Robust Location Estimators with Applications in HEP Data Analysis

R. Frühwirth, HEPHY, Vienna, Austria

University of Oslo, 16 June 2006

Outline

- **Introduction to robust location estimators**
- **Robust location estimators in 1D**
- **Application to energy loss data**
- **Robust location estimators and regression in 3D**
- **Application to vertex reconstruction**
- **Summary and outlook**

Introduction to robust location estimators

Introduction to robust location estimators

- A function $\mu(X)$ of a (continuous) random variable X is called a measure of location if

  1. $\mu(aX + b) = a\mu(X) + b$
  2. $X \geq 0 \Longrightarrow \mu(X) \geq 0$

  Property 1. is called affine equivariance.

- Examples: Mean, median, quantiles, linear combinations of quantiles, mode

- An estimator of $\mu(X)$ is called a location estimator.

- Examples: Sample mean, sample median, order statistics, linear combinations of order statistics, sample mode

Introduction to robust location estimators

- If the parent distribution has long tails or if the sample is contaminated by outliers a location estimator should be robust.
- Robustness of an estimator can be quantified by its breakdown point, the smallest proportion of outliers needed to make the estimator useless (unbounded).
- Examples:
  - The sample mean is not robust. Its finite-sample breakdown point is $\epsilon_n^* = 1/n$, and its asymptotic breakdown point is $\epsilon^* = 0$.
  - The sample median is highly robust. Its finite-sample and its asymptotic breakdown point is $\epsilon_n^* = \epsilon^* = 0.5$.

Robust location estimators in 1D

Robust location estimators in 1D

- Estimators based on modal intervals
  - In-depth study in Bickel and Frühwirth, CSDA 50 (2006).
  - LMS (least median of squares): minimize the median of the squared residuals.

$$\mu_{\mathrm{LMS}} = \arg\min_{\mu} \operatorname*{med}_{i} r_i^2(\mu), \quad r_i(\mu) = x_i - \mu$$

    - In 1D μ_{LMS} is the midpoint of the shortest interval containing 50% of the data.
    - The FSBP and the ABP are $\epsilon_n^* = \epsilon^* = 0.5$.

Robust location estimators in 1D

- Estimators based on modal intervals (cont)
  - Shorth: mean of the data in the shortest interval containing 50% of the data.
    - Breakdown point $\epsilon_n^* = \epsilon^* = 0.5$.
  - HSM (half sample mode): iterated LMS.
    - Find the shortest interval containing 50% of the data.
    - Keep only data in this interval and repeat until only two observations are left.
    - The HSM is the mean of the last two observations.
    - Breakdown point $\epsilon_n^* = \epsilon^* = 0.5$.

Robust location estimators in 1D

- Estimators based on modal intervals (cont)
  - FSM (fraction of sample mode): generalized HSM.
    - Each successive modal interval contains a fraction p of the data.
    - Breakdown point $\epsilon^* = \min(p, 1-p)$.
  - FSMW (fraction of sample mode with weights): generalized FSM.
    - Each observation has a weight assigned to it.
    - Each successive modal interval contains a fraction p of the total sum of the weights.

Robust location estimators in 1D

- Other estimators of the mode
  - EPDFM (mode of the empirical density function): maximum of the smoothed data histogram (kernel estimator).

$$\hat{f}(x) = \frac{1}{nh\sqrt{2\pi}} \sum_{i=1}^{n} \exp\left[-\frac{1}{2}\left(\frac{x - x_i}{h}\right)^2\right]$$

    - h is a smoothing parameter determined from a robust scale estimator, e.g. MAD (median of the absolute deviations from the median).

Robust location estimators in 1D

- Other estimators of the mode (cont)
  - PM (parametric mode): maximum of a pdf (e.g. normal) with location and scale estimated from the data.
    - If necessary, the data can be transformed for a better approximation by the chosen family.
    - The location and scale estimates can be classical (mean and standard deviation) or robust (median and MAD).

Robust location estimators in 1D

- Estimators based on order statistics
  - L-estimators: linear combinations of order statistics.

    $$\mathcal{L}(x_1, \ldots, x_n) = w_1 x_1 + \ldots + w_n x_n$$

    Affine equivariance requires $\sum w_i = 1$.
    - The breakdown point depends on which weights are different from zero. If i is the smallest and j is the largest index with a weight different from zero, the FSBP is

      $$\epsilon_n^* = \min(i, n - j + 1)/n$$

    - Special cases: median, truncated means

Robust location estimators in 1D

- Estimators based on trimming or down-weighting
  - LTS (least trimmed squares): minimize the sum of the h smallest squared residuals.

$$\mu_{\mathrm{LTS}} = \arg\min_{\mu} \sum_{i=1}^{h} r_{[i]}^2(\mu), \quad r_i(\mu) = (x_i - \mu)/\sigma_i$$

    - Exact computation by exhaustive search (prohibitive), approximate computation by iteration.
    - Maximum breakdown point is $\epsilon_n^* = 0.5$ for $h = n/2 + 1$.

Robust location estimators in 1D

- Estimators based on trimming or down-weighting (cont)
  - M-estimator: minimize the sum of a function of the residuals.

$$\mu_{\mathrm{M}} = \arg\min_{\mu} \sum_{i=1}^{n} \rho(r_i(\mu)), \quad r_i(\mu) = (x_i - \mu)/\sigma_i$$

    - $\rho(x) = x^2$ gives the mean, $\rho(x) = |x|$ gives the median (if all σ_i are equal).
    - $\rho(x)$ should grow less quickly than a quadratic function.
    - Computation by iterated re-weighted mean, with weights

$$w_i = \frac{\psi(r_i)/\sigma_i}{r_i}, \quad \psi(x) = \rho'(x)$$

Robust location estimators in 1D

- Estimators based on trimming or down-weighting (cont)
  - Adaptive estimator: minimize the sum of weighted squares residuals.

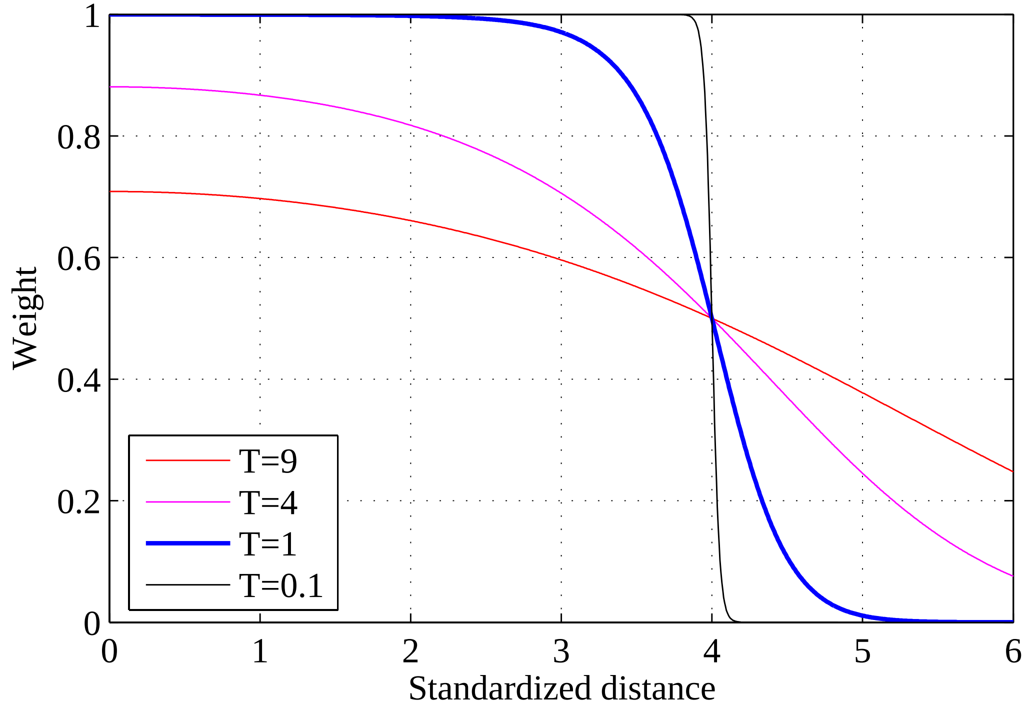
$$\mu_{\mathrm{M}} = \arg\min_{\mu} \sum_{i=1}^{n} w_i \, r_i^2(\mu), \qquad r_i(\mu) = (x_i - \mu)/\sigma_i$$

    - The weights describe the compatibility of the observation with the estimate:

$$w_i = \frac{\exp(-r_i^2/2T)}{\exp(-r_i^2/2T) + \exp(-C^2/2T)}$$

Robust location estimators in 1D

- C is a cut-off, T is a temperature.
- Computation as an iterated re-weighted mean.



