Int. J. Comp. Vision, Submitted 3-22-00, Revised 11-13-00, now in press, 1 (2001)

Implementing a multi-model estimation method.

T. VIEVILLE, D. LINGRAND AND F. GASPARD
INRIA, Sophia, BP93, 06902 Valbonne, France.
tel : +33 4 92 38 76 88 fax : +33 4 92 38 78 45
e-mail : Thierry.Vieville@inria.fr

;

Abstract. We revisit the problem of parameter estimation in computer vision, reconsidering and implementing what may be called the Kanatani's estimation method, presented here as a simple optimisation problem, so (a)without any direct reference to a probabilistic framework but (b)considering (i) non-linear implicit measurement equations and parameter constraints, plus (ii) robust estimation in the presence of outliers and (iii) multi-model comparisons.

Here, (A)a projection algorithm based on generalisations of square-root decompositions allows an efficient and numerically stable local resolution of a set of non-linear equations. On the other hand, (B)a robust estimation module of a hierarchy of non-linear models has been designed and validated.

A step ahead, (C)the software architecture of the estimation module is discussed with the goal of being integrated in reactive software environments or within applications with time constraints, while an experimentation considering the parameterisation of retinal displacements between two views is proposed as an illustration of the estimation module.

Keywords: Non-Linear Estimation, Robust Estimation, Multi-Model

.1. Introduction

Estimation of parameters in computer-vision is a recurrent and somehow "never-solved" problem (a didactic introduction about this topic may be found in [48]), since many different aspects are to be taken into account such as (i) nonlinear equations and constraints, (ii) approximate measures and outliers elimination, (iii) singularities in the equations, with the requirement to use different models as alternatives. Let us illustrate these aspects by an example.

The two-views "motion" problem as a typical example.

In this paper, we consider as a typical example the two-views "motion" estimation problem: given two views of a 3D scene we have to recover the *physical parameters* (calibration, Euclidean displacement), say $\mathbf{q}$, defining the disparity between 2D data points in the images.

Let us briefly review the problem. Refer, for instance, to [49] for more details. We consider two images of a rigid object, with singular points (in fact corners) detected on this object and matched in the two views. A bilinear constraint, which characterises the retinal displacement, exists between the homogeneous coordinates of these pairs of points $(\mathbf{p}_i, \mathbf{p}'_i)$.

This is written : $\mathbf{p}_i'^T \mathbf{F}(\mathbf{q}) \mathbf{p}_i = 0$ and it constitutes the *measurement equation* provided by the match $(\mathbf{p}_i, \mathbf{p}'_i)$. It is used to evaluate the parameter $\mathbf{q}$. Here the "fundamental matrix" $\mathbf{F}(\mathbf{q})$ is defined up to a scale factor and subject to the algebraic cubic constraint $det(\mathbf{F}(\mathbf{q})) = 0$.

Hence, the components of the fundamental matrix are "homogeneous" in the sense that they are linear with respect to the measurement equation and defined up to a scale factor. Their estimation may thus be much simpler than estimating the physical parameters. From a theoretical point of view, this corresponds to analysing the projective structure of the scene (see [17, 30]).

This parameterisation is undefined in the case of a "planar" displacement (i.e. all points are in the same plane or it is a pure rotation) whereas another equation holds : $\mathbf{p}'_i \wedge \mathbf{H}(\mathbf{q}) \mathbf{p}_i = 0$, given another matrix $\mathbf{H}(\mathbf{q})$, also defined up to scale factor, but not subject to constraint.

Here, indeed, points may belong to another rigid objects or may be incorrectly matched and thus act as outliers for this estimation. Robust estimate is thus mandatory.

Furthermore, the rigid displacement or the camera intrinsic parameters may be specific [43] thus yielding particular forms of $\mathbf{F}(\mathbf{q})$ or $\mathbf{H}(\mathbf{q})$. Several models must thus be evaluated concurrently.

Estimating "homogeneous" parameters.

