Track-based alignment using a Kalman filter technique

R. Frühwirth and E. Widl[a]

[a] *Institute for High-Energy Physics*
Vienna, Austria

Abstract
An iterative method for track-based global alignment is proposed. It is derived from the Kalman filter and designed to avoid the inversion of large matrices. The update formulas for the alignment parameters and for the associated covariance matrix are described. The implementation and the computational complexity are discussed, and we show how to limit the latter to an acceptable level by restricting the update to detectors that are close in the sense of a certain metric. The performance of the Kalman filter in terms of precision and speed of convergence is studied with simulated tracks. Results from an implementation in the CMS reconstruction program CMSSW are presented, using two sections of the barrel part of the CMS Tracker.

2.1 Introduction

This note describes a method for global alignment with tracks that does not require solving a large system of linear equations. The method is iterative, based on the Kalman filter equations [1–3]. The alignment parameters are updated immediately after a track is processed, and the current best estimates of the alignment parameters are used in the track fit. The update of the alignment parameters is not restricted to the modules crossed by the track, but limited to modules with significant correlations to the ones in the current track. In order to keep track of the correlations some bookkeeping is required.

The Kalman filter equations offer the possibility to use prior information about the alignment from mechanical or laser alignment, and it is easy to fix the position of reference modules. The method is also highly suitable for alignment relative to another detector.

Table 2.1: Notation for alignment-related objects

N	total number of alignable detector modules
$\boldsymbol{d}_\mathrm{t}$	vector of true alignment parameters
$\boldsymbol{d}_0$	expansion point of alignment parameters
$\boldsymbol{d}$	current estimate of alignment parameters
$\boldsymbol{d}_i$	alignment parameters of detector module i
$\boldsymbol{D}$	covariance matrix of $\boldsymbol{d}$
$\boldsymbol{D}_{ij}$	submatrix of covariances between modules i and j
$\widehat{\boldsymbol{d}}$	updated estimate of alignment parameters
$\widehat{\boldsymbol{D}}$	covariance matrix of $\widehat{\boldsymbol{d}}$

2.2 Sequential updating

Tables 2.1 and 2.2 show a synopsis of the notation that is used in the following. All vectors are supposed to be column vectors.

Table 2.2: Notation for track-related objects

I	list of modules crossed by the current track
k	size of I
$\boldsymbol{m}$	observations of the current track
$\boldsymbol{V}$	covariance matrix of $\boldsymbol{m}$
$\boldsymbol{x}_\mathrm{t}$	true track parameters of the current track
$\boldsymbol{x}_0$	expansion point of track model
$\boldsymbol{x}$	predicted track parameters of the current track
$\boldsymbol{C}$	covariance matrix of $\boldsymbol{x}$
$\widehat{\boldsymbol{x}}$	updated track parameters of the current track

The observations $\boldsymbol{m}$ depend on the track parameters $\boldsymbol{x}_\mathrm{t}$ via the track model f:

$$\boldsymbol{m} = \boldsymbol{f}(\boldsymbol{x}_\mathrm{t}) + \boldsymbol{\varepsilon}, \quad \mathsf{cov}(\boldsymbol{\varepsilon}) = \boldsymbol{V} \ .$$

The stochastic vector $\boldsymbol{\varepsilon}$ contains the effects of the observation error and of multiple scattering. Its variance–covariance matrix $\boldsymbol{V}$ can be assumed to be known. Energy loss is considered as deterministic and is included in the track model. A preliminary track fit gives a provisional estimate $\boldsymbol{x}$ of the track parameters [4]. The current estimates of the alignment parameters are used at this stage.