Robust location estimators in 1D

- Estimators based on residuals can be immediately generalized to robust regression.
  - LMS regression
  - LTS regression
  - M regression
  - Adaptive regression
- Combinatorial optimization
- Dependent on starting value

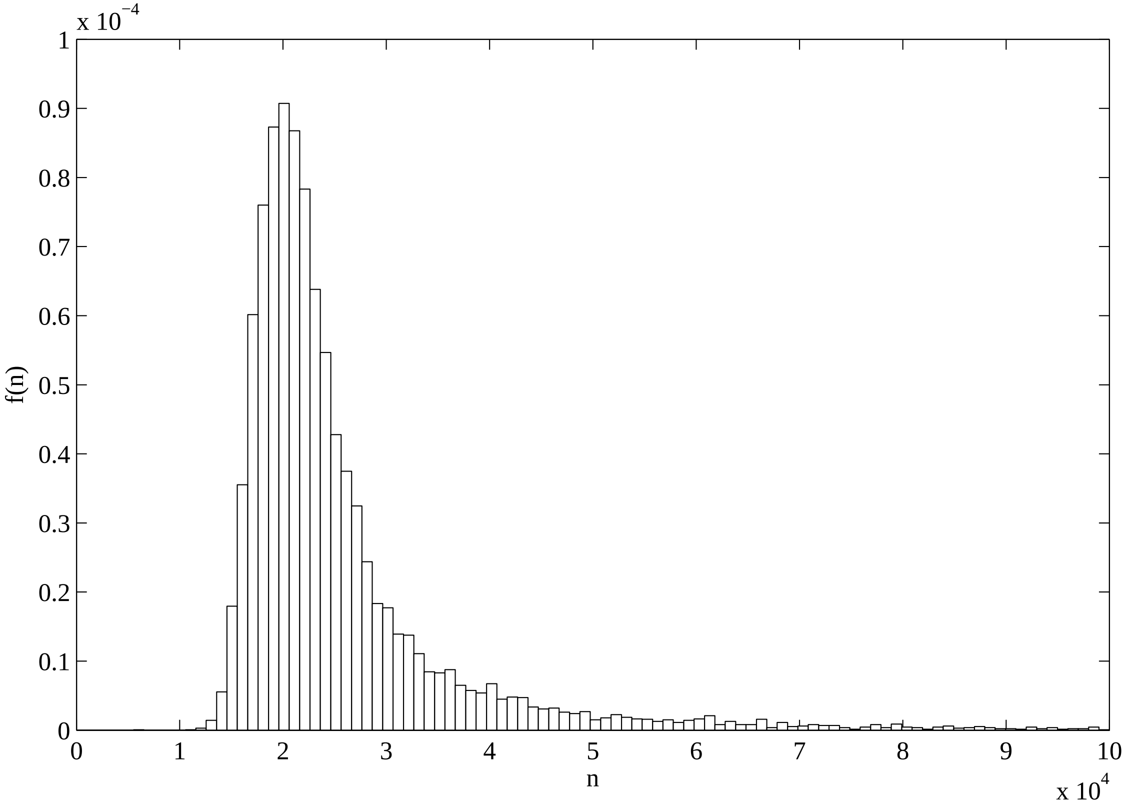
Application to energy loss data

Application to energy loss data

- Joint work with M. Regler and M. Friedl (ICATTP 2005).
- The experimental data
  - Beam (KEK): Pions with $p = 4\ \mathrm{GeV}/c$
  - Detector: 300 micron silicon wafer, 51 micron pitch
  - Double-sided, data from p side
  - Amplification by the APV25 front-end chip (developed for CMS)
  - Low electronic noise (739 e-h pairs)
  - Pedestal subtraction
  - Common mode correction

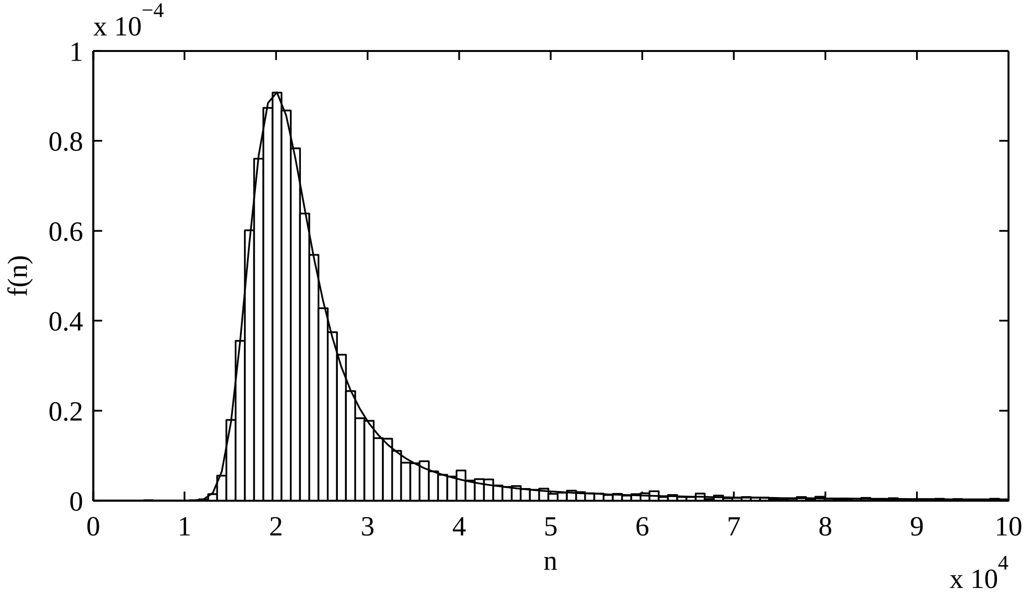
Application to energy loss data

- Signal data



Application to energy loss data

- Data modelling
  - Model: Bichsel convoluted with Gaussian noise
  - Perfect fit
  - Fitted noise width (723) agrees well with experiment (739)



Application to energy loss data

- Generation of random samples
  - Generate 50000 random samples of size 15 from the experimental distribution.
  - Add various amounts of uniform noise.
  - Study (parametric) maximum likelihood and (non-parametric) L-estimators.
  - Maximum likelihood uses Bichsel model, convoluted with Gaussian

Application to energy loss data

- Optimal L-estimator
  - Given a location to be estimated, there is an equivariant L-estimator $\mathcal{L}_{\mathrm{opt}}$ with the smallest variance.
  - The optimal weights are estimated from the data, using the first two moments of the order statistics:

$$\vec{\mu} = E(\vec{x}), \quad C = \mathrm{cov}(\vec{x})$$

  - The two largest order statistics have very large variance and are dropped.
  - More order statistics can be dropped for higher robustness.

Application to energy loss data

- Under the constraints $E(\mathcal{L}_{\rm opt}) = t$ and $\sum w_i = 1$ the optimal weights are determined by minimizing

$$M(\vec{w}) = \vec{w}^{\rm T} C \vec{w} - 2\alpha(\vec{w}^{\rm T}\vec{\mu} - t) - 2\beta(\vec{w}^{\rm T}\vec{e} - 1)$$

