Developments in other math and statistical classes

Anna Kreshuk,

PH/SFT, CERN

Contents

- News in fitting
  - Linear fitter
  - Robust fitter
  - Fitting of multigraphs
- Multidimensional methods
  - Robust estimator of multivariate location and scatter
- New methods in old classes
- Future plans

Linear Fitter (0)

- To fit functions linear in parameters
  - Polynomials, hyperplanes, linear combinations of arbitrary functions
- TLinearFitter can be used directly or through TH1, TGraph, TGraph2D::Fit interfaces
- When used directly, can fit multidimensional functions

Linear Fitter (1)

- Special formula syntax:
  - Linear parts separated by “++” signs:
    - “1 ++ sin(x) ++ sin(2*x) ++ cos(3*x)”
    - “[0] + [1]*sin(x) + [2]*sin(2*x) + [3]*cos(3*x)”
  - Simple to use in multidimensional case
    - “x0 ++ x1 ++ exp(x2) ++ log(x3) ++ x4”
- Polynomials (pol0, pol1…) and hyperplanes (hyp1, hyp2, …) are the fastest to compute
- By default, polynomials in TH1, TGraph::Fit functions now go through Linear Fitter
- Data to be used for fitting is not copied into the fitter

Linear Fitter (2)

- Advantages in separating linear and non-linear fitting:
  - Doesn't require setting initial parameter values
  - The gain in speed

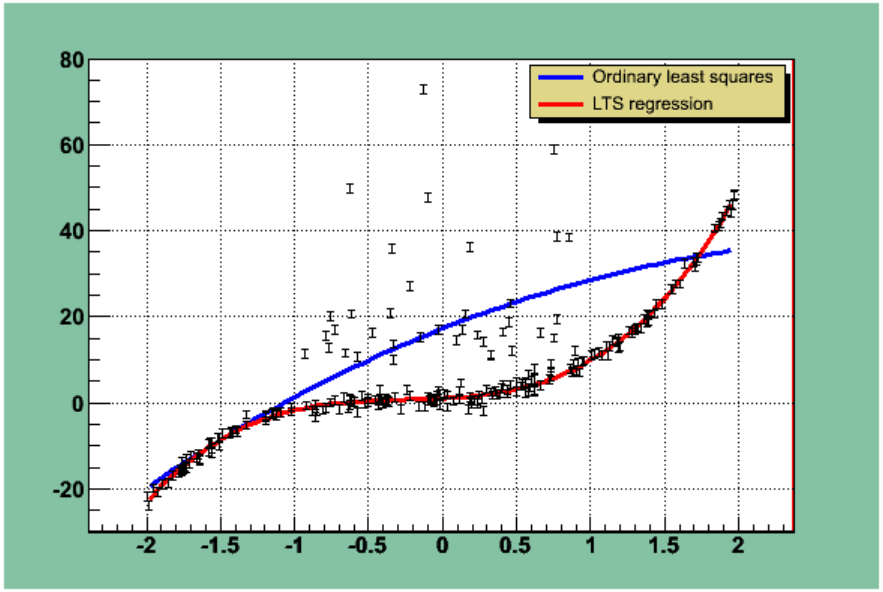
Function	Linear fitter	Minuit
Pol3 in TGraphErrors 1000 fits of 1000 points	Average CPU time 1.95	Average CPU time 30.54
TMath::Sin(x) + TMath::Sin(2*x)	Average CPU time 2.39	Average CPU time 21.34

Robust fitting (0)

- Least Trimmed Squares regression – extension of the TLinearFitter class
- Motivation: least-squares fitting is very sensitive to bad observations
- Robust fitter is used to fit datasets with outliers
- The algorithm tries to fit *h* points (out of *N*) that have the smallest sum of squared residuals

Robust fitting (1)

- High breakdown point
  - smallest proportion of outliers that can cause the estimator to produce values arbitrarily far from the true parameters