The observations also depend on the alignment parameters $\boldsymbol{d}_\mathrm{t}$, and the track model is extended accord-

ingly:

$$\boldsymbol{m} = \boldsymbol{f}(\boldsymbol{x}_\mathrm{t}, \boldsymbol{d}_\mathrm{t}) + \boldsymbol{\varepsilon}, \quad \mathsf{cov}(\boldsymbol{\varepsilon}) = \boldsymbol{V} \ .$$

The track model is linearized by a first-order Taylor expansion at expansion points $\boldsymbol{d}_0$ and $\boldsymbol{x}_0$:

$$\boldsymbol{m} = \boldsymbol{c} + \boldsymbol{A}\boldsymbol{d}_\mathrm{t} + \boldsymbol{B}\boldsymbol{x}_\mathrm{t} + \boldsymbol{\varepsilon} = \boldsymbol{c} + \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \end{pmatrix} \begin{pmatrix} \boldsymbol{d}_\mathrm{t} \\ \boldsymbol{x}_\mathrm{t} \end{pmatrix} + \boldsymbol{\varepsilon} \ .$$

The corresponding Jacobians are given by

$$\boldsymbol{A} = \partial \boldsymbol{m} / \partial \boldsymbol{d}_\mathrm{t} \big|_{\boldsymbol{d}_0}, \quad \boldsymbol{B} = \partial \boldsymbol{m} / \partial \boldsymbol{x}_\mathrm{t} \big|_{\boldsymbol{x}_0}$$

with the constant

$$\boldsymbol{c} = f(\boldsymbol{x}_0, \boldsymbol{d}_0) - \boldsymbol{A}\boldsymbol{d}_0 - \boldsymbol{B}\boldsymbol{x}_0 \ .$$

The expansion point $\boldsymbol{d}_0$ is either the nominal or the currently estimated module alignment, while the expansion point $\boldsymbol{x}_0$ is the result of a preliminary track fit. In addition, the Kalman filter requires a prediction $\boldsymbol{x}$ of the track parameters, along with its variance–covariance matrix $\boldsymbol{C}$. This prediction has to be stochastically independent of the observations in the track.

We first consider the case where an independent prediction of the track parameters exists. This can be an external prediction from an already aligned detector, or external information from a vertex or from kinematical constraints. In this case, the update equation of the Kalman filter reads

$$\begin{pmatrix} \widehat{\boldsymbol{d}} \\ \widehat{\boldsymbol{x}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{d} \\ \boldsymbol{x} \end{pmatrix} + \boldsymbol{K} \left(\boldsymbol{m} - \boldsymbol{c} - \boldsymbol{A}\boldsymbol{d} - \boldsymbol{B}\boldsymbol{x} \right)$$

with the following gain matrix:

$$\begin{aligned} \boldsymbol{K} &= \begin{pmatrix} \boldsymbol{D} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{C} \end{pmatrix} \begin{pmatrix} \boldsymbol{A}^T \\ \boldsymbol{B}^T \end{pmatrix} \underbrace{\left(\boldsymbol{V} + \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^T + \boldsymbol{B}\boldsymbol{C}\boldsymbol{B}^T \right)^{-1}}_{\boldsymbol{G}} \\ &= \begin{pmatrix} \boldsymbol{D}\boldsymbol{A}^T\boldsymbol{G} \\ \boldsymbol{C}\boldsymbol{B}^T\boldsymbol{G} \end{pmatrix} \ . \end{aligned}$$

In the second case, no independent prediction exists. The prediction $\boldsymbol{x}_0$ gets zero weight in order not to bias the estimation. This is accomplished by multiplying $\boldsymbol{C}$ by a scale factor α and letting α tend to infinity [5]:

$$\begin{aligned} \boldsymbol{G} &= \lim_{\alpha \longrightarrow \infty} \left(\boldsymbol{V} + \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^T + \alpha \boldsymbol{B}\boldsymbol{C}\boldsymbol{B}^T \right)^{-1} \\ &= \boldsymbol{V}_D^{-1} - \boldsymbol{V}_D^{-1}\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{V}_D^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{V}_D^{-1} \end{aligned}$$

with

$$\boldsymbol{V}_D = \boldsymbol{V} + \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^T \ .$$