- The explicit solution is given by

$$\vec{w}_{\rm opt} = \Delta^{-1} B(t\vec{e} - \vec{\mu})$$

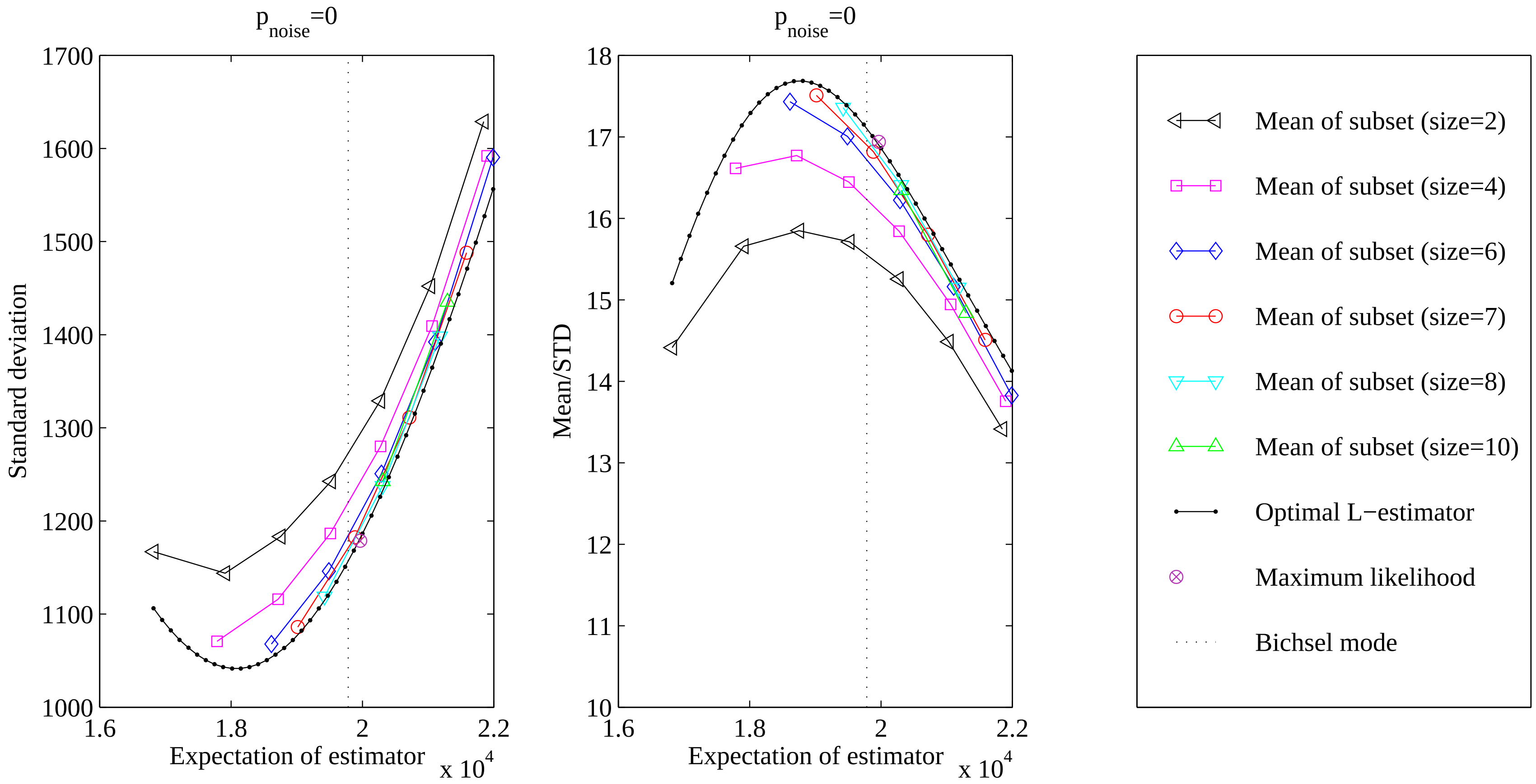
with

$$B = C^{-1}(\vec{\mu}\vec{e}^{\rm T} - \vec{e}\vec{\mu}^{\rm T})C^{-1},\ \Delta = \vec{\mu}^{\rm T} B\vec{e},\ \vec{e} = (1, \ldots, 1)^{\rm T}$$

Application to energy loss data

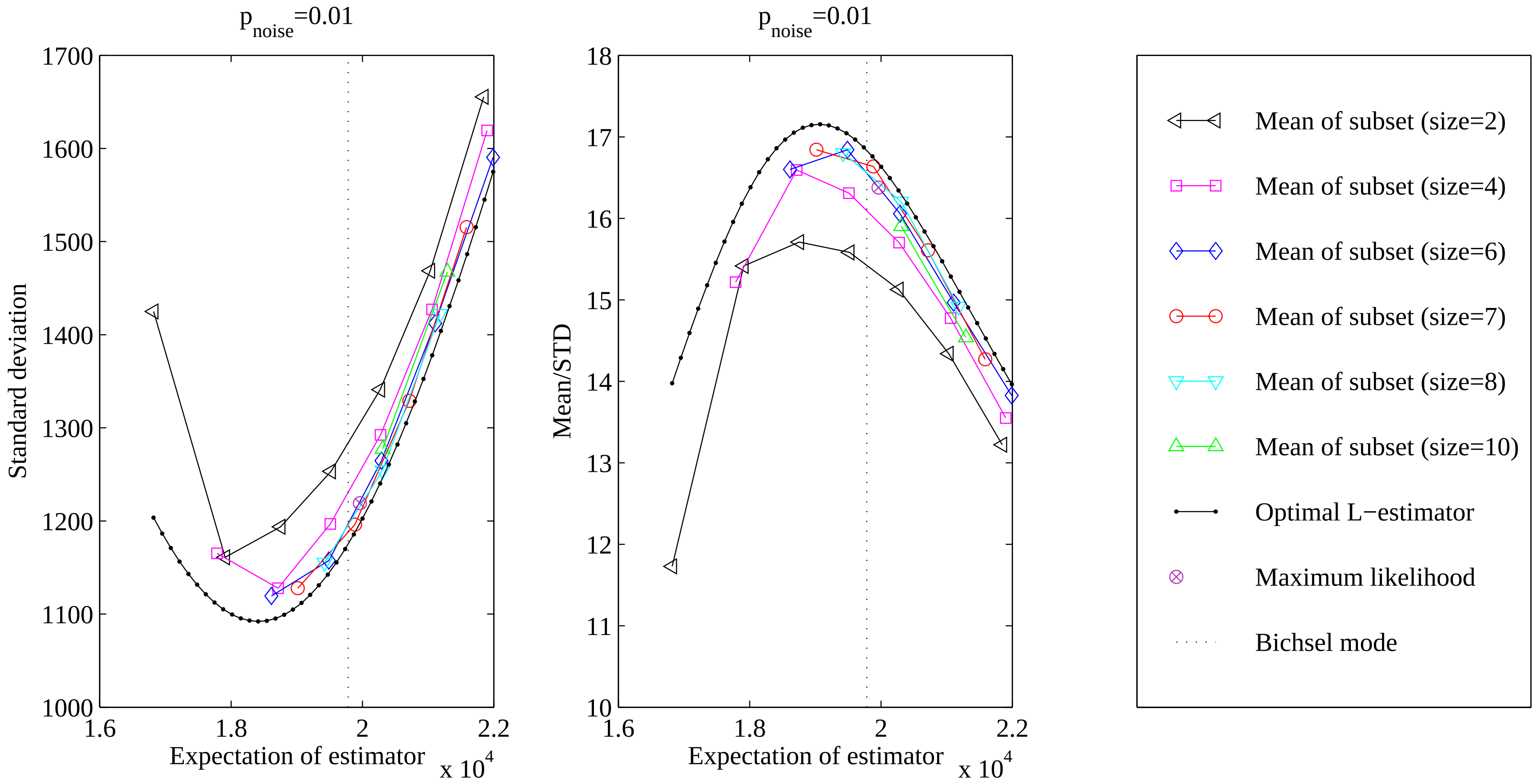
- Comparison of the separation power (ratio of mean to standard deviation)
  - Truncated means of subsets of size 2,4,6,7,8,10
  - Optimal L-estimator
  - ML-estimator
- Contamination with various amounts of noise (uniform in the range of the observations)

