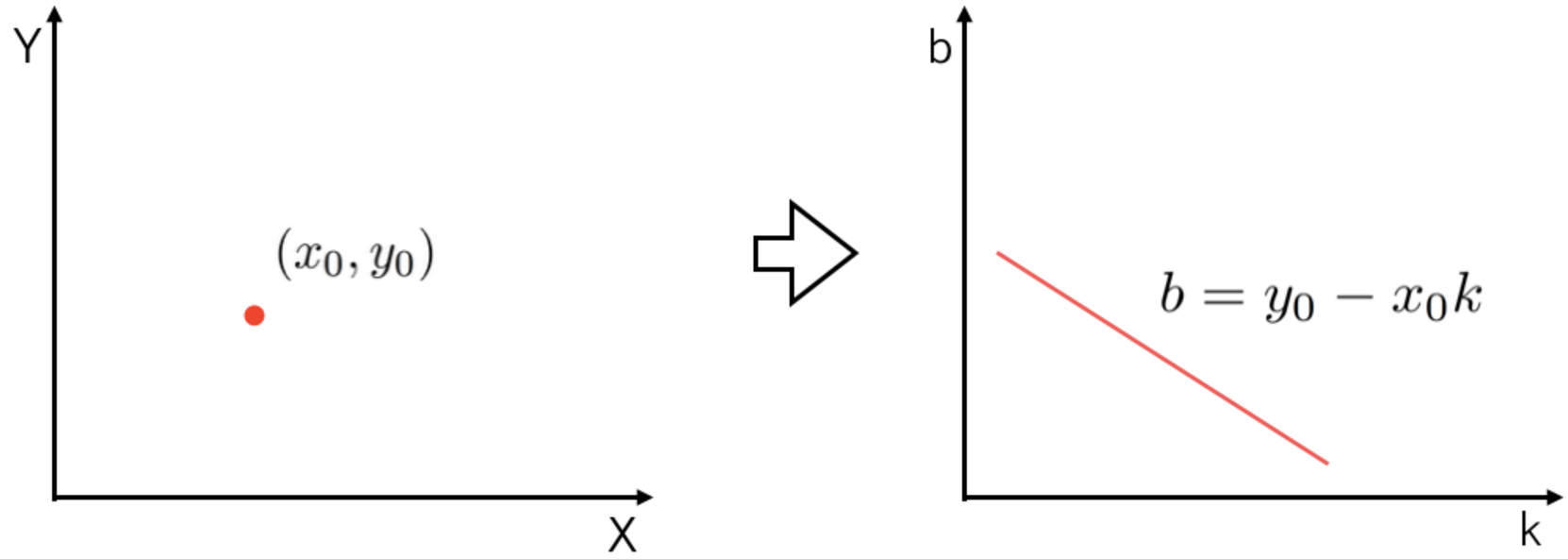
$$y = kx + b$$

where k, b are parameters of the line.



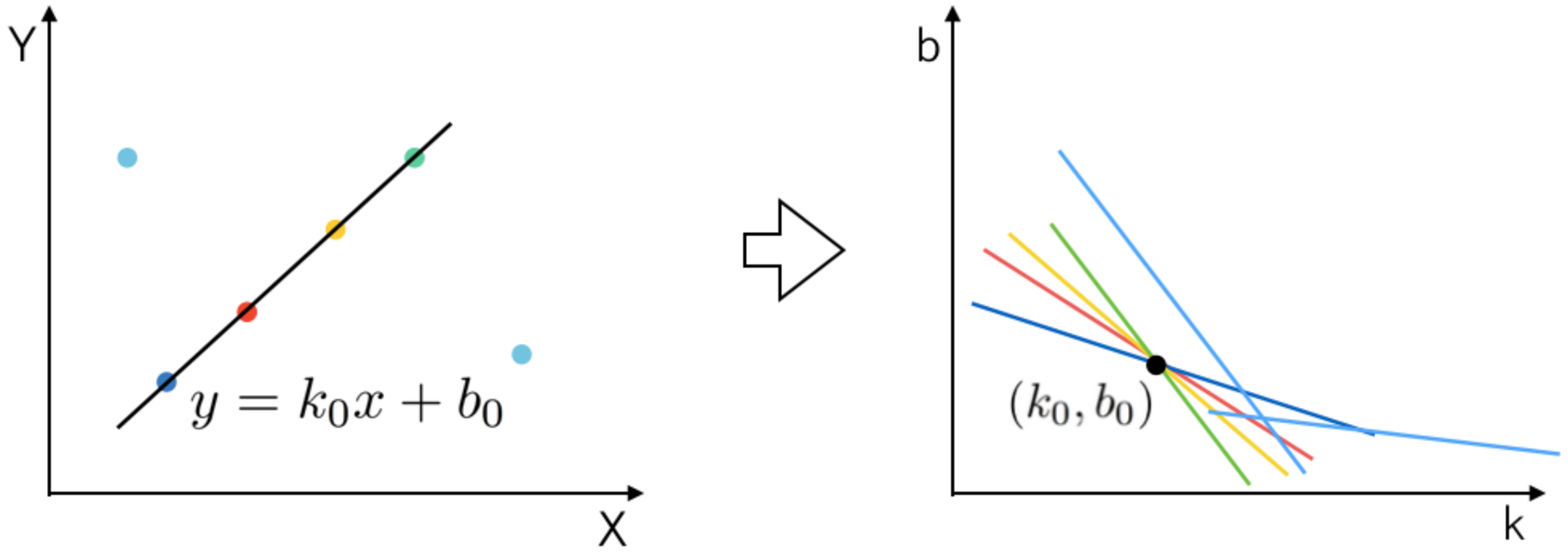
The Hough Transform

The Hough transform convert a point from (x,y) space to the curve in (k, b) space of the parameters. Each point of the curve in (k, b) space represents parameters of the line, that can go trough the point in (x, y) space.



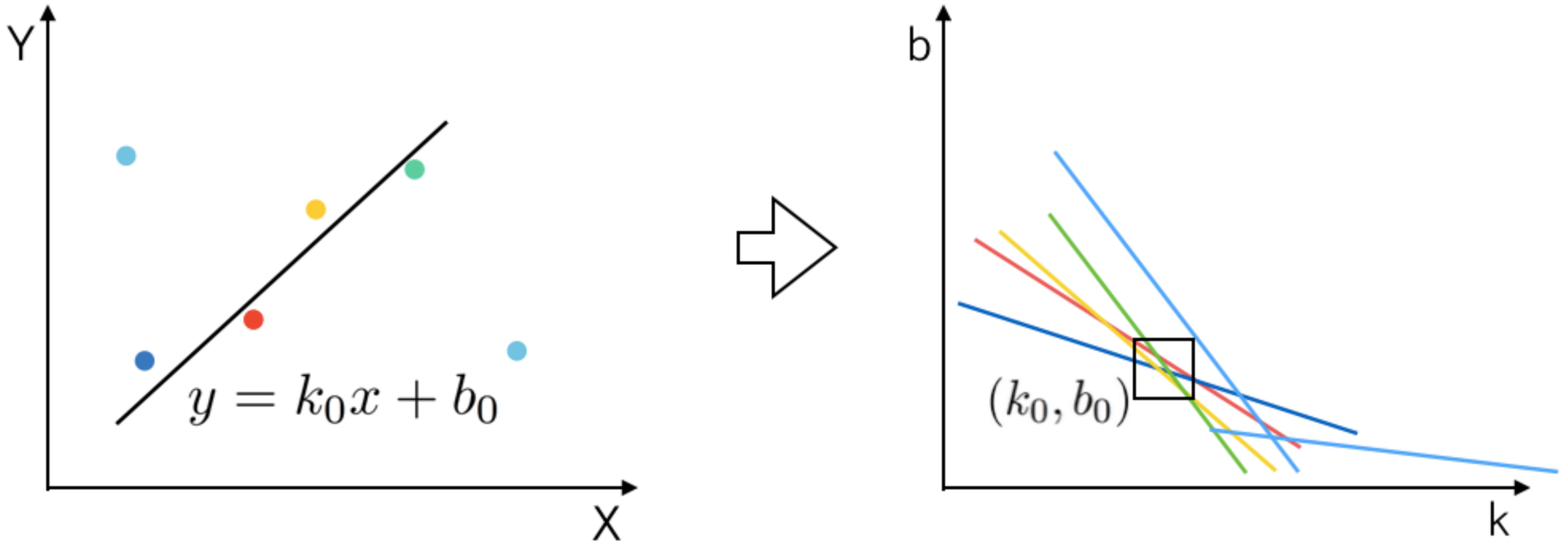
The Hough Transform

Since set of hits lie on the straight line, the curves in the parameter space should intersect in one point. This point corresponds the line's parameters.