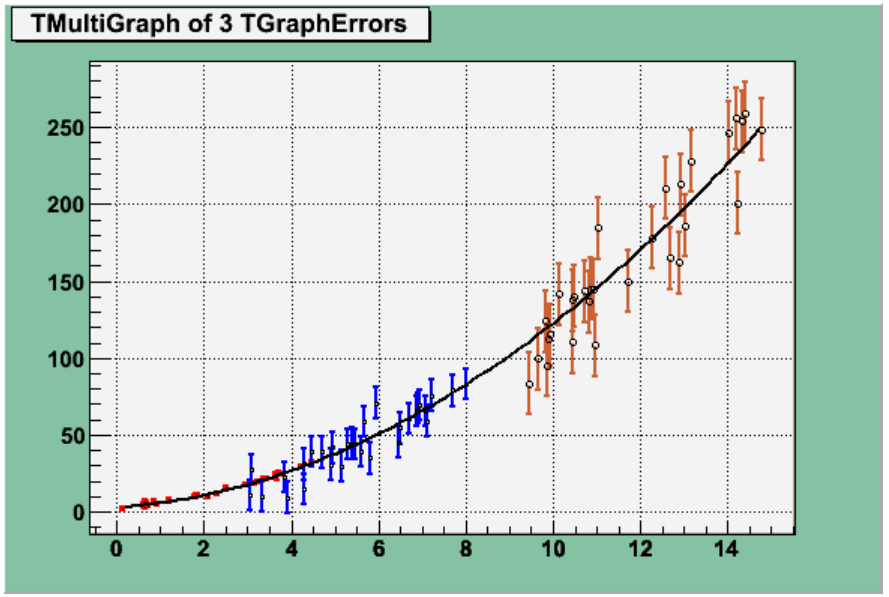

```
Graph.Fit("pol3", "rob=0.75", -2, 2);
```

2nd parameter – fraction *h* of the good points

TMultiGraph::Fit

A multigraph is a collection of graphs

- Fit function, implemented in this class, allows to fit all graphs simultaneously, as if all the points belong to the same graph
- All options of TGraph::Fit supported

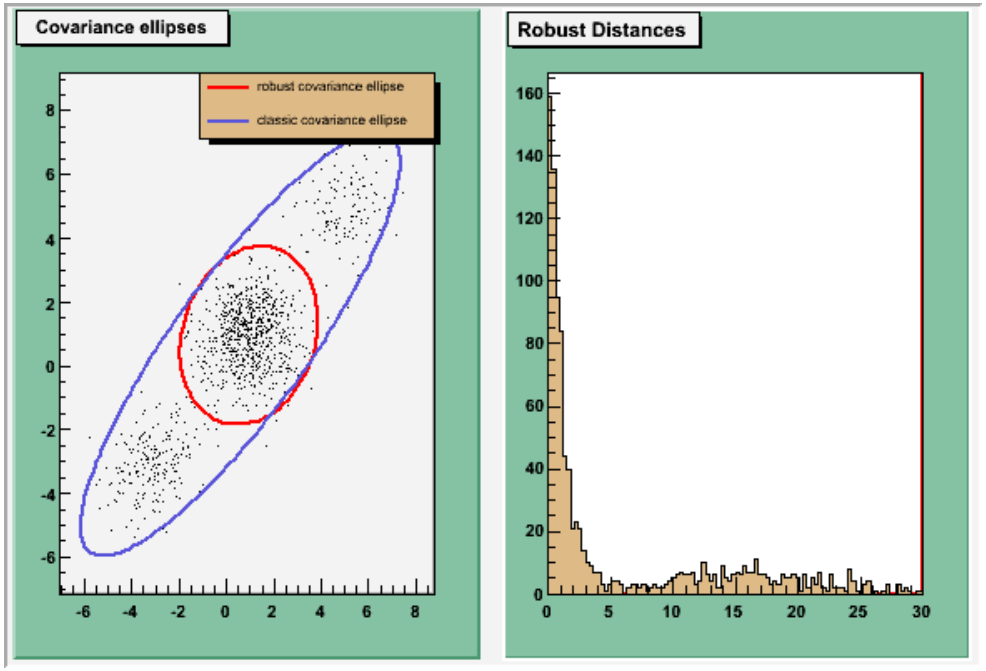


Multivariate covariance

- Minimum Covariance Determinant Estimator – a highly robust estimator of multivariate location and scatter
- Motivation: arithmetic mean and regular covariance estimator are very sensitive to bad observations
- Class TRobustEstimator
- The algorithm tries to find a subset of *h* observations (out of *N*) with the minimal covariance matrix determinant