Application to energy loss data



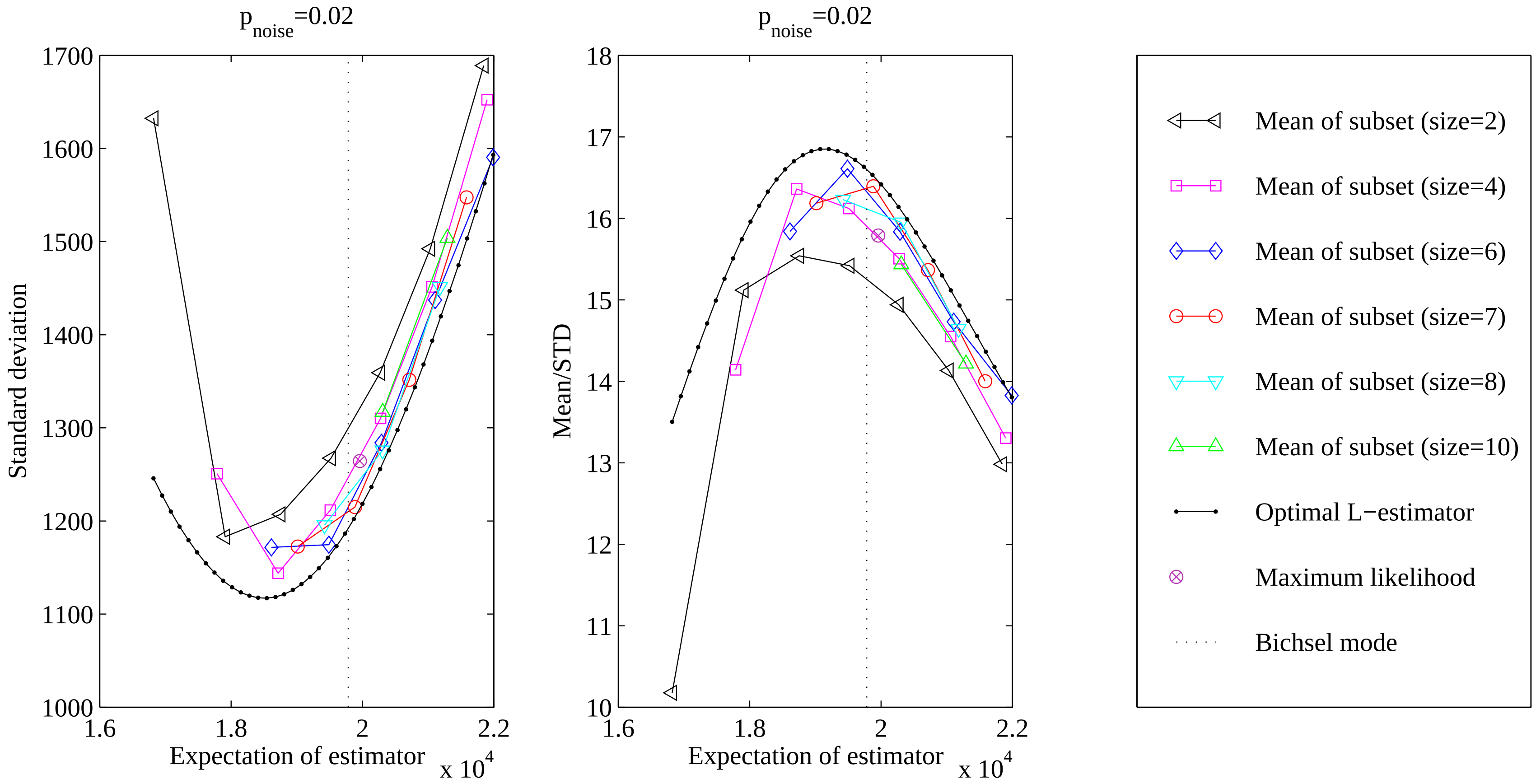
No noise

Application to energy loss data



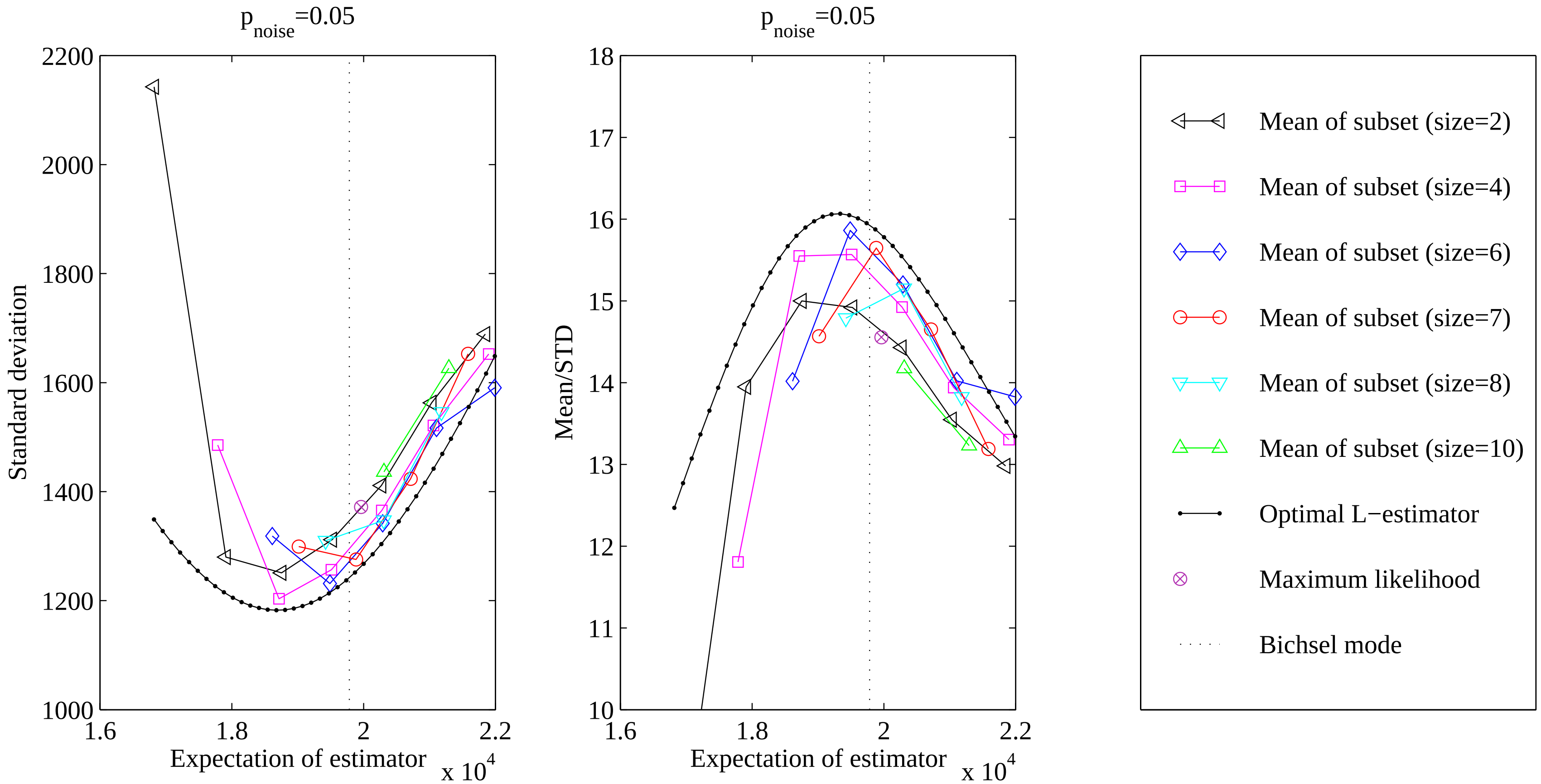
1% noise

Application to energy loss data



2% noise

Application to energy loss data



5% noise

Application to energy loss data

- Separation ratio ρ drops by about 5% when 2% noise are added
- Optimal L-estimator slightly better than truncated means
- Without noise, means of observations from 1 to 6 and 1 to 7 are nearly as good
- With noise, mean of observations from 2 to 7 is nearly as good
- Estimation of median (mean of observations 7 and 8) gives inferior separation
- ML estimator is not equivariant, slightly worse than the best L-estimators and much slower to compute

Robust location estimators and regression in 3D

Robust location estimators and regression in 3D

- True 3D location estimators:
  - LMS: the point with minimizes the median of the (squared) distances to all data points, center of the smallest sphere covering half of the data
  - MVE (Minimum volume ellipsoid): center of the minimal ellipsoid covering h data points ($h > n/2$)
  - MCD (Minimum covariance determinant): Mean of the h points for which the determinant of the covariance matrix is minimal ($h > n/2$).
- Computation tends to be slow

Robust location estimators and regression in 3D

- Coordinatewise 1D location estimators:
  - Coordinatewise median, LMS, HSM, FSM, . . .
- Cannot be affine equivariant (only exception: arithmetic mean!)
- Need not be inside the convex hull of the observations
- But: much faster to compute
- Important for fast and robust initialization of vertex fit

Robust location estimators and regression in 3D

- LTS, M-estimator, Adaptive estimator:
  - Straightforward generalization from 1D case, both for robust location estimation and for robust regression
  - Implemented as iterated re-weighted LS-estimator (regression or Kalman Filter)
- Used for robust precision estimates of interaction vertices

Application to vertex reconstruction

Application to vertex reconstruction

- Need fast and robust location estimators for approximate initial estimate (expansion point)
- Need precise but robust regression for estimation of vertex position and track momenta
- Finding of secondary vertices by iterated robust estimation
  - Compute robust estimate of vertex position
  - Find tracks not compatible with estimated vertex
  - Repeat on set of incompatible tracks

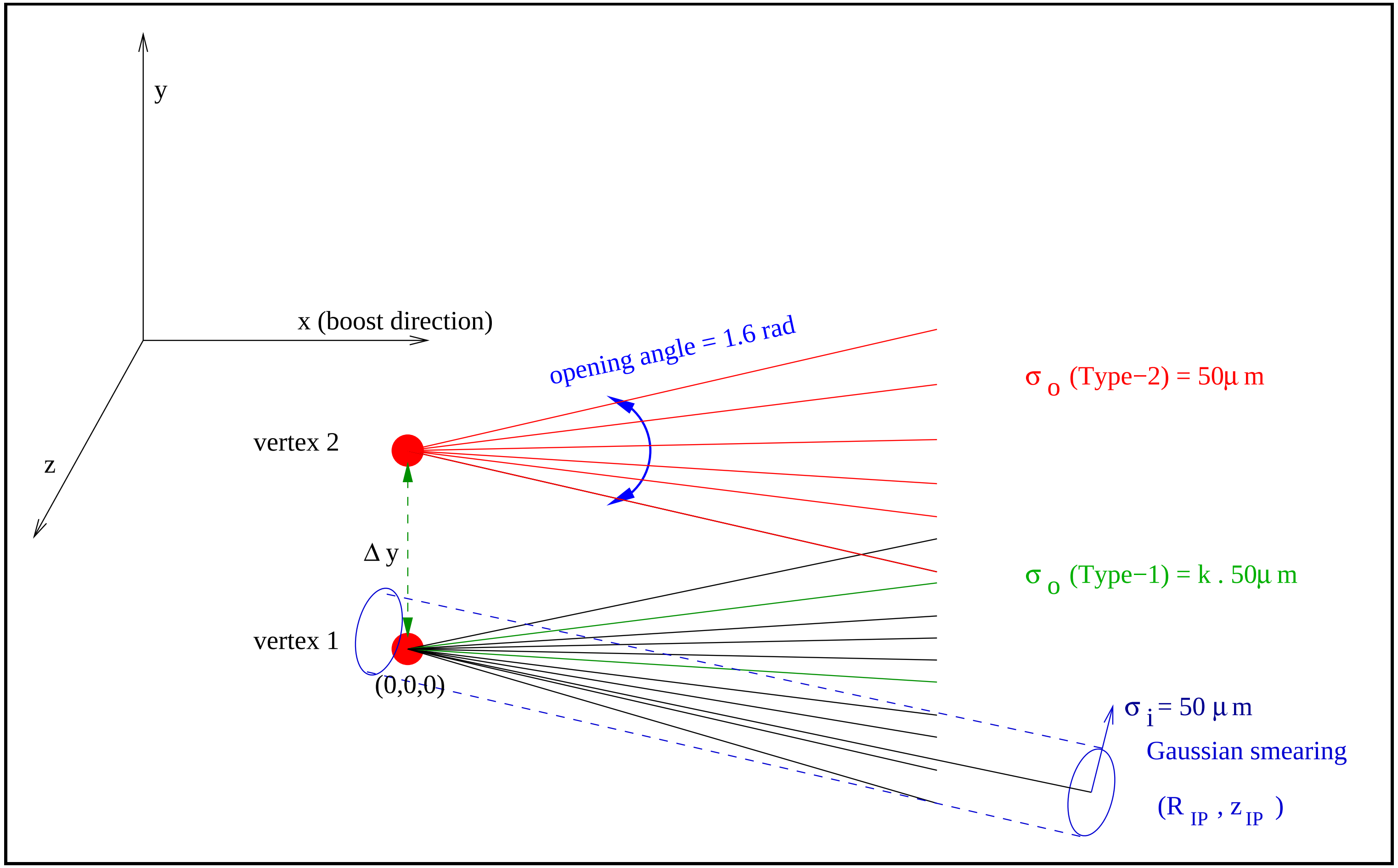
Application to vertex reconstruction

- Studies
  - Study of LS, LTS and adaptive estimator with synthetic data (J. D'Hondt et al., IEEE TNS)
  - Robust vertex reconstruction in CMS with simulated data (W. Waltenberger, thesis)
  - Gaussian-sum filter for vertex reconstruction (R. Fruehwirth and T. Speer, CPC)