The Hough Transform

Due to noise in coordinates of the hits, the optimal parameters of the track are in a cell in (k, b) space where the largest number of curves pass through.



The Hough Transform Implementation

Algorithm:

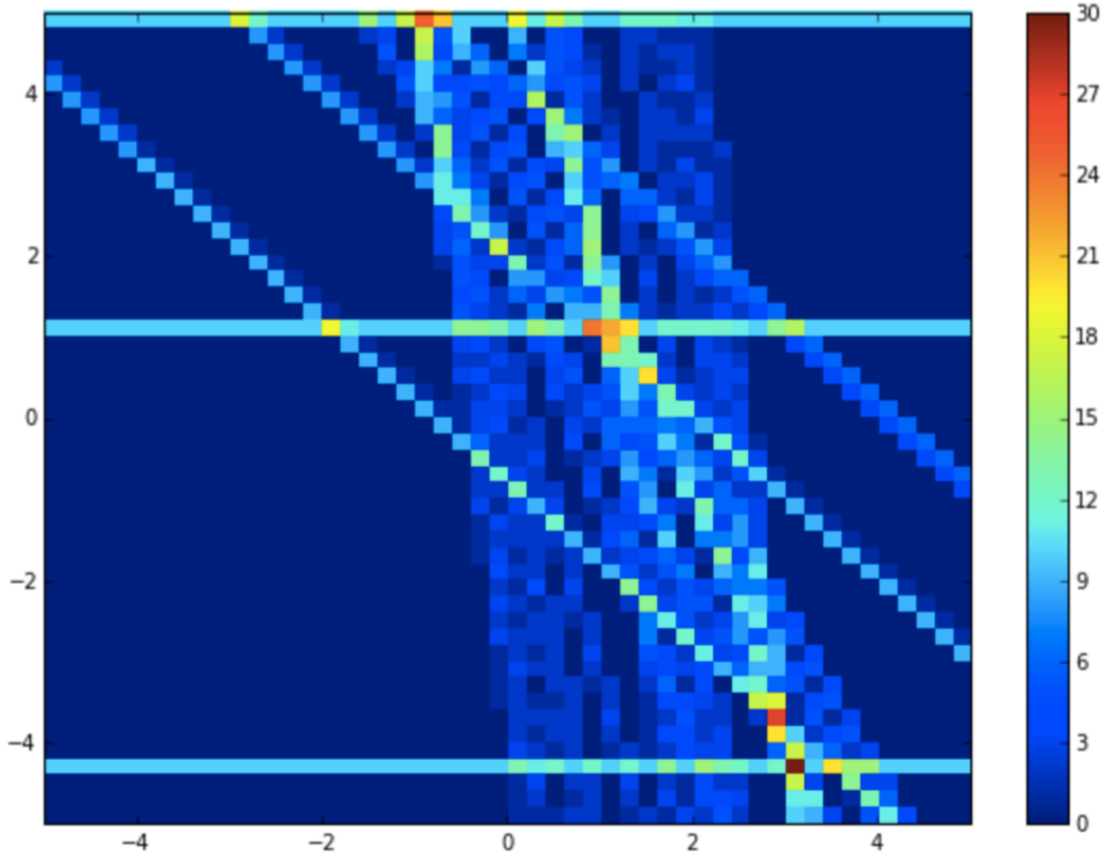
1. Transform each hit using the Hough transform.

2. Histogram the curves of the transformed hits.

3. Find bin with max intensity. This bin corresponds to a track parameters. Hits which are belong to this bin are marked as reconstructed track.

4. Exclude the curves for the marked hits from the histogram.

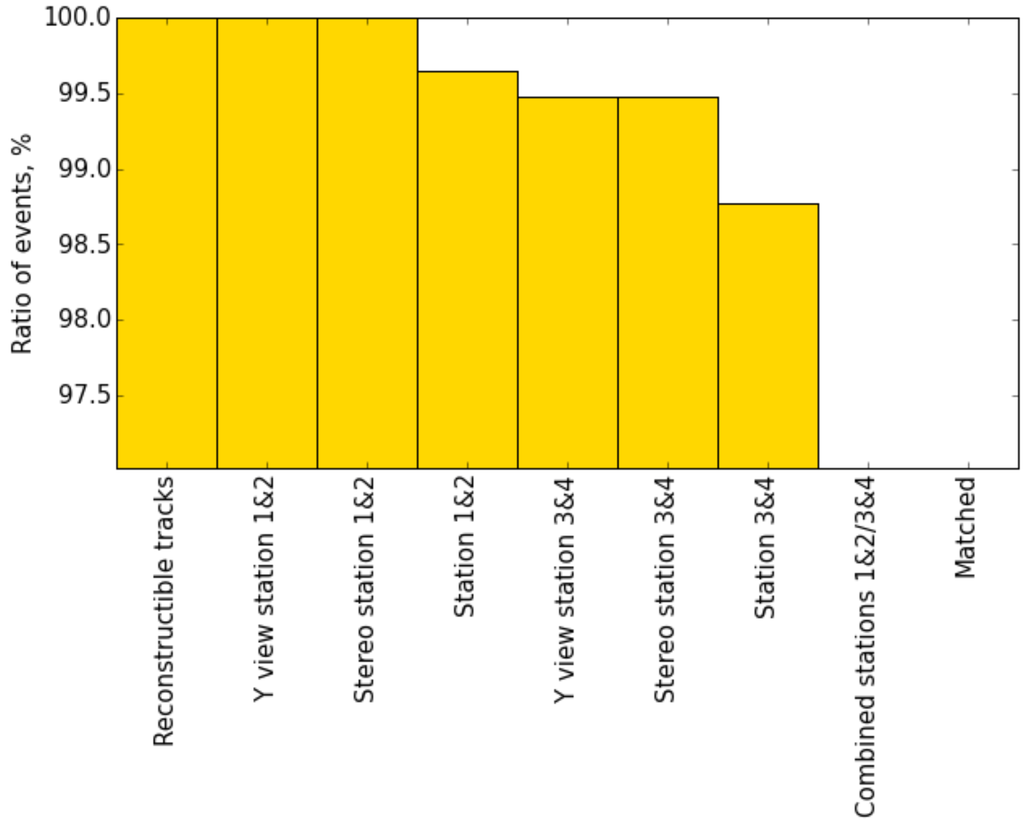
5. Repeat 3-4 steps until desirable number of tracks (2) is not found.



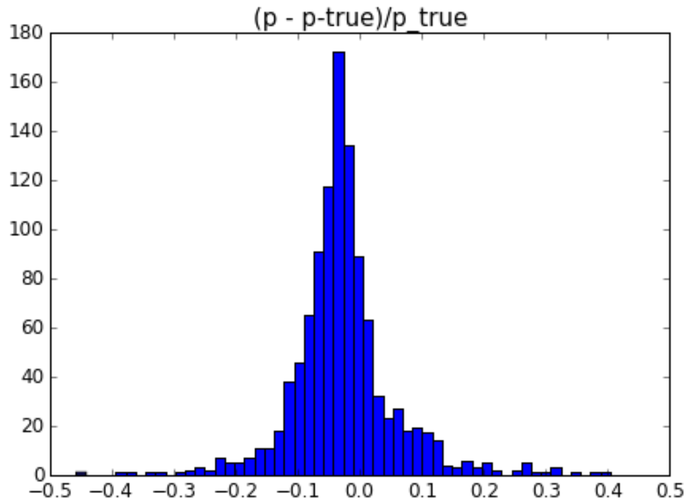
Quality Metrics

	Avg. Tracks Efficiency	Avg. Reconstruction Efficiency	Ghost Rate	Clone Rate
Y-views, stations 1&2	0,994	1	0,000	0,002
Stereo-views, stations 1&2	0,988	0,998	0,001	0,000
All-views, stations 1&2	0,989	0,998	0,001	0,001
Y-views, stations 3&4	0,995	0,999	0,001	0,011
Stereo-views, stations 3&4	0,990	0,994	0,001	0,002
All-views, stations 3&4	0,992	0,994	0,001	0,006
Combination	-	0,985	0,009	0,001

Quality Metrics



Efficiency is 97.0 %



Avg. accuracy is 7 %

Time

10 sec. / event.

SHiP Tracks Recognition

Retina. The First Approach.