Because of $\boldsymbol{G}\boldsymbol{B} = \boldsymbol{0}$ the update equation of the alignment parameters can be simplified to

$$\widehat{\boldsymbol{d}} = \boldsymbol{d} + \boldsymbol{D}\boldsymbol{A}^T\boldsymbol{G} \left(\boldsymbol{m} - \boldsymbol{c} - \boldsymbol{A}\boldsymbol{d} \right) \ .$$

The update of the covariance matrix can be calculated by linear error propagation:

$$\begin{aligned} \widehat{\boldsymbol{D}} = &\left(\boldsymbol{I} - \boldsymbol{D}\boldsymbol{A}^T\boldsymbol{G}\boldsymbol{A} \right) \boldsymbol{D} \left(\boldsymbol{I} - \boldsymbol{A}^T\boldsymbol{G}\boldsymbol{A}\boldsymbol{D} \right) \\ &+ \boldsymbol{D}\boldsymbol{A}^T\boldsymbol{G}\boldsymbol{V}\boldsymbol{G}\boldsymbol{A}\boldsymbol{D} \ . \end{aligned}$$

Both terms on the right-hand side are positive definite, so the left-hand side is guaranteed to be positive definite as well.

The iterative update of the alignment parameters needs some starting values. Mechanical and laser alignment can be used to obtain suitable starting values. Reference modules can be fixed by giving them very small initial errors.

2.3 Implementation and computational complexity

Let us assume that the current track crosses k detector modules. Their indices are denoted by the set $I = \{i_1, \ldots, i_k\}$. The dimension $n = 2\,k$ of the observation vector $\boldsymbol{m}$ is small, in the order of 30 for the CMS Inner Tracker. The matrix $\boldsymbol{B}$ is of size $n \times 5$ and is therefore small. The matrix $\boldsymbol{A}$ is a row of N blocks $\boldsymbol{A}_i$ of size $n \times m$, where m is the number of alignment parameters per detector module (usually 6). However, only k out of these N blocks are different from zero:

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{0} \ldots \boldsymbol{0} \ \boldsymbol{A}_{i_1} \ \boldsymbol{0} \ldots \boldsymbol{0} \ \boldsymbol{A}_{i_2} \ \boldsymbol{0} \ldots \boldsymbol{0} \ \boldsymbol{A}_{i_k} \ \boldsymbol{0} \ldots \boldsymbol{0} \end{pmatrix}$$

The only large matrix in the update formulas is the product $\boldsymbol{D}\boldsymbol{A}^T$. It is a column of N blocks each of which has size $m \times n$. Complete computation of $\boldsymbol{D}\boldsymbol{A}^T$ would lead to an algorithm that scales with N^2. This is too slow for practical purposes. There are two alternatives:

- Algorithm A: Compute only the blocks of the modules in the current track, neglecting all correlations.
- Algorithm B: Compute the blocks of the modules having significant correlations with the modules in the current track.

Algorithm A gives an unbiased estimate, but is suboptimal because of the missing correlations. Algorithm B is nearly optimal, but it has to be guaranteed that $\widehat{\boldsymbol{D}}$ is positive definite all the time. This problem is being studied, but there is not yet a foolproof solution. In any case, there has to be a tradeoff between speed and precision.

In order to keep track of the necessary updates, a list L_i is attached to each detector module i, containing the detector modules that have significant correlations with i. This list may contain only i itself in the beginning and grows as more tracks are processed. If there is prior knowledge about correlations, for instance because of mechanical constraints, it can be incorporated in the list and in the initial covariance matrix. The length

of the list can be restricted to a fairly small number, as the correlations between detector modules that are far from each other tend to be small. This leads to the following procedure for computing the updated alignment parameters:

1. Update the list L_i for every $i \in I$ (see below).
2. Form the list L of all detector modules that are correlated with the ones crossed by the current track: $L = \bigcup_{i\in I} L_i$. The size of L should be much smaller than N.
3. For all $j \in L$ compute: $(\boldsymbol{DA}^T)_j = \sum_{i\in I} \boldsymbol{D}_{ji}\boldsymbol{A}_i^T$. Each block $\boldsymbol{D}_{ji}$ is of size $m \times m$.
4. Compute: $\boldsymbol{ADA}^T = \sum_{i\in I} \boldsymbol{A}_i(\boldsymbol{DA}^T)_i$.
5. Compute: $\boldsymbol{V}_D = \boldsymbol{V} + \boldsymbol{ADA}^T$ and $\boldsymbol{G}$. All matrices involved are of size $n \times n$.
6. Compute: $\boldsymbol{m'} = \boldsymbol{G}\left(\boldsymbol{m} - \boldsymbol{c} - \sum_{i\in I} \boldsymbol{A}_i\boldsymbol{d}_i\right)$.
7. For all $j \in L$ compute: $\widehat{\boldsymbol{d}}_j = \boldsymbol{d}_j + (\boldsymbol{DA}^T)_j\boldsymbol{m'}$.

In the beginning, the covariance matrix $\boldsymbol{D}$ is block-diagonal and contains the prior uncertainty of the alignment parameters, derived from laser alignment and mechanical measurements. If required, it may also contain prior correlations between different detector modules. After each track, only the blocks in the list $L = \bigcup_{i\in I} L_i$ need to be updated. This is done in the following way:

For all $j, l \in L$ compute:
$\widehat{\boldsymbol{D}}_{jl} = \boldsymbol{D}_{jl} + (\boldsymbol{DA}^T)_j(\boldsymbol{GV}_D\boldsymbol{G} - 2\boldsymbol{G})[(\boldsymbol{DA}^T)_l]^T$.

The computational complexity of the parameter update is of the order $|L| \cdot |I|$, and the computational complexity of the update of the covariance matrix is of the order $|L|^2$. Restricting the size of the lists L_i is therefore of crucial importance.

The current proposal for building the lists L_i is based on the concept of a distance between two modules i and j. Let us define the following relation:

$i \sim j \Longleftrightarrow i$ and j have been crossed by the same track.

Then the relation "$\sim$" is symmetric, but not transitive. Based on this relation, we define the distance $d(i, j)$ between detector modules i and j by:

1. $d(i, i) = 0$
2. If $i \neq j$ and $i \sim i_1 \sim i_2 \sim \cdots \sim i_n \sim j$ is the shortest chain connecting i to j, the distance is $d(i, j) = n + 1$.

It is clear from the definition that $d(i, j) = 1$ if and only if $i \sim j$. Also, it is easy to show that d has the formal properties required by a metric:

1. $d(i, j) = 0$ if and only if $i = j$
2. $d(i, j) = d(j, i)$
3. $d(i, j) \leq d(i, k) + d(k, j)$ for all k.

There are several possibilities for building the list L_i:

- The list L_i contains all modules k with $d(k, i) \leq d_{\max}$.
- The list L_i grows until the correlations have stabilized. Then all correlations below an upper limit are dropped. This approach is dynamic and adapts to the track sample used.
- The 'optimal' correlation structure is determined from simulated data. This approach is static and has to be done separately for every potential track sample (cosmics, beam halo, interactions).

Clearly, more detailed studies are required in order to select the best approach.

2.4 The two-track fitter

With track pairs, vertex and mass constraints can be used to improve the momentum resolution. Typical examples are the decays $Z \longrightarrow \mu^+\mu^-$ and $J/\psi \longrightarrow \mu^+\mu^-$. Track pairs can be treated in the same manner as single tracks if the five track parameters are replaced by nine decay parameters:

- the position $\boldsymbol{v}$ of the decay vertex (3 parameters),
- the momentum $\boldsymbol{p}$ of the mother particle in the lab frame (3 parameters),
- the two decay angles (θ, φ) in the rest frame of the mother particle (2 parameters),
- and the mass m of the mother particle (1 parameter).