Application to vertex reconstruction

- Sensitivity of Robust Vertex Fitting Algorithms
  J. D'Hondt, R. Frühwirth, P. Vanlaer, W. Waltenberger
  IEEE Trans. Nucl. Sci. 51 (2004) 2037.
  - Performance of LS, LTS and AVF on synthetic vertices
  - Type-1 outliers: variance inflation
  - Type-2 outliers: shift
  - Various multiplicities

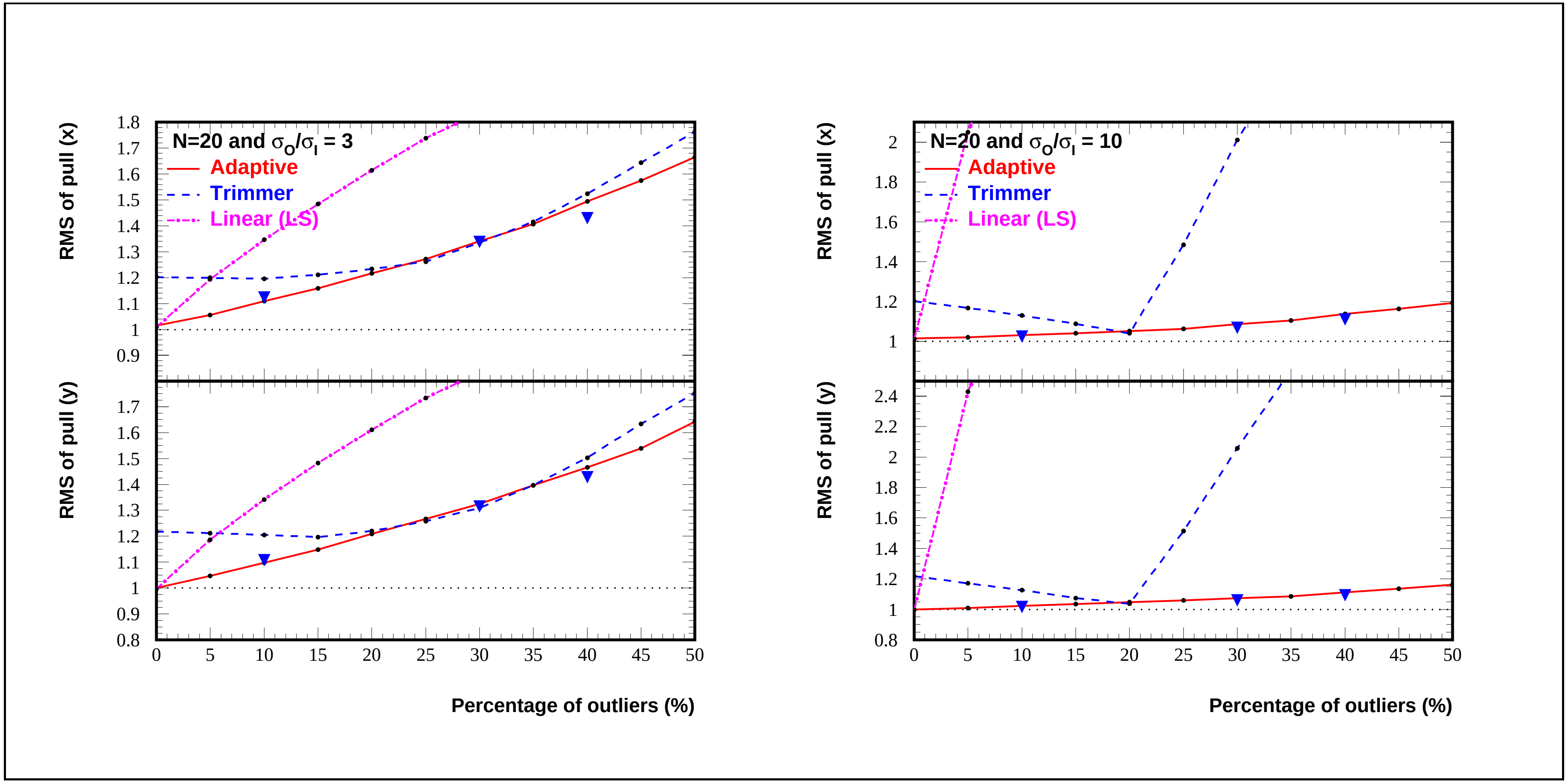
Application to vertex reconstruction



Application to vertex reconstruction

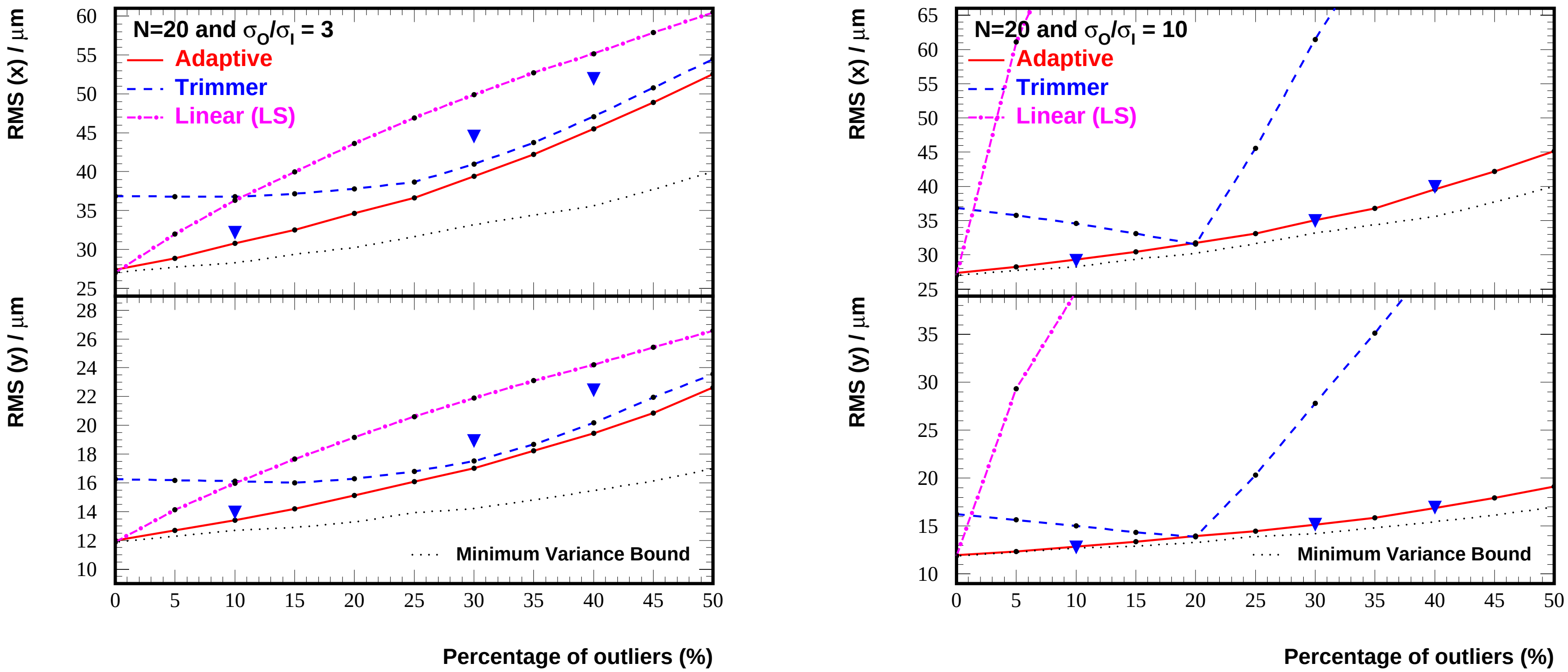
- Study of type-1 outliers
  - $\sigma_{\text{out}}/\sigma_{\text{inl}} = 3, 10$
- 20 tracks in total
- Look at pulls, residuals, χ^2-probability