The Retina Function

The retina function is defined as:

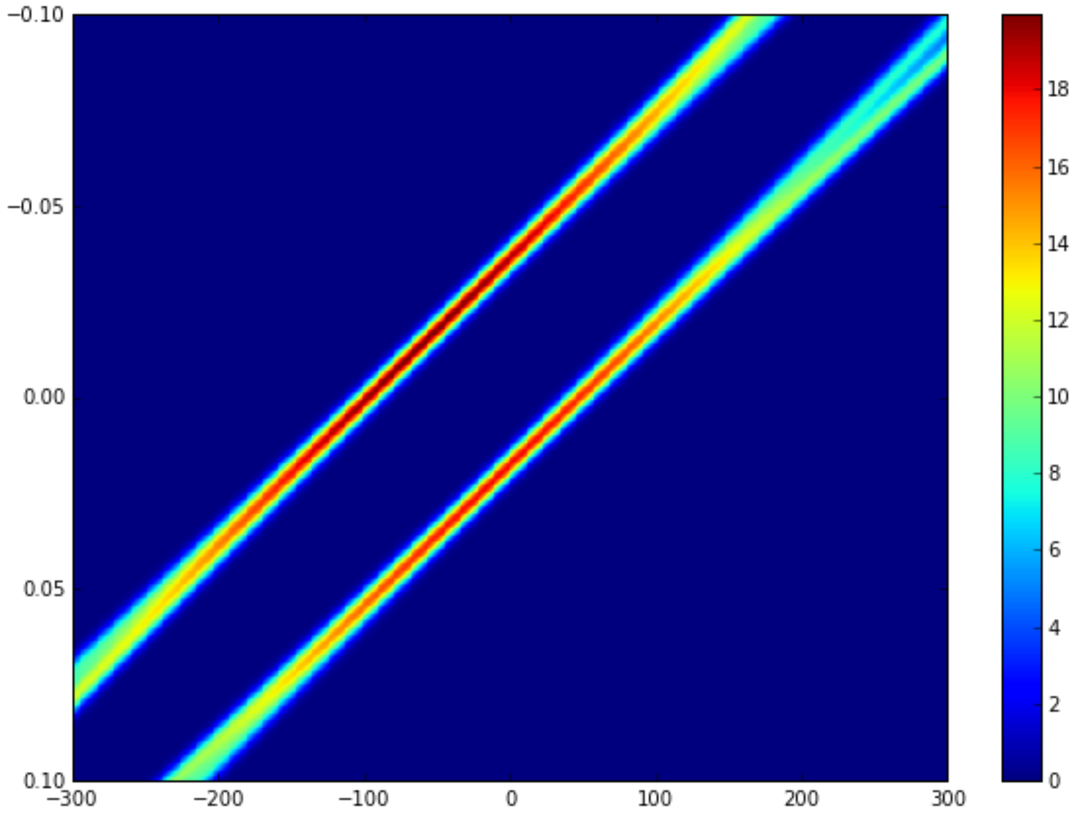
$$R(\theta) = \sum_i e^{-\frac{\rho^2(\theta, x_i)}{\sigma^2}}$$

where $\rho(\theta, x_i)$ is distance between the i-th hit and a track with parameters θ .

For 2D tracks:

$$\rho(\theta, x_i) = y_i - (kx_i + b)$$

$$\theta = [k, b]$$

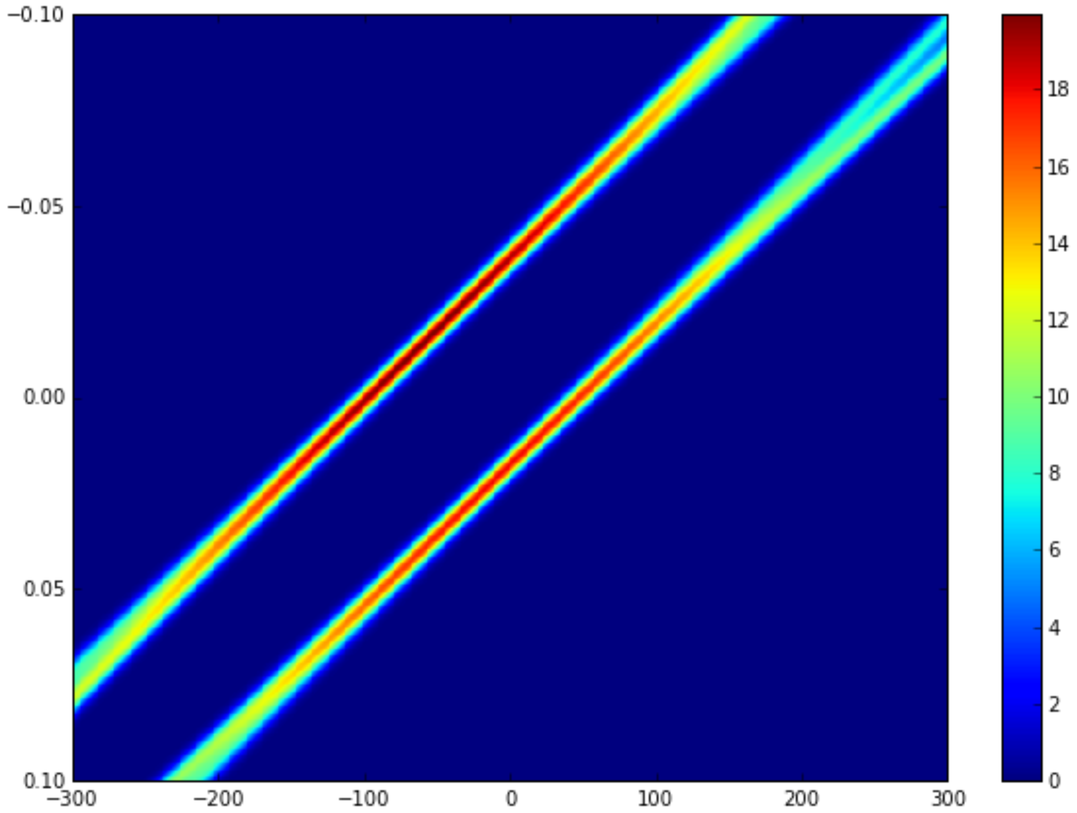


The Retina Function

The retina function maxima corresponds to a track parameters.

Then, the tracks recognition can be considered as an optimization problem.

The tracks reconstruction performance depends on the way of searching the retina function maxima.



The Retina Function

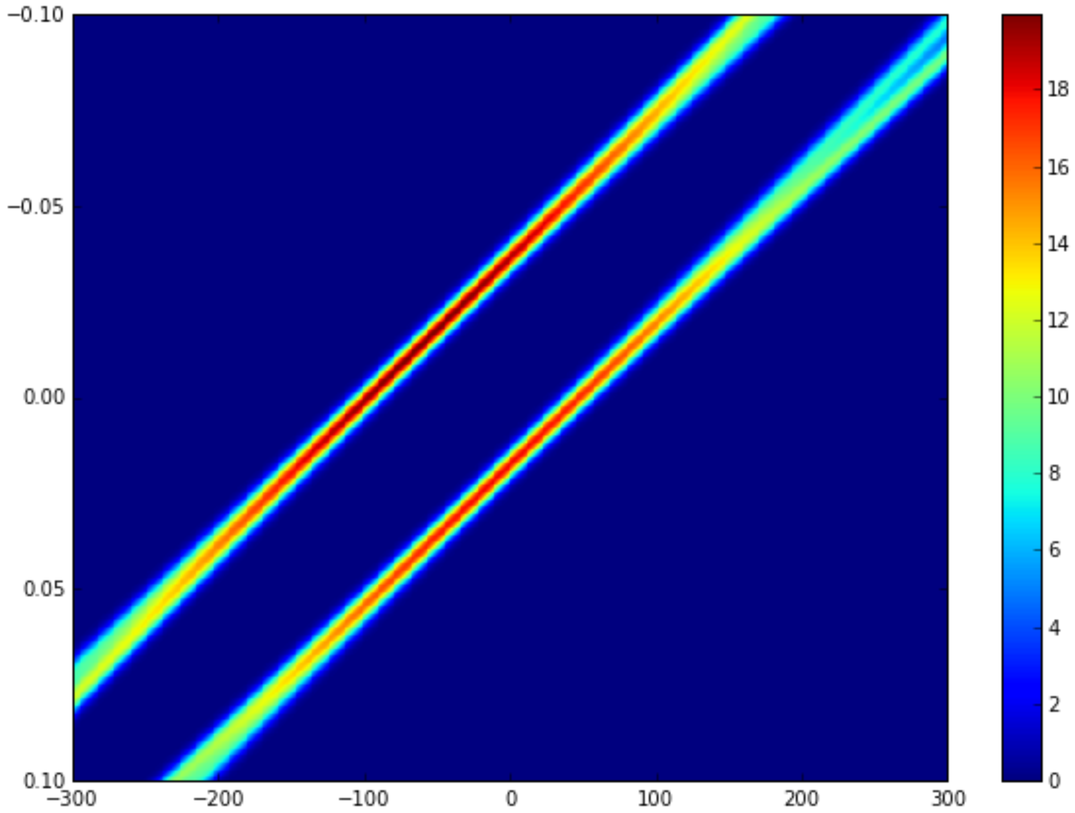
Algorithm:

1. Set init track parameters.

2. Find the retina function maxima using an optimization technique (GD, AdaDelta, BFGS, etc.). The sigma parameter can be decreased during the optimization.