Multivariate covariance

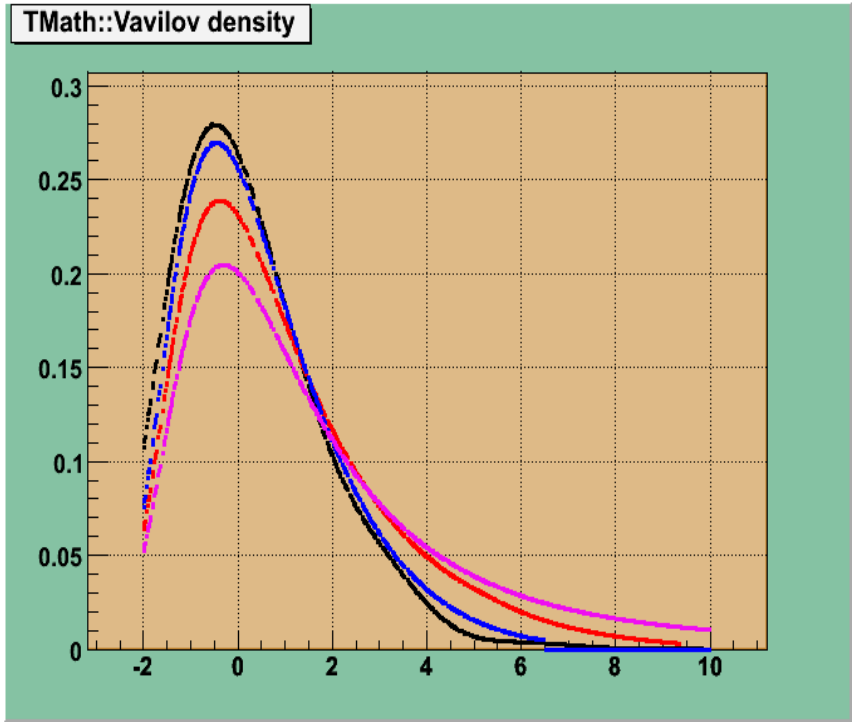
- High breakdown point
- Left – covariance ellipses of a 1000-point dataset with 250 outliers
- Right – distances of points from the robust mean, calculated using robust covariance matrix



- Indices of outlying points can be returned

News in TMath

- More distribution functions, densities and quantile functions
- Median (for weighted observations) and K-th order statistic
- Kolmogorov test for unbinned data

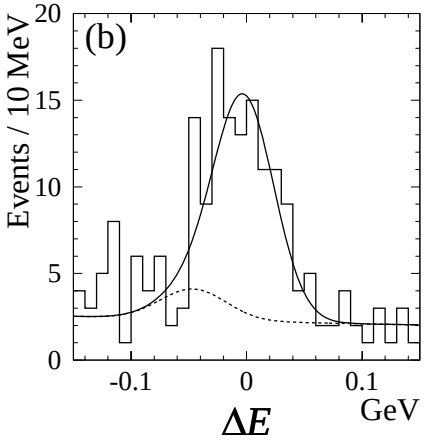


News in TH1 and TF1

- TH1:
  - Chi2 test
  - Mean & RMS error, skewness and kurtosis
- TF1:
  - Derivatives (1st, 2nd and 3rd)
  - Improved minimization – a combination of grid search and Brent's method (golden section search and parabolic interpolation)