Application to vertex reconstruction



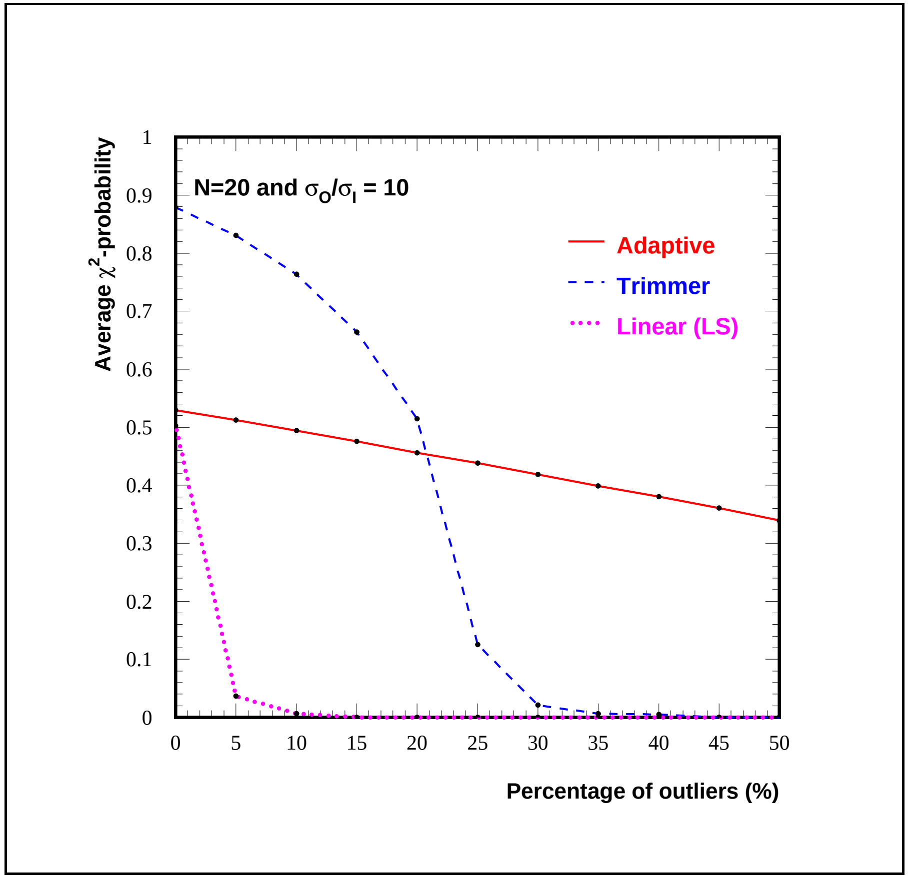
RMS of pulls

Application to vertex reconstruction



RMS of residuals

Application to vertex reconstruction

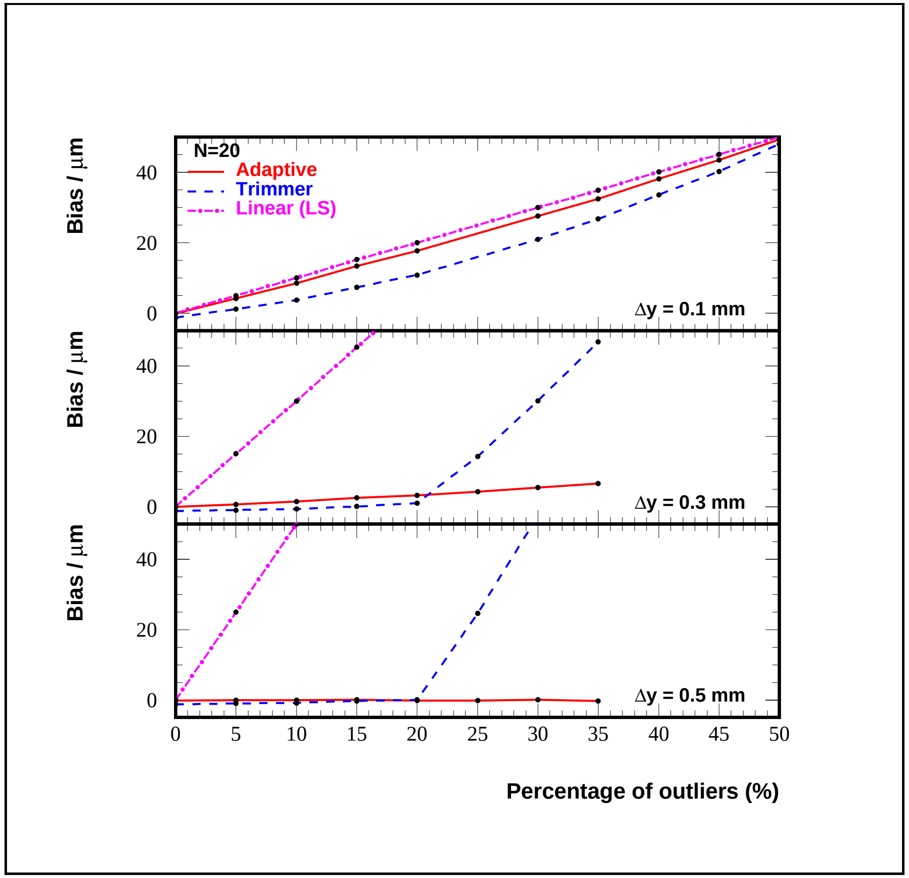


Average χ^2-probability

Application to vertex reconstruction

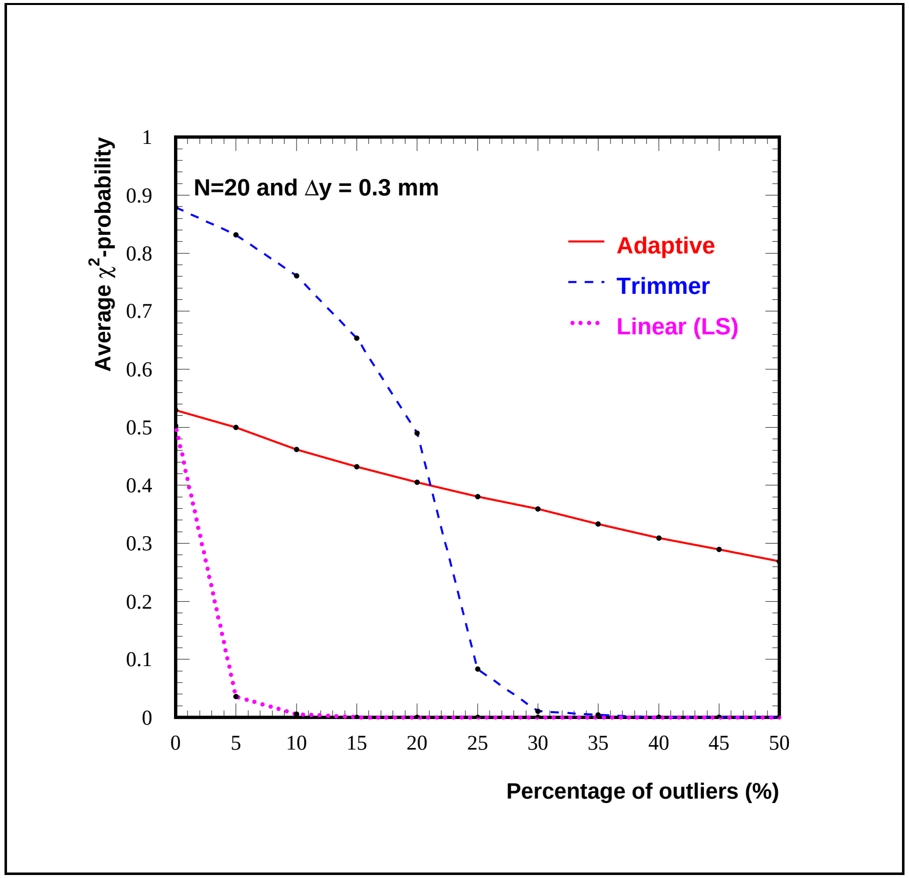
- Study of type-2 outliers
  - 20 tracks, $\Delta y = 100, 300, 500\mu$
- Look at bias, χ^2-probability

Application to vertex reconstruction



Average bias

Application to vertex reconstruction



Average χ^2-probability

Application to vertex reconstruction

- Fast initial vertex estimate
  - Should be where the majority of the tracks is
  - Based on crossing point of track pairs
  - The crossing point is halfway between the two points of closest approach
  - Each crossing point gets a weight inversely proportional to the distance
- Coordinatewise estimators
  - LMS, HSM, FSMW
  - Use all crossing points or only a subset

Application to vertex reconstruction

- 3D estimators
  - SMS (approximate LMS): crossing point that minimizes the median of the squared distances
  - ISMS (iterated SMS): repeat SMS on the data points below the median
- Baseline estimators
  - Zero: returns $(0, 0, 0)$
  - Monte Carlo: simulated vertex
  - vertex fitted by LS

Application to vertex reconstruction

- Comparison with $c\bar{c}$ and $q\bar{q}$ events

Estimator	$c\bar{c}$ RMS $[\mu m]$	$c\bar{c}$ σ_{Fit} $[\mu m]$	$c\bar{c}$ fail %	$c\bar{c}$ t $[ms]$	$q\bar{q}$ RMS $[\mu m]$	$q\bar{q}$ σ_{Fit} $[\mu m]$	$q\bar{q}$ fail %	$q\bar{q}$ t $[ms]$
SubsetHSM(-1)	38	29	0	3.5	37	25	0	4
HSM(-1)	37	27	0	3.4	32	24	0	3.9
LMS(-1)	196	41	0.2	3.4	237	44	1.1	3.9
HSM(100)	54	38	0	0.7	48	32	0	0.8
FSMW(-1)	38	27	0	16.1	34	24	0	22.9
ISMS(200)	33	27	0	8.2	30	26	0.1	8.1
FSMW(200)	49	33	0	3	43	29	0	3