3. Select the found track's hits based on their distances to the track.

4. Repeat 1-3 steps until the desired number of tracks is not found, or until the tracks has min number of hits.



The Retina Function

The algorithm parameters for Y-views:

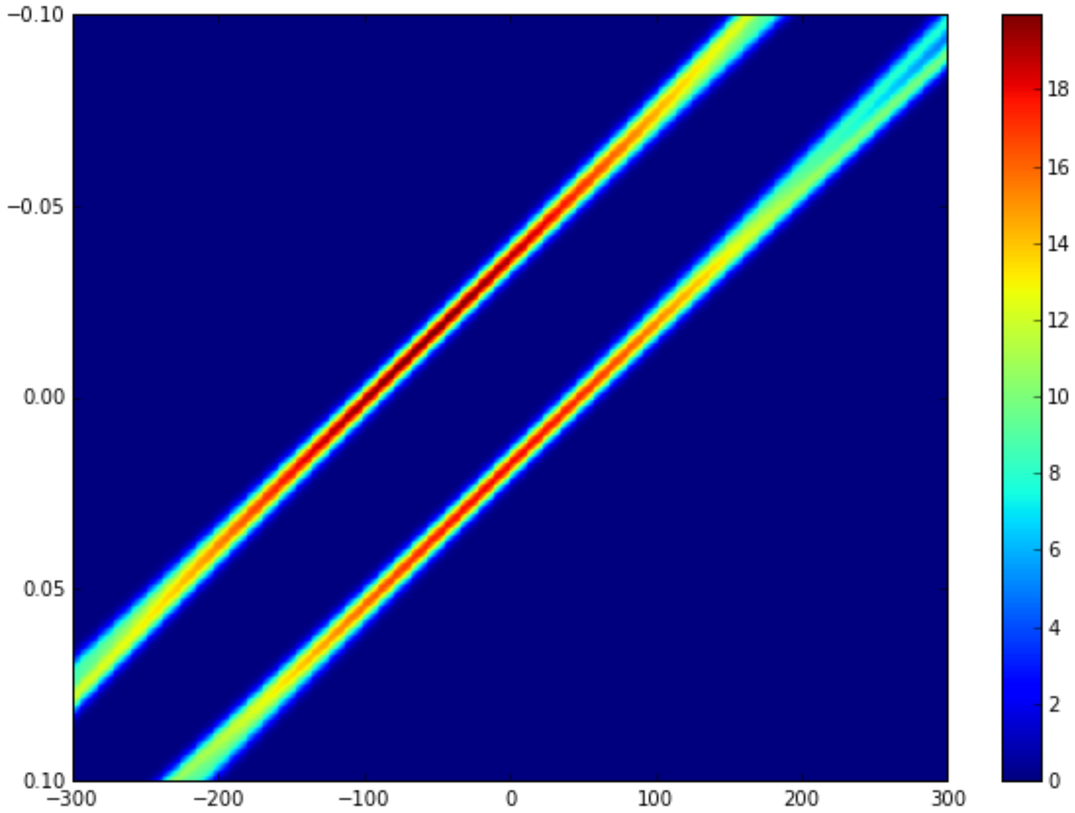
- n_tracks = 2;
- residuals_threshold = 0.9;
- sigma_max = 1000;
- sigma_min = 1;
- sigma_decay_rate = 0.5;
- min_hits = 2;

The algorithm parameters for Stereo-views:

- n_tracks = 1;
- residuals_threshold = 15;
- sigma_max = 1000;
- sigma_min = 10;
- sigma_decay_rate = 0.5;
- min_hits = 2;

The New Tracks Combination was used.

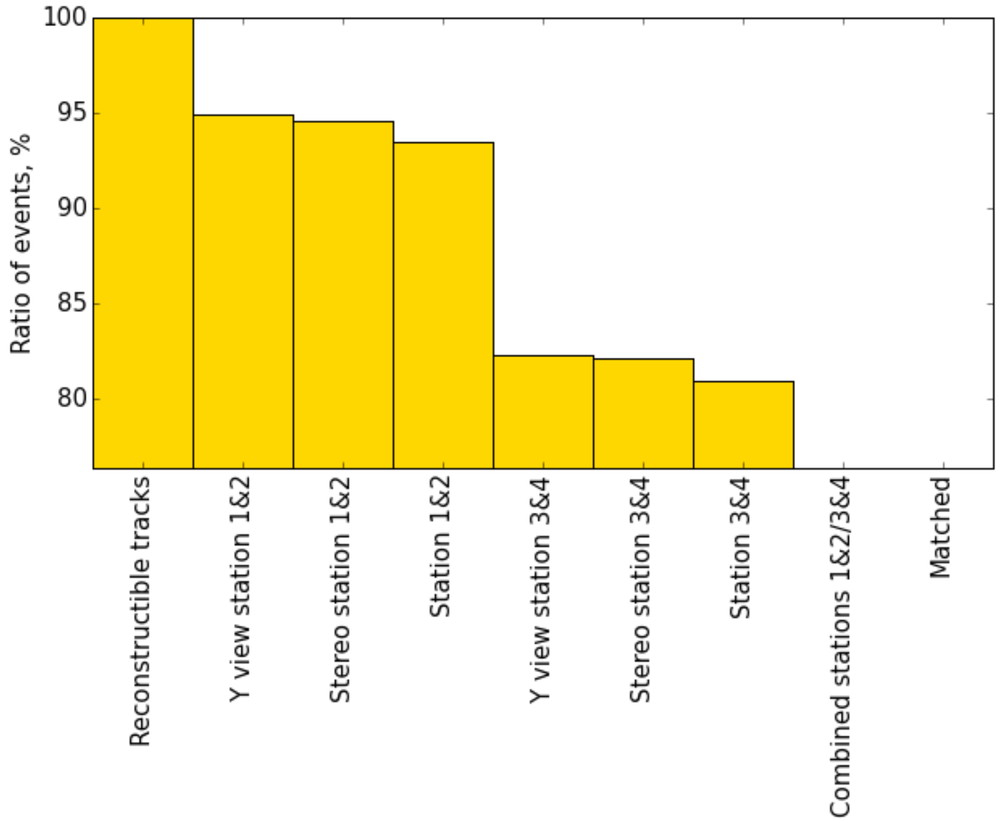
The Double Hits trick was used.

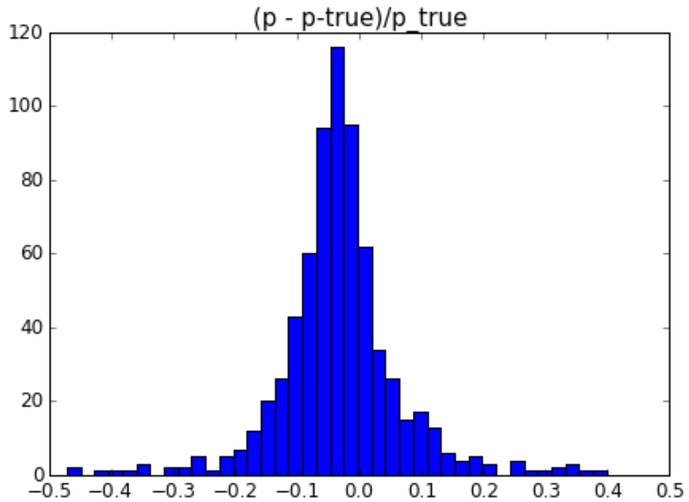


Quality Metrics

	Avg. Tracks Efficiency	Avg. Reconstruction Efficiency	Ghost Rate	Clone Rate
Y-views, stations 1&2	0,962	0,97	0,015	0,014
Stereo-views, stations 1&2	0,952	0,955	0,018	0,025
All-views, stations 1&2	0,946	0,967	0,017	0,015
Y-views, stations 3&4	0,948	0,931	0,052	0,021
Stereo-views, stations 3&4	0,941	0,922	0,043	0,032
All-views, stations 3&4	0,926	0,939	0,039	0,018
Combination	-	0,824	0,046	0,004