The mass is constrained by adding a virtual observation of the mass (theoretical value plus width). The Jacobians of the observation vectors $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ of the two tracks w.r.t. the decay parameters are obtained by the chain rule:

$$\frac{\partial \boldsymbol{m}_i}{\partial(\boldsymbol{v}, \boldsymbol{p}, \theta, \varphi, m)} = \frac{\partial \boldsymbol{m}_i}{\partial \boldsymbol{x}_i} \cdot \frac{\partial \boldsymbol{x}_i}{\partial(\boldsymbol{v}, \boldsymbol{p}_i)} \cdot \frac{\partial(\boldsymbol{v}, \boldsymbol{p}_i)}{\partial(\boldsymbol{p}, \theta, \varphi, m)} \ .$$

Otherwise the formalism remains unchanged.

2.5 Results

The algorithm has been implemented in CMSSW [6]. We have tested its performance in two set-ups, a smaller one with about 500 alignable modules, and a larger one with more than 5000 modules.

2.5.1 Small set-up

The small set-up is a wheel-shaped subset of the barrel part of the tracker [7]. It consists of 3 pixel layers (192 modules), 4 TIB layers (162 modules), and 6 TOB layers (344 modules). The modules of TIB and TOB were aligned relative to the Pixel barrel, yielding a total

of 506 alignable modules. The misalignment was simulated according to

$$\sigma(\Delta u) = 100\,\mu\text{m}, \sigma(\Delta v) = 100\,\mu\text{m}, \sigma(\Delta\gamma) = 5\,\text{mrad}$$

u and v are the local coordinates, respectively orthogonal and parallel to the strips, and γ is the angle of rotation around the local z axis. Alignment w.r.t. v is of course restricted to the double-sided modules. 10 000 muon pairs (20 000 tracks) from the decay $Z \longrightarrow \mu^+\mu^-$ were used for the alignment. The momentum of the Z was distributed uniformly between 50 and 80 GeV/c.

The performance was investigated in four cases:

1. No correlations,
2. correlations up to $d_{\max} = 5$,
3. all correlations,
4. all correlations with Z-mass constraint.

Figure 2.1 shows the final residuals of alignment in u, i.e., the estimated shifts minus the true shifts, for the four cases just mentioned. The residuals are centred around zero. The RMS width of the distribution depends on the number of correlations used, in just the way one would expect. There is a significant improvement from case 1 (no correlations) to case 2 (correlations up to $d_{\max} = 5$). Using all correlations yields very little improvement, but the Z-mass constraint shrinks the r.m.s. width of the distribution by almost 10%. The corresponding results for the rotation angle γ look very similar (Fig. 2.2). Using correlations up to $d_{\max} = 5$ shrinks the width by a factor of 2. Using all correlations is only slightly better, and the Z-mass constraint again gives an improvement of about 10%.

The performance also depends on how many alignment parameters are estimated concurrently. Figure 2.3 shows the evolution of residuals of the u-shifts as a function of the number of updates. In the left-hand panel only one parameter, namely u, has been misaligned and estimated, whereas in the right-hand panel all three parameters (u, v and γ) have been misaligned and estimated. The final r.m.s. width is 11.0 μm in the 1-parameter case, and 20.4 μm in the 3-parameter case. In both cases all correlations have been used.

The improvement from using the correlations has to be paid for by a substantial increase in computing time. Table 2.3 shows the CPU time spent in various parts of the algorithm. At first glance it may seem surprising that using track pairs with a mass constraint is faster than using single tracks. The explanation is the following. First, the alignment interface is accessed only once per track pair, and as a consequence the time spent there is reduced by about a factor of two. Second, the update is also called only once per track pair. Although more modules have to be updated per track pair, there remains a net gain of about seven per cent.

Table 2.3: Timing for the alignment of the small set-up. Shown are the times for the algorithmic update (T_{alg}), the retrieval of data from the alignment interface (T_{int}) and the update of the correlation lists (T_{lis}). All times are CPU times and are given in seconds.