Resolutions and failure rates for quark pairs

Application to vertex reconstruction

- Comparison with low-multiplicity vertices

Estimator	$J/\psi\ \phi \rightarrow K^+K^-\mu^+\mu^-$ RMS $[\mu\text{m}]$	σ_{Fit} $[\mu\text{m}]$	fail %	t $[ms]$	$\tau^\pm \rightarrow \pi^\pm\pi^+\pi^-$ RMS $[\mu\text{m}]$	σ_{Fit} $[\mu\text{m}]$	fail %	t $[ms]$
SubsetHSM(-1)	744	61	19.4	0.2	1018	780	37.9	0.1
HSM(-1)	744	61	19.4	0.2	1020	780	37.8	0.1
LMS(-1)	833	139	17.0	0.2	1035	937	38.4	0.1
HSM(100)	744	61	19.4	0.2	1020	780	37.8	0.1
FSMW(-1)	578	63	14.0	0.3	967	755	32.3	0.2
ISMS(200)	736	60	16.2	0.4	1010	876	38.2	0.2
FSMW(200)	578	64	14.0	0.2	967	755	32.3	0.1

Resolutions and failure rates for low-multiplicity vertices

Application to vertex reconstruction

- Influence of the initial estimate on the final (adaptive) vertex fit
  - vtx: reconstructed vertex can be associated to the simulated vertex
  - $<$cut: reconstructed vertex closer than 200μ from the simulated vertex
  - FSMW(200) with $p = 0.4$ is the default algorithm

Application to vertex reconstruction

Estimator	$c\bar{c}$ (jetfilter)		$q\bar{q}$ (jetfilter)		$J/\psi\ \phi$		$\tau \longrightarrow \pi\pi\pi$	
	vtx	$<$ cut	vtx	$<$ cut	vtx	$<$ cut	vtx	$<$ cut
LMS	1969	1849	1963	1881	1901	1120	1578	250
FSMW(200)	1986	1892	1993	1942	1948	1285	1778	272
Zero	443	381	435	385	407	147	167	23
MonteCarlo	1993	1897	2000	1999	1984	1409	1678	348
LinVtxFit	1955	1840	1935	1854	1926	1305	1648	261

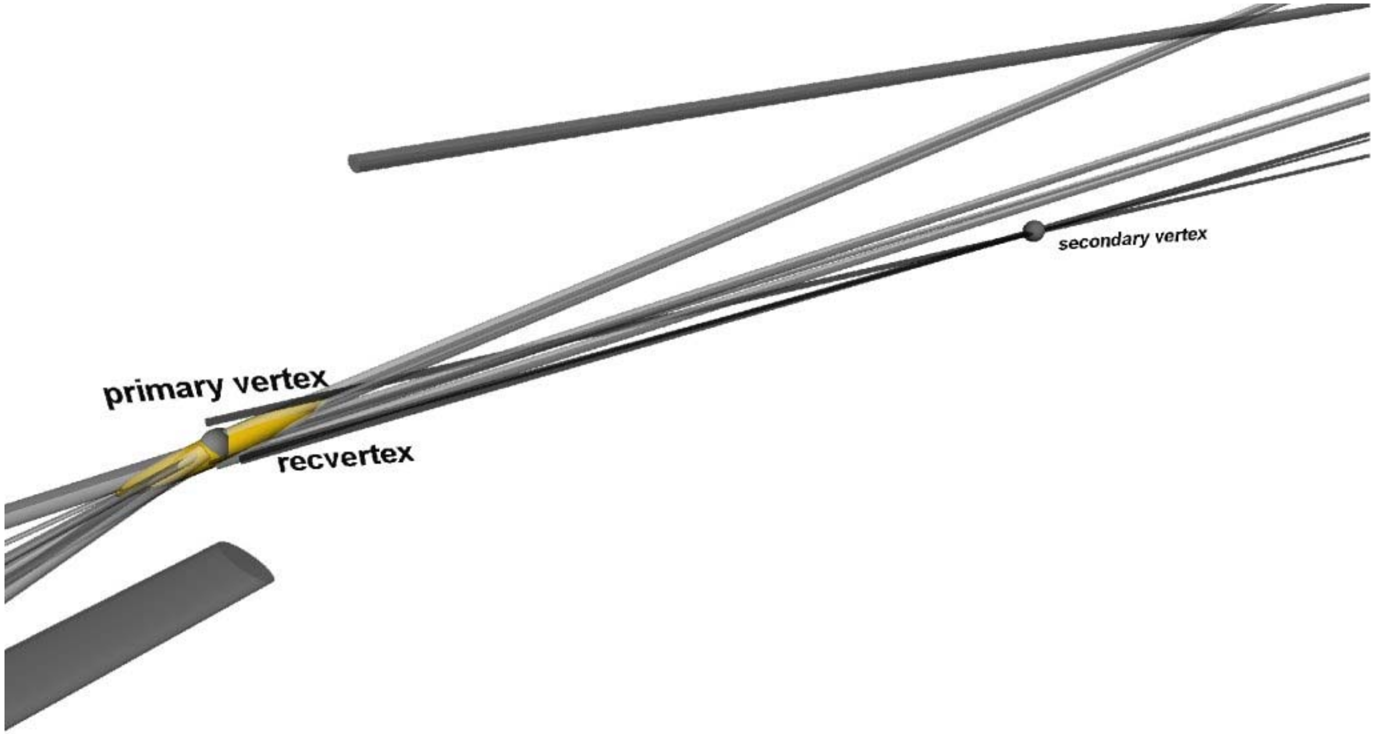
Influence of the initial estimate on the final (adaptive) vertex fit

Application to vertex reconstruction

- Final precision vertex estimate: Adaptive fitter
  - Iterated re-weighted Kalman filter
  - Linearization point and initial weights from fast robust initial estimate
  - Fast annealing: $T = 256, 64, 4, 1, 1, \ldots$
  - Iteration stops when $T = 1$ and $\Delta v < 1\mu\mathrm{m}$

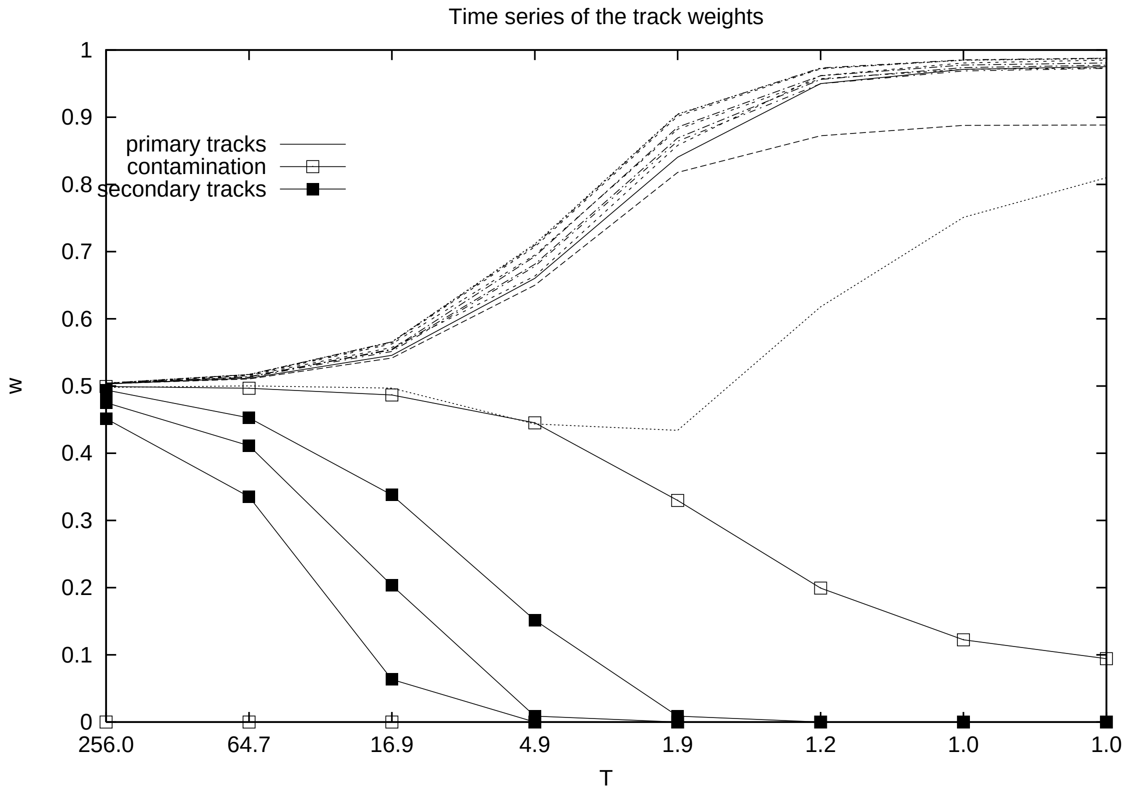
Application to vertex reconstruction

- Detailed study and comparison with Kalman filter in PhD thesis by W. Waltenberger