Quality Metrics





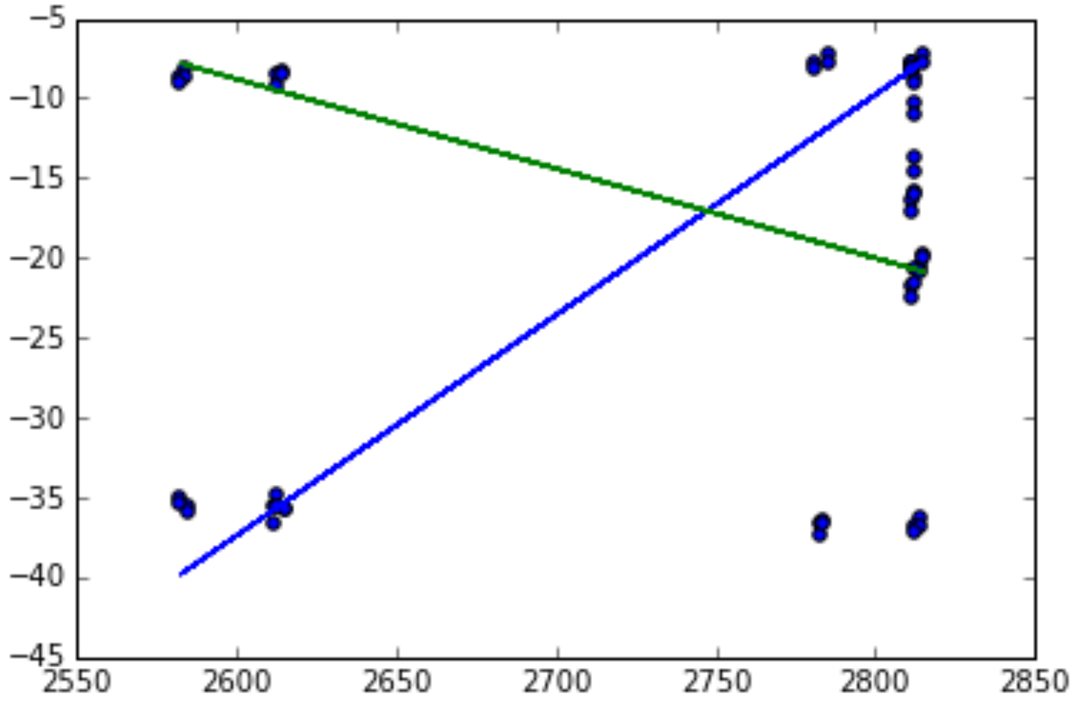
Efficiency is 76.3 %

Avg. accuracy is 16 %

Time

0.3 sec. / event.

Why so bad?



SHiP Tracks Recognition

Retina. The Second Approach.

The Retina Function

The retina function is defined as:

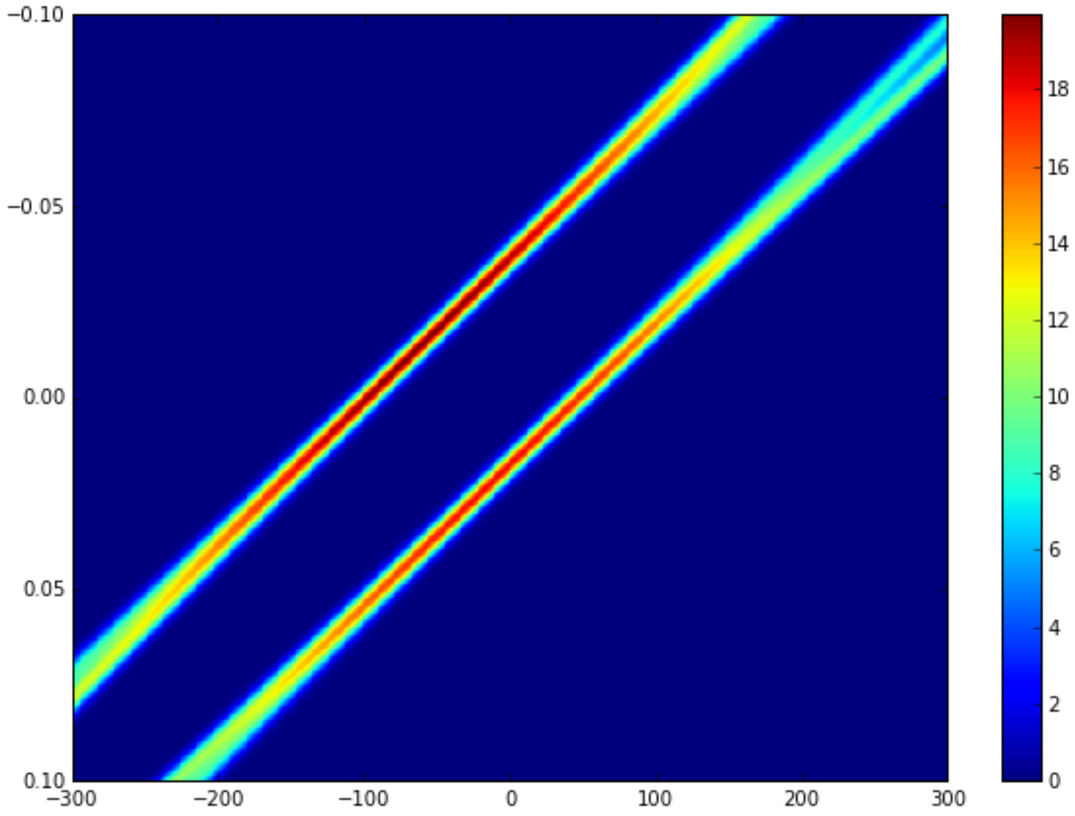
$$R(\theta) = \sum_i e^{-\frac{\rho^2(\theta, x_i)}{\sigma^2}}$$

where $\rho(\theta, x_i)$ is distance between the i-th hit and a track with parameters θ .

For 2D tracks:

$$\rho(\theta, x_i) = y_i - (kx_i + b)$$

$$\theta = [k, b]$$



The Retina Function

Algorithm:

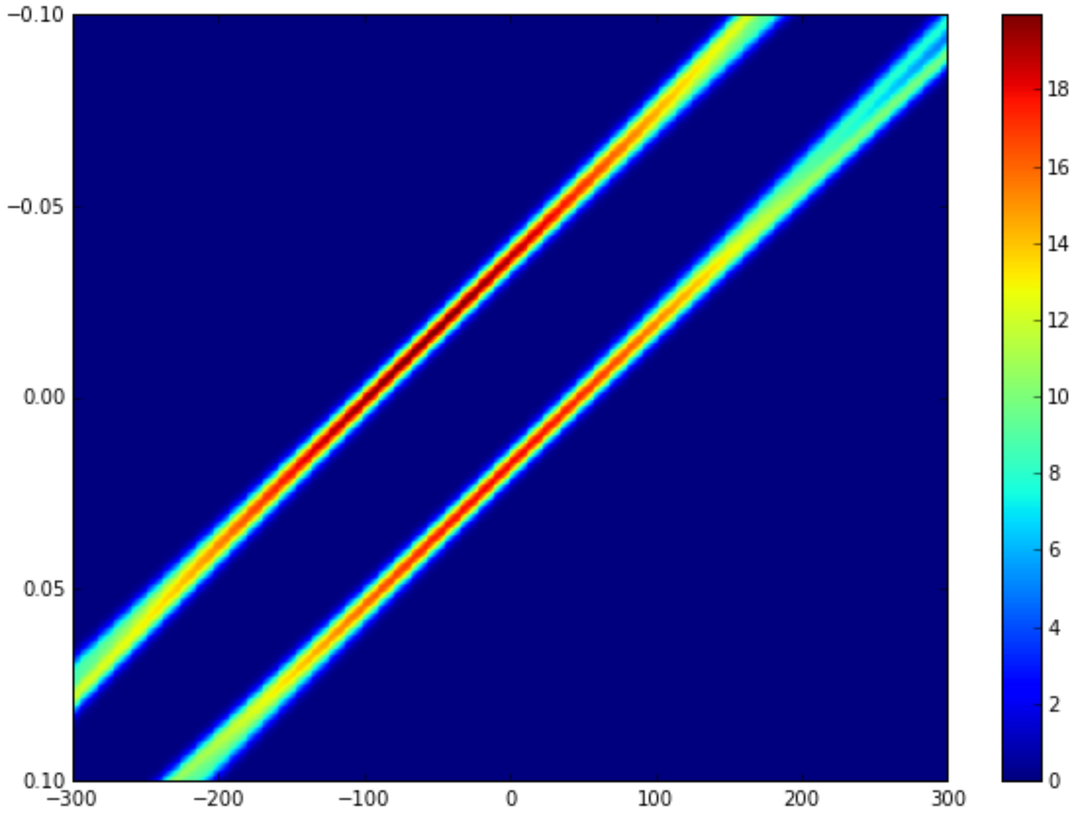
1. For each pairs of unmarked hits calculate a line parameters and the retina function value for its parameters.

2. Take the line parameters correspond to the max retina function values as the init point for the optimization problem.

3. Find the retina function maxima using an optimization technique (GD, AdaDelta, BFGS, etc.). The sigma parameter can be decreased during the optimization.