Three estimated parameters			
	T_{alg}	T_{int}	T_{lis}
No correlations	65	14	–
Correlations up to $d_{\max} = 5$	1591	1228	73
All correlations	5260	3316	–
All correlations with mass constraint	4827	1694	–

One estimated parameter			
	T_{alg}	T_{int}	T_{lis}
All correlations	603	1870	–

2.5.2 Large set-up

The larger set-up comprises about a third of the entire tracker. It consists of 3 pixel layers (505 modules), 4 TIB layers (1739 modules), and 6 TOB layers (3961 modules). The modules of TIB and TOB were aligned relative to the Pixel barrel, yielding a total of 5700 alignable modules. The misalignment was simulated according to

$$\sigma(\Delta u, \Delta v) = 100\,\mu\text{m}, \sigma(\Delta\gamma) = 5\,\text{mrad}\ .$$

2000 muon pairs (4000 tracks) from the decay $Z \longrightarrow \mu^+\mu^-$ were used for the alignment. The momentum of the Z was distributed uniformly between 50 and 80 GeV/c.

The alignment was studied including correlations up to $d_{\max} = 3$. Figure 2.4 shows the final residuals of alignment in u, i.e., the estimated shifts minus the true shifts, separately for TIB and TOB. Note that modules with less than 4 hits have been excluded, leaving 1530 modules in the TIB and 3250 modules in the TOB. In the TIB, the r.m.s. width is down to about 20 μm with no more than 4000 tracks, while in the TOB it is about 40 μm. Figure 2.5 shows the corresponding residuals for the rotation angle γ. The r.m.s. width is about 0.9 mrad in the TIB, and about 1 mrad in the TOB. The CPU time required for the 4000 tracks is shown in Table 2.4.

Table 2.4: Timing for the alignment of the large set-up. Shown are the times for the algorithmic update (T_{alg}), the retrieval of data from the alignment interface (T_{int}) and the update of the correlation lists (T_{lis}). All times are CPU times and are given in seconds.

Three estimated parameters			
	T_{alg}	T_{int}	T_{lis}
Correlations up to $d_{\max} = 3$	4860	5196	97

2.6 Summary and outlook

A Kalman filter for sequential estimation of alignment constants has been developed and successfully tested on a small scale and on a large set-up. Its advantages are

- No large system of equations has to be solved.
- The depth of the correlations that are taken into account can be tailored to the problem.
- The errors of the estimated alignment constants are available and can be used as a stopping criterion.

The disadvantages are

- The computational expense per track is large, especially if all correlations are used.
- More bookkeeping is required.

Clearly further work is required to make the method operational on the full tracker:

- The various approaches to the correlation lists have to be studied.
- The method has to be tested on full scale and further optimized for speed.
- The estimation has to be extended to the full set of angles and shifts.

Acknowledgements

We wish to thank Wolfgang Adam (HEPHY Vienna) for technical support and CERN for financial support to one of the authors (E. Widl).

References

[1] D.E. Catlin, *Estimation, Control, and the Discrete Kalman Filter* (Springer, New York, 1989).

[2] R. Frühwirth, *Nucl. Instrum. Methods* **A262** (1984) 444.

[3] R. Frühwirth, T. Todorov and M. Winkler, *J. Phys. G: Nucl. Part. Phys.* **29** (2003) 561.

[4] W. Adam *et al.*, CMS Note 2006/041 (2006).

[5] R. Zurmühl and S. Falk, *Matrizen und ihre Anwendungen*, 5th edition (Springer, Berlin–Heidelberg, 1984).

[6] `https://twiki.cern.ch/twiki/bin/view/CMS/CMSSW`

[7] CMS Collaboration, CERN/LHCC 98-006 (1998); CERN/LHCC 2000-016 (2000).