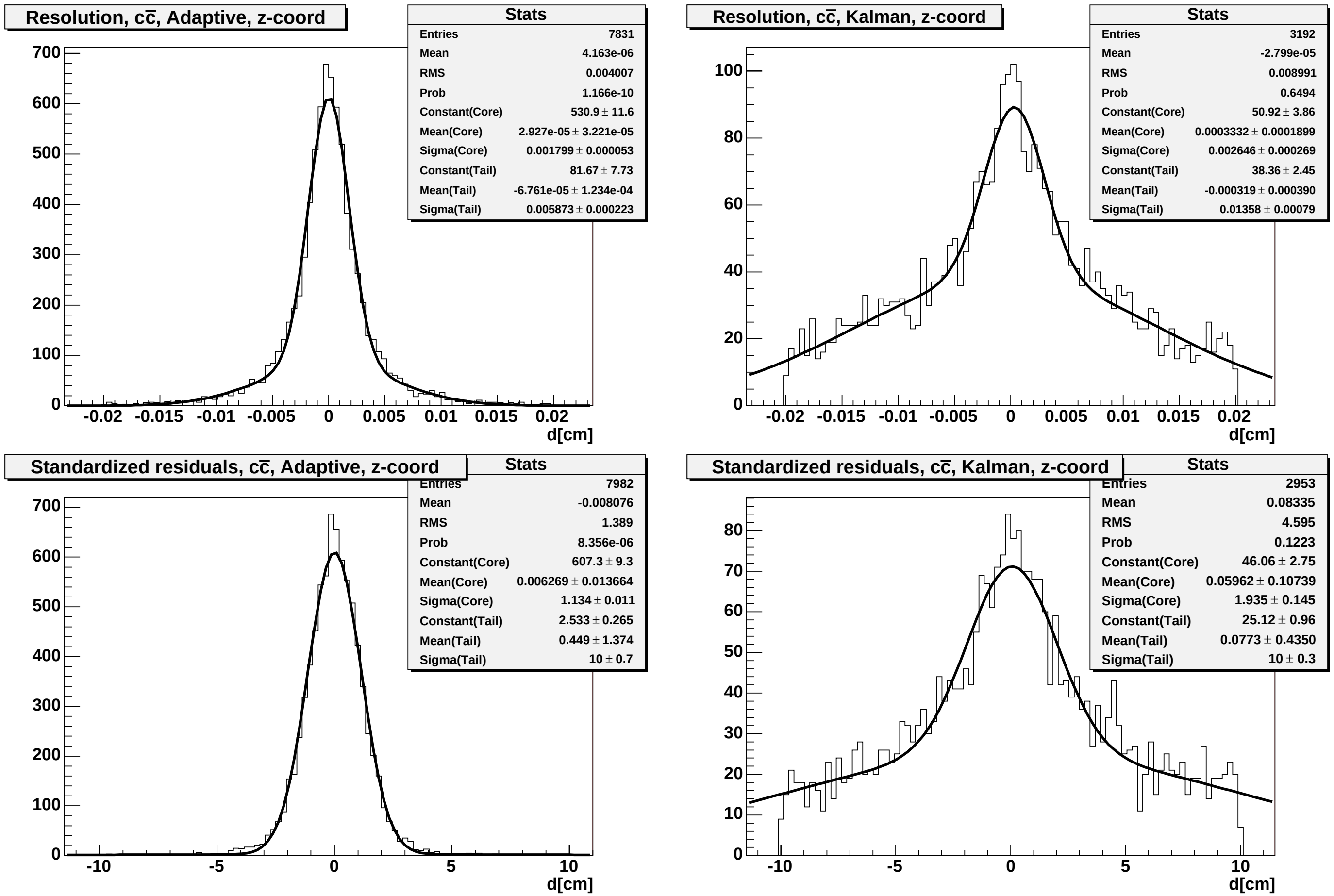
Example of a $c\bar{c}$ event

Application to vertex reconstruction



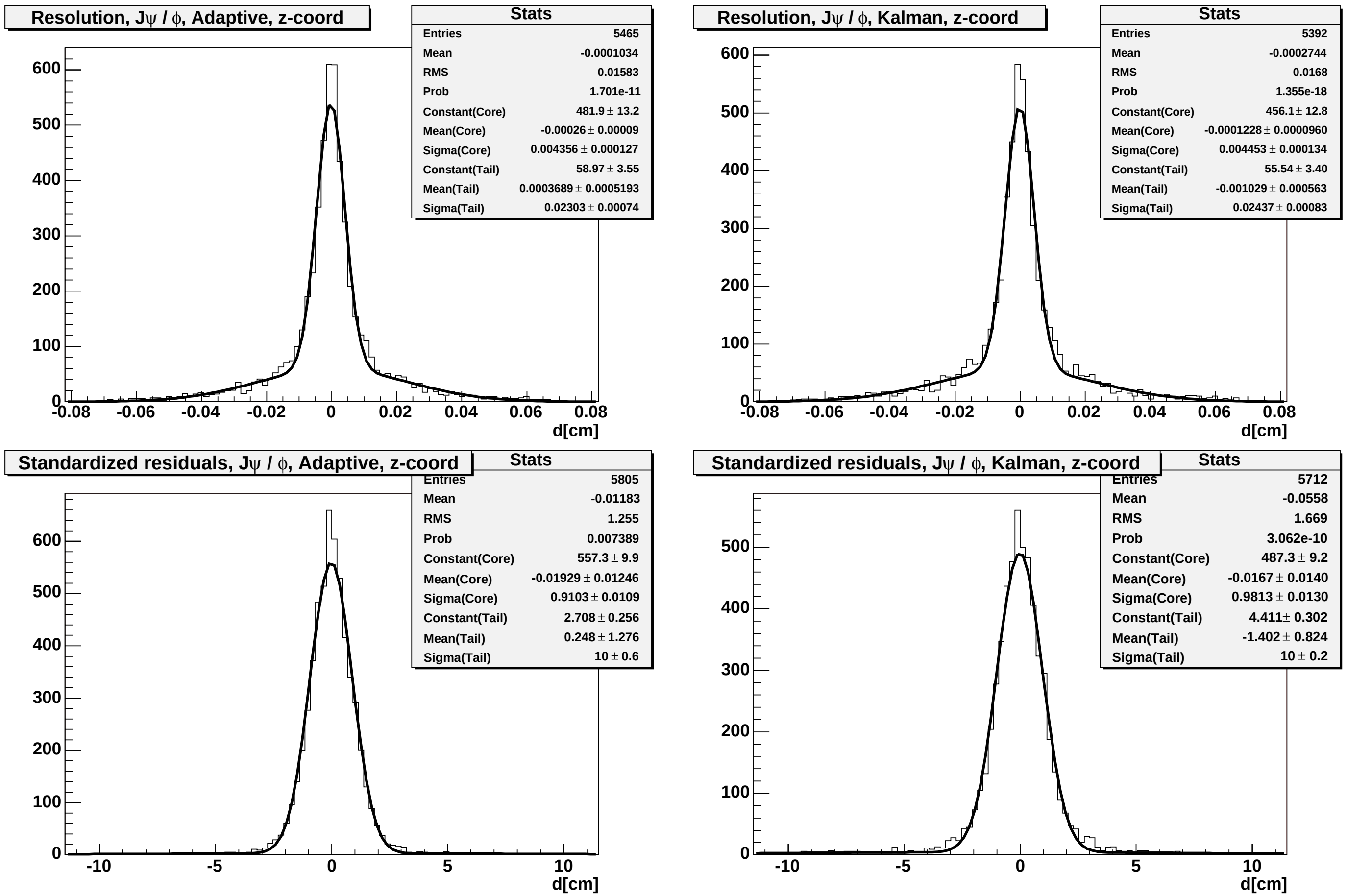
Evolution of the track weights

Application to vertex reconstruction



Comparing AVF with KVF, $c\bar{c}$ channel

Application to vertex reconstruction



Comparing AVF with KVF, $J/\psi\ \phi$ channel

Summary and outlook

Summary and outlook

- Robust location estimators useful in many situations
- L-estimators can be optimized to the distribution
- Mode estimators find the concentration point of the observations without any assumption on the distribution
- Adaptive estimators downweight outliers without assumption on the type or level of the contamination
- Without contamination adaptive estimators are very close to the optimal linear estimator

Summary and outlook

- Iterated adaptive estimator can be used for vertex finding
  - Adaptive fit of "main" vertex
  - Repeat on set of rejected tracks
  - Currently under test
- Adaptive estimator can be generalized for simultaneous fitting of several vertices (multi-vertex fit)
  - Start with several vertex candidates
  - Weights are modified such that vertices compete for the tracks
  - Some simple tests with synthetic data

Summary and outlook

- ORCA implementation being ported to new framework CMSSW
- Spin-off: Vertex reconstruction toolkit RAVE (W. Waltenberger, F. Moser, W. Mitaroff)
  - Currently source-code compatible with CMSSW
  - Experiment independent, open to extensions
  - Vertex finding (clustering) and vertex fitting (estimation)
  - Intended for use in ILC and BELLE(?)
- Simple stand-alone framework (VERTIGO) for fast debugging, visualization and analysis

Summary and outlook