4. Select the found track's hits based on their distances to the track. Mark these hits.

5. Repeat 1-4 steps until the desired number of tracks is not found, or until the tracks has min number of hits.

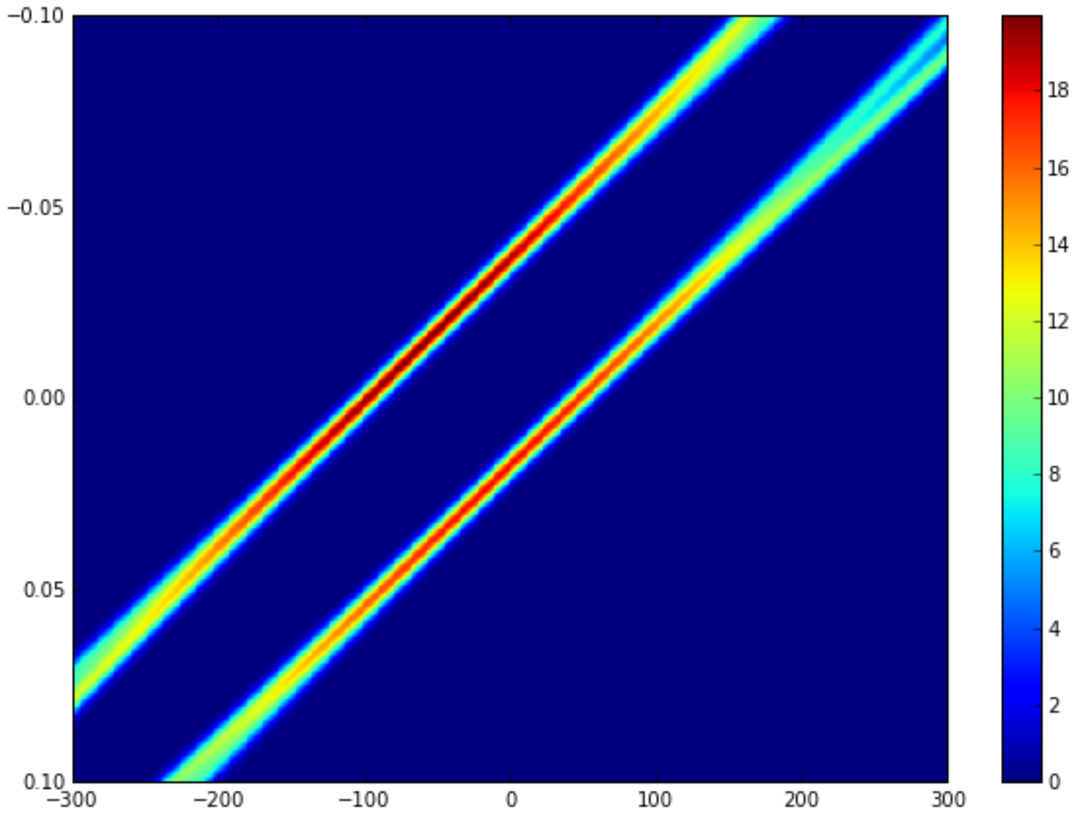


The Retina Function

The algorithm parameters for Y-views:
- n_tracks = 2;
- residuals_threshold = 0.5;
- sigma = 0.5;
- min_hits = 2;

The algorithm parameters for Stereo-views:
- n_tracks = 1;
- residuals_threshold = 5;
- sigma = 5;
- min_hits = 2;

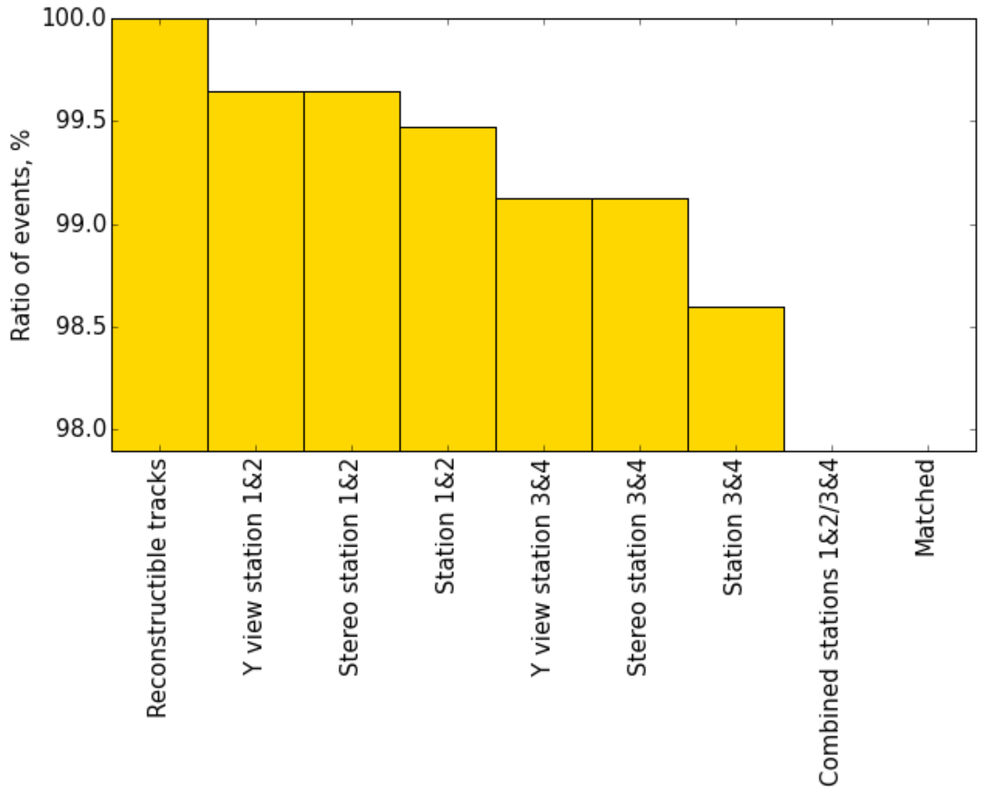
The New Tracks Combination was used.
The Double Hits trick was used.

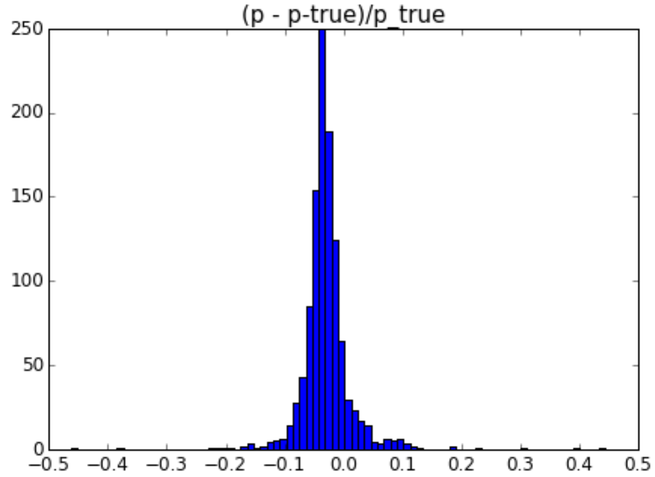


Quality Metrics

	Avg. Tracks Efficiency	Avg. Reconstruction Efficiency	Ghost Rate	Clone Rate
Y-views, stations 1&2	0,992	0,998	0,000	0,004
Stereo-views, stations 1&2	0,991	0,998	0,001	0,001
All-views, stations 1&2	0,989	0,998	0,000	0,002
Y-views, stations 3&4	0,993	0,998	0,001	0,010
Stereo-views, stations 3&4	0,994	0,996	0,001	0,004
All-views, stations 3&4	0,990	0,996	0,001	0,004
Combination	-	0,988	0,010	0,003

Quality Metrics





Efficiency is 97.9 %

Avg. accuracy is 5 %

Time

0.2 sec. / event.