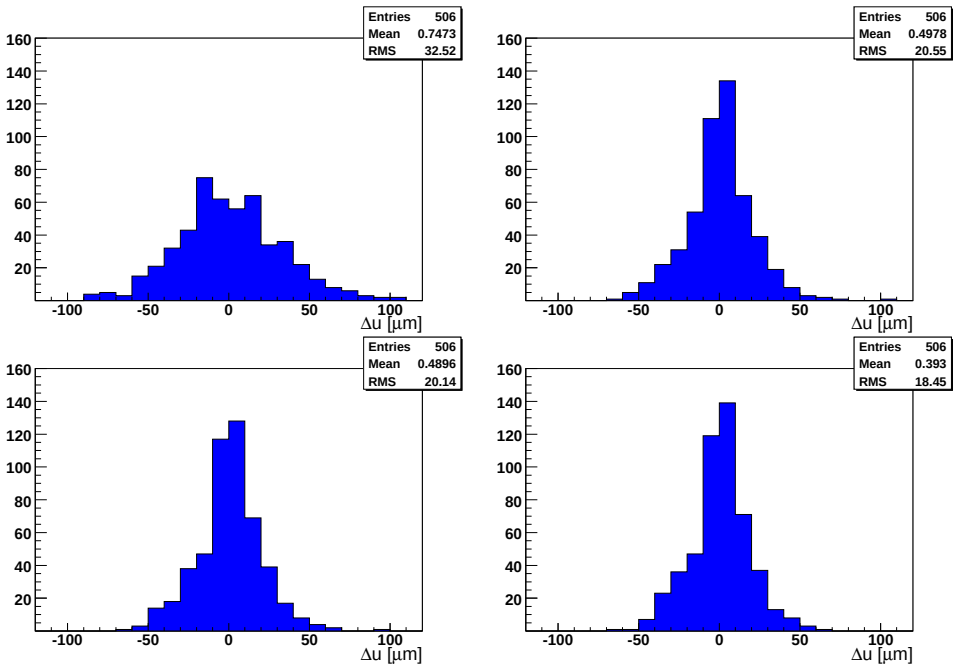


Fig. 2.1: Residuals of the shift in u after alignment (small set-up). Top left: no correlations. Top right: correlations up to $d_{\max} = 5$. Bottom left: all correlations. Bottom right: all correlations with Z-mass constraint.

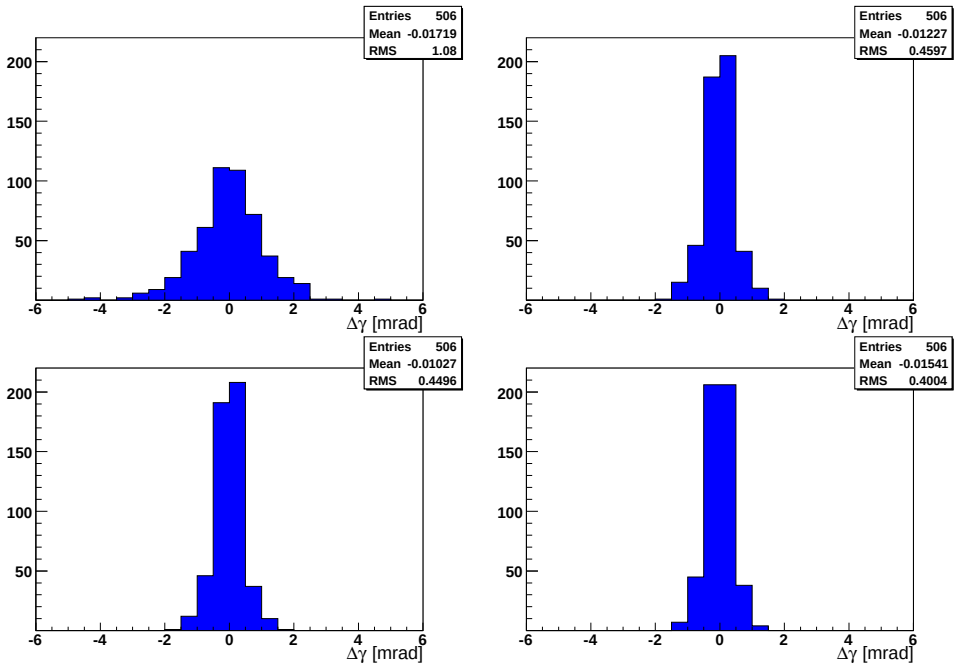


Fig. 2.2: Residuals of the rotation angle γ after alignment (small set-up). Top left: no correlations. Top right: correlations up to $d_{\max} = 5$. Bottom left: all correlations. Bottom right: all correlations with Z-mass constraint.