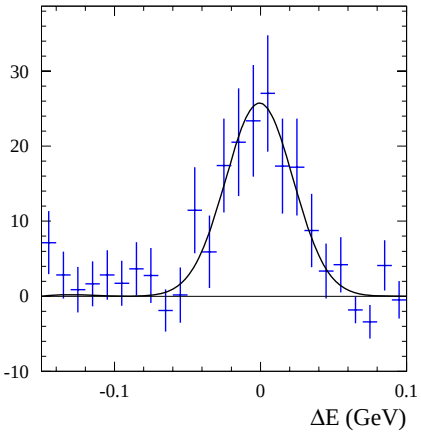
Future plans (0)

- Short term:
  - sPlot – A statistical tool to unfold data distributions
    - class TSPlot to be added soon



Left – Projection plot
cut on the likelihood ratio
Excess of events –
Signal? Background?

Right – *sPlot* – no cut
getting rid of background
by statistical methods
Signal!!!



More about *sPlot* in Muriel Pivk's presentation at 11:05

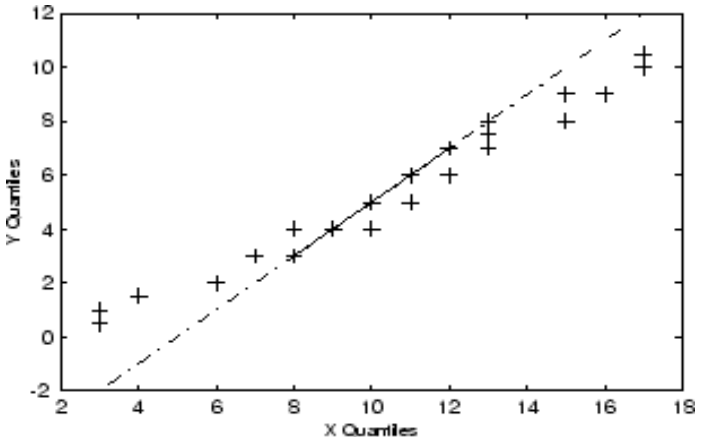
Future plans (1)

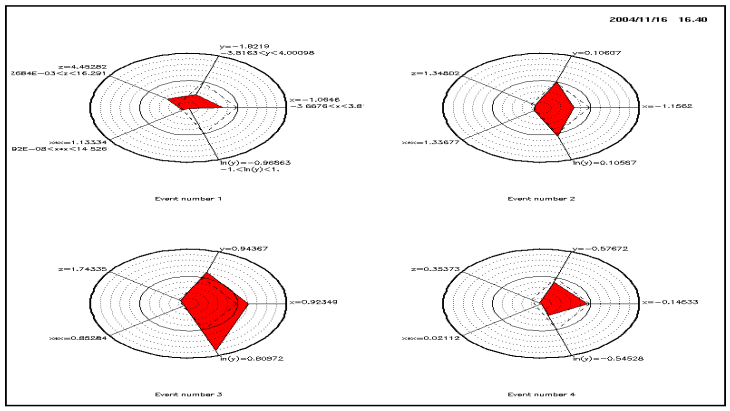
- From PHYSTAT05 conference in Oxford, September 2005:
  - Following a talk by Marc Paterno, an interface with R was discussed – ROOT Trees can already be read from R promt
  - Following a talk by Jim Linnemann, a repository of physics-oriented statistical software in Fermilab was discussed.

  - Rajendran Raja – Goodness of fit for unbinned likelihood fits
  - Nikolai Gagunashvili – Chi2 test for comparison of weighted and unweighted histograms
  - Martin Block – Outlier rejection and fitting with Lorentz weights
  - F.Tegenfeldt & J.Conrad – More on confidence intervals
  - Kyle Cranmer – More on hypothesis testing and confidence intervals
  - A lot of other interesting suggestions…

Future plans (2)

- Statistical plots
  - Quantile-quantile plot – useful for determining if 2 samples come from the same distribution
  - Boxplot
  - Spiderplot





Future plans (3)

- Loess – locally weighted regression
  - A procedure for estimating a regression surface by multivariate smoothing
- FFT
- Cluster analysis