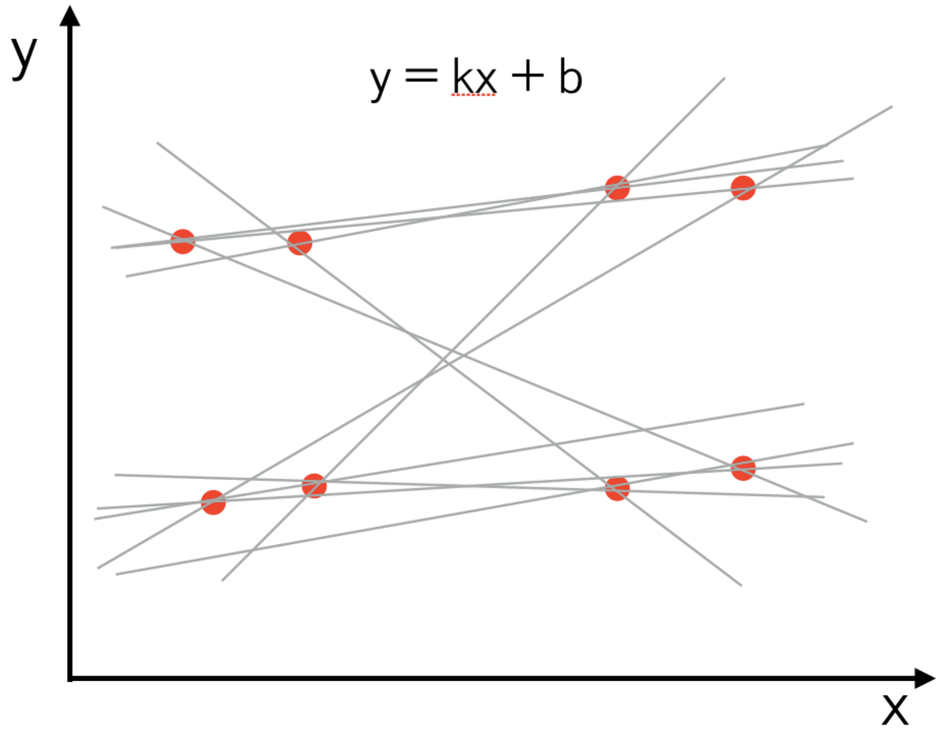
SHiP Tracks Recognition

New Model

New Model

1. Select a random subset of the hits. For example, 2 hits.

2. Fit linear models for each subset. (Classification or threshold based on chi-square value and parameters of the model helps to reject bad ones.)



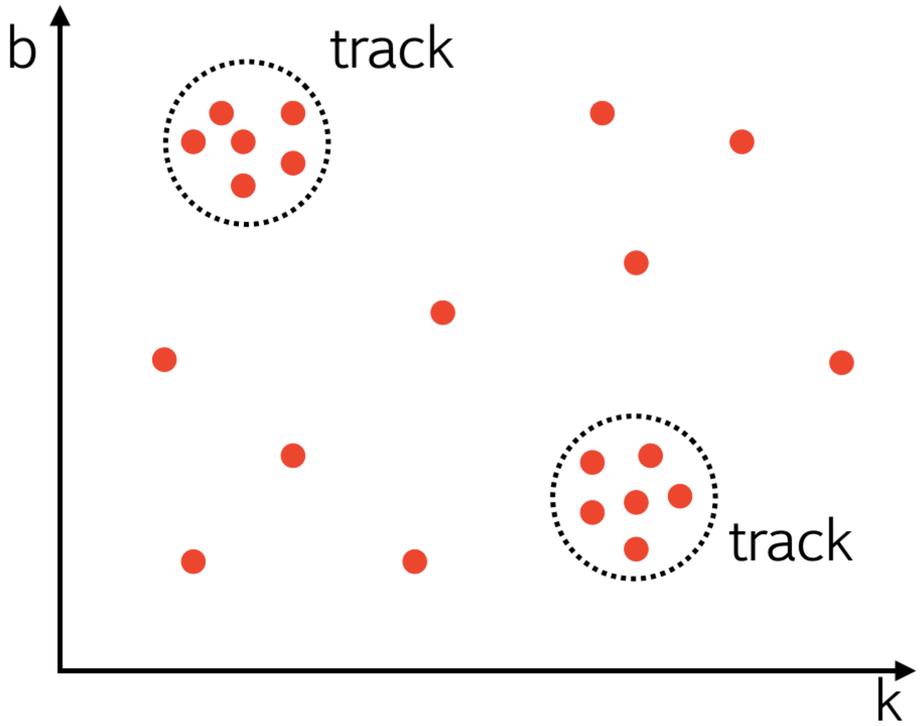
New Model

3. Plot all the linear models in parameter space (one point for a linear model).

4. Find the most dense region using kNN method. This region corresponds to a track. Find hits which correspond to this track (clustering, classification, threshold value). Mark them as reconstructed track.

5. Remove linear models which was fitted using the marked hits.

6. Repeat 4-5 steps until N tracks are not found or until dense regions exist.



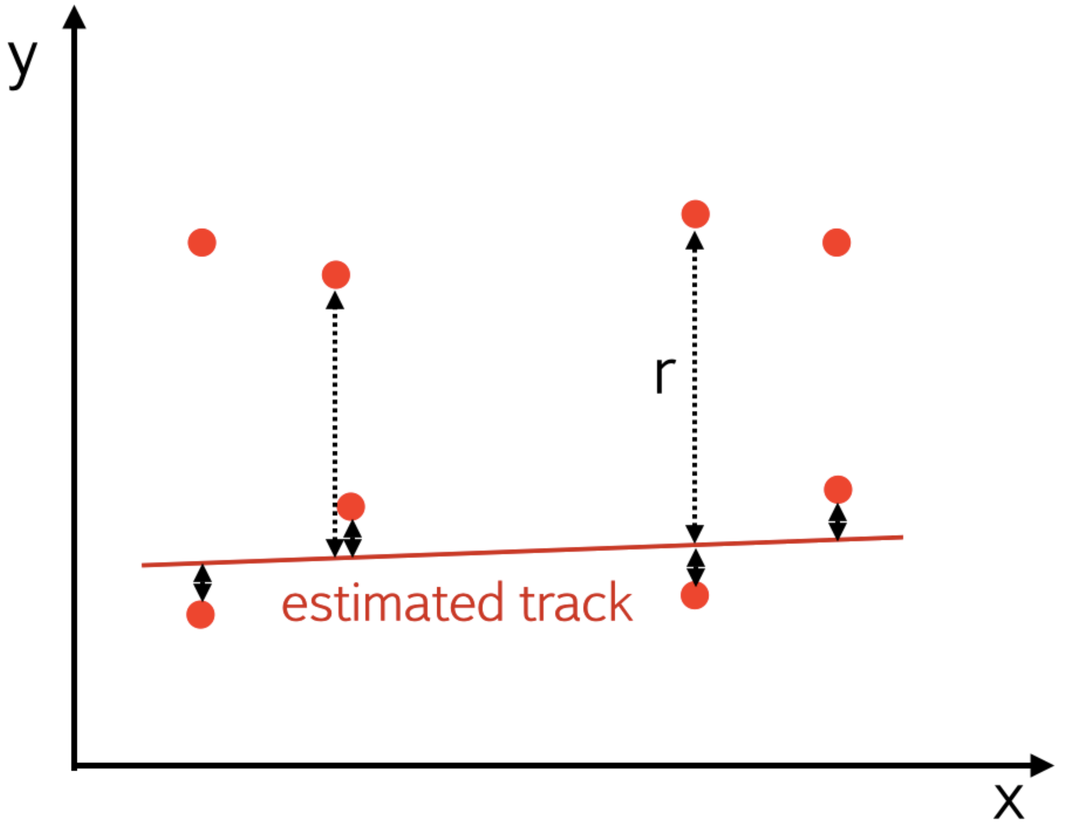
New Model

Hits corresponding to the estimated track can be found based on distances to the track by the following ways:

〉 Distance threshold value

〉 Clustering on several clusters based of Silhouette score (default)

〉 Classification using the distances (chi-square impact) and the track parameters



New Model

1. The following parameters were used:

〉 2 hits in random subset

〉 k = 20 in kNN method

〉 All combinations

〉 No classifiers

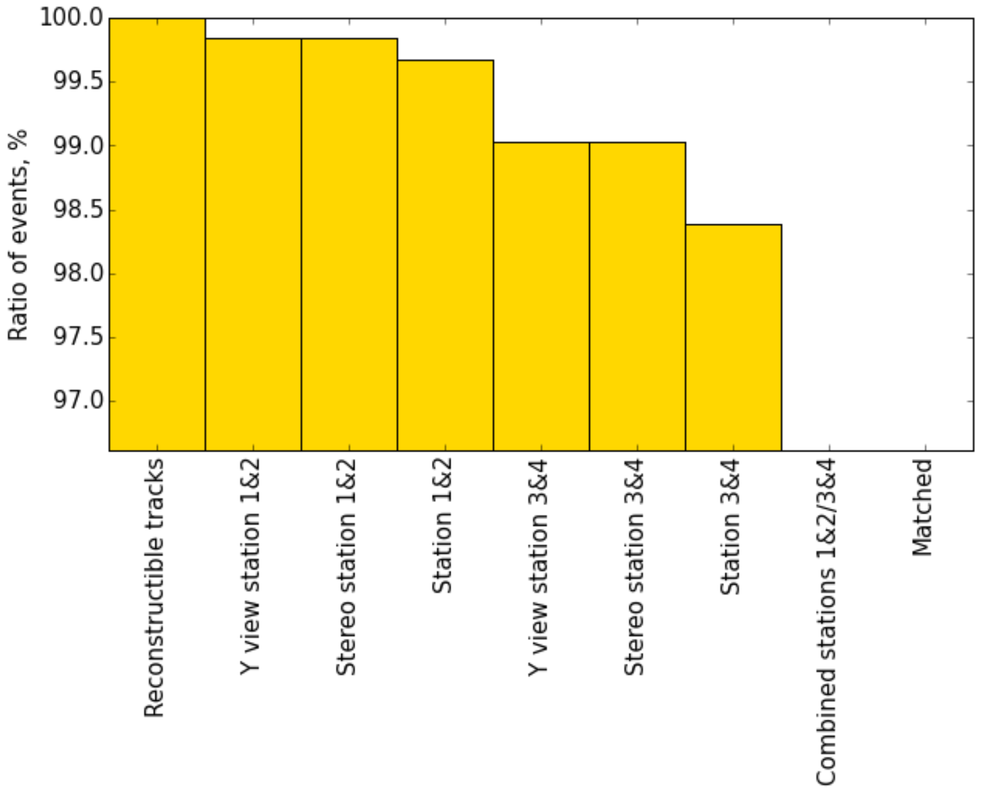
2. Each hit was doubled by using distance from a track to a straw tube's wire.