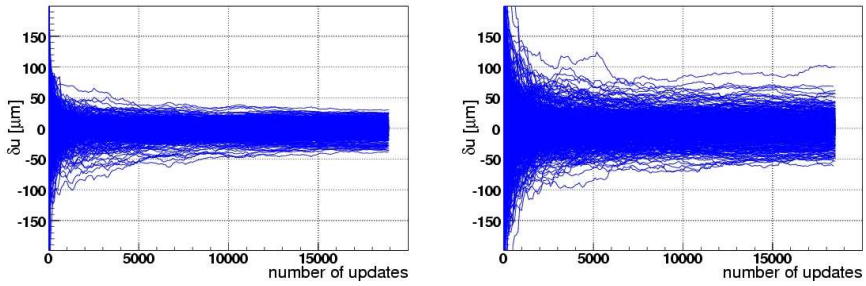


Fig. 2.3: Evolution of the residuals of the shift in u versus the number of updates (small set-up). Left: one parameter estimated. Right: three parameters estimated.

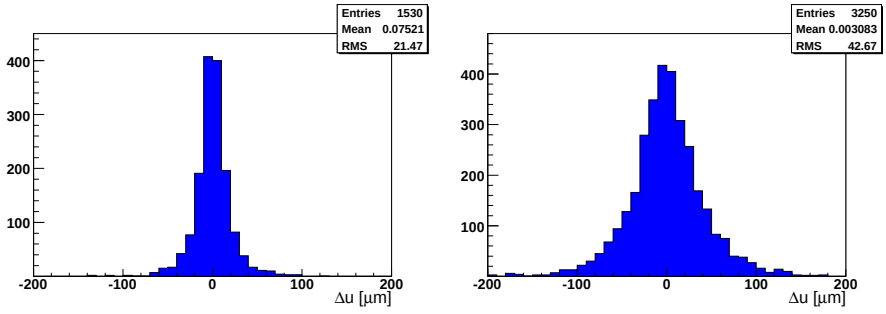
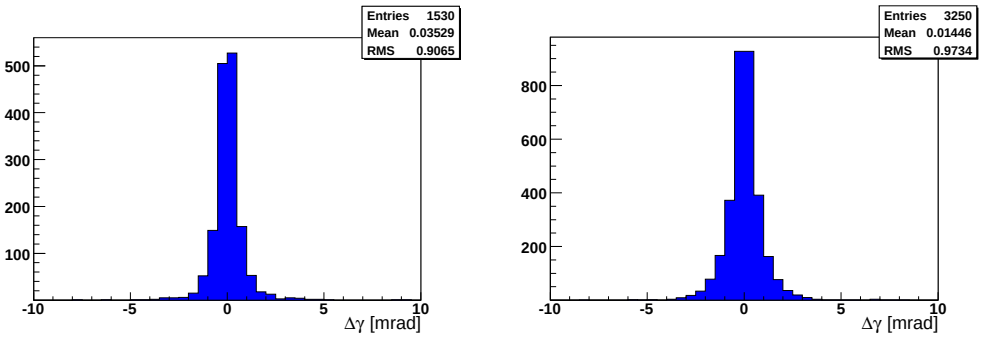


Fig. 2.4: Residuals of the shift in u after alignment (large set-up). Left: inner barrel (TIB) Right: outer barrel (TOB).



Inner barrel (TIB) Outer barrel (TOB)

Fig. 2.5: Residuals of the rotation angle γ after alignment (large set-up)