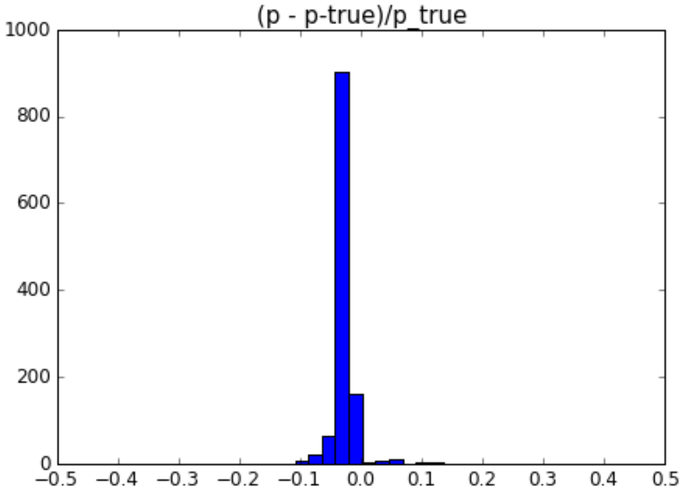
3. With new tracks combination.

Quality Metrics

	Avg. Tracks Efficiency	Avg. Reconstruction Efficiency	Ghost Rate	Clone Rate
Y-views, stations 1&2	0,996	0,999	0,000	0,002
Stereo-views, stations 1&2	0,990	0,993	0,002	0,006
All-views, stations 1&2	0,989	0,998	0,000	0,002
Y-views, stations 3&4	0,992	0,997	0,003	0,010
Stereo-views, stations 3&4	0,986	0,994	0,002	0,004
All-views, stations 3&4	0,986	0,994	0,002	0,004
Combination	-	0,983	0,014	0,005

Quality Metrics





Efficiency is 96.6 %

Avg. accuracy is 3 %

Time

12. sec. / event.

New Model with Classifiers

Classifiers can be used to:

- Reject bad linear models on 2nd step
- Reject bad hits for the estimating track on 4th step

The results of the models with the classifiers are close to that one's without the classifiers. But it needs more events for the more precise comparison.

SHiP Tracks Recognition

Summary Results

Summary Results

	Avg. Efficiency, %	Avg. Momentum Accuracy, %	Time, sec/event
Baseline	88.7	4	0.4
Baseline + New Tracks Comb.	91.8	4	0.4
Baseline + Double Hits Trick	97.2	6	0.9
RANSAC	97.0	6	0.8
Hough	97.0	7	10
Retina One	76.3	16	0.3
Retina Two	97.9	5	0.2
New Model	96.6	3	12

SHiP Tracks Recognition

Next Steps

Limitations

The current method which uses the two 2D projections has the following limitations:

- Estimated hits' (z, x) coordinates has high errors due to the errors of the track in y-z plane and small slope of the U,V-views with respect to the Y-view.

- Some of the tracks have few hits in one of the 2D projections. This makes it hard to recognize them properly.

The First New Idea

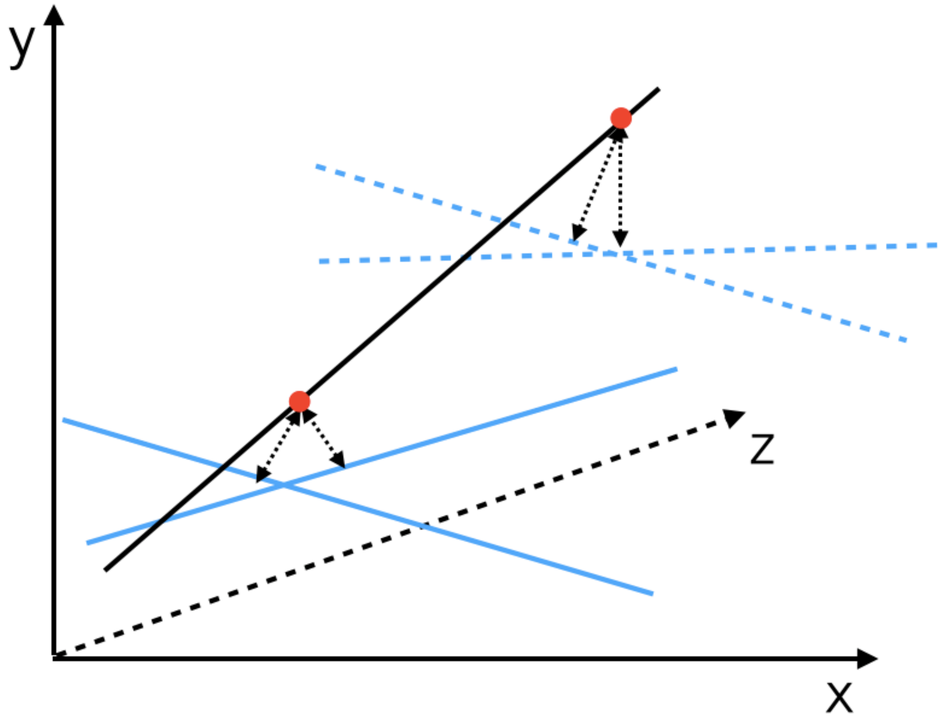
Find tracks using distances from a track model in 3D to the active straw tubes. Without any projections.

Try the following models in 3D:

〉 Retina-like model

〉 Elastic arms

〉 Robust two-models linear regression. This model fits two linear models simultaneously.

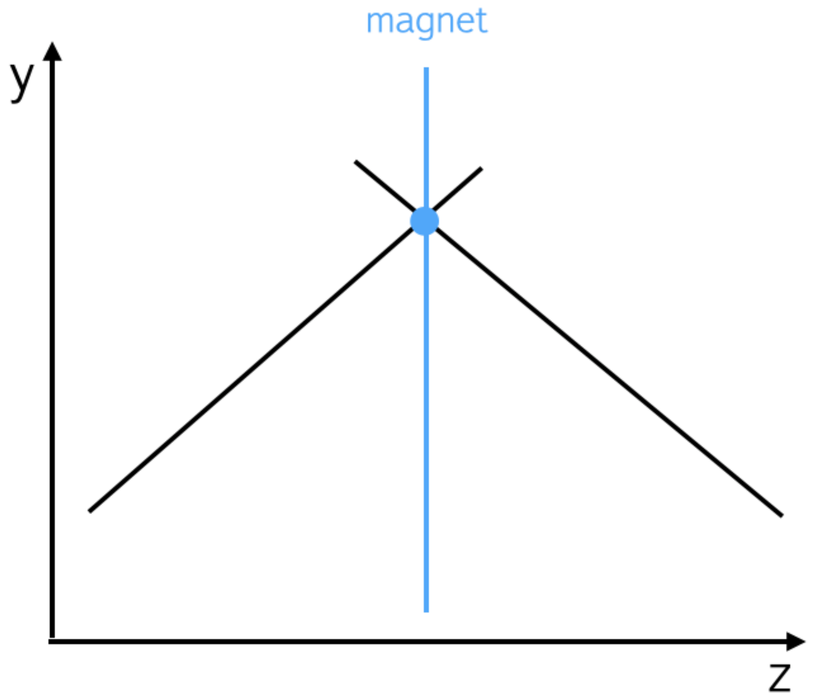


The Second New Idea

Fit 3D tracks before and after the magnet simultaneously using the following assumptions:

〉 Extrapolations of the tracks should pass through the same point in the center of the magnet

〉 The track is not deflected by the magnetic field in x-z plane.



Thanks for your attention!