Kanatani[1] may be the first computer-vision scientist who has really attacked the double problem of non-linear statistical estimation [20, 22, 23] and multi-model statistical inference [21, 23] using a pioneer work [1] developed in another domain.

More recently, Meer and his group [26, 25, 29] have developed a very powerful formalism for non-linear statistical estimation, providing that the parameter to estimate is homogeneous with respect to the measurement equation. This corresponds to the estimation of the $\mathbf{F}$ or $\mathbf{H}$ matrices components in our example.

Similarly, Brooks and his group, estimating an homogeneous parameter with measurement equations which are quadratic functions of the measurement variables [7] develop an effective method to obtain an unbiased estimate for the Kanatani estimation scheme.

More generally, several authors (e.g. [40, 16]) have developed methods to deal with this class of problem.

All these "re-normalisation" methods assume that rejection of outliers has been done elsewhere in a earlier module. This may be a caveat, since rejection of outliers requires a reasonable estimation of the parameter itself. As a consequence, both methods may have to be mixed, as we attempt to do here.

In particular, the presence of inliers of a different "object" (i.e. belonging to another set of measures coherent with a different parameter value) may breakdown the estimation, even for robust estimators (see [37] for a quantitative study). As studied by this author, the observed bias is intrinsically due to the fact that the used criteria are only based on the residual error. Although the present framework is not intended to solve this problem, we will discuss this aspect for the proposed method.

Using "physical" parameters.

It appears that, in our context [11, 46, 42, 41, 43], we are not able to re-use this formalism for the following reason : we *must* estimate not the homogeneous *but* the physical parameters and such a parameterisation of computer-vision parameters is NOT homogenous with respect to measurement equations.

This corresponds to perceptual tasks in which the Euclidean geometry of the camera and/or the scene has to be recovered (localisation, visual measurements, tracking involving robotic degrees of freedom) or is a part of the problem (calibration tasks, camera with constrained displacements, assumptions on specific configurations or displacements, etc..).

Estimating the physical parameters allows to introduce specific knowledge about the visual system [43] and leads to much accurate estimations, as in [10]. It is well-known, for calibration problems for instance (e.g. [47, 6]), that non-linear estimation is much more stable and precise, considering physical parameters.

At an applicative level, the precision of the data input and output is easier to specify by the end-user on physical parameters [12]. Furthermore, the estimation is directly optimised with respect to the desired parameters, which helps analysing the obtained results.

Using physical parameters instead of homogeneous ones has also two "technical" advantages:

(a) In this context, physical parameters induce a parameterisation of the homogeneous parameters. In the two-views motion example, the matrix $\mathbf{F}(\mathbf{q})$ must verify $det(\mathbf{F}(\mathbf{q})) = 0$ but this is always the case if we write it in function of $\mathbf{q}$. We thus may avoid some complexity here.

(b) Authors dealing with homogeneous parameters have demonstrated that the metric of the related criterion (for instance [26] using a Mahanalobis distance in which bias is corrected) is deeply dependent upon the data and the estimated parameters. This is because they have to estimate the characteristics of a non-linear transformation from the physical to the homogeneous parameters. On the reverse [22], if we keep using physical parameters and the raw measures as input to the estimation algorithm, the related metric remains constant.

What is the paper about

This paper thus describes a potential alternative as a comprehensive computational system for solving nonlinear parametric fitting problems that are frequently encountered in computer vision applications, here using "physical parameters":

- in the next section we will define the *estimation problem* as an *optimisation problem* [3, 14] trying to find a "minimal" formulation ..
- allowing to solve it as a *projection problem*. Properties of such a problem are well known [32]. This allows us to propose a rather efficient implementation,
- while it will be specialised to our estimation problem, including robust estimation and multi-model estimation in the subsequent sections.

This finally will allow us to describe an effective software implementation and experiment it to validate this approach.

.2. Estimating a parameter with non-linear constraints.

.2.1. Position of the problem

We consider the "simple" problem of estimating a static quantity $\mathbf{q}$ from a set of M measures.

More precisely[2], we want to estimate a n-dimensional real vectorial quantity, say a *parameter*, $\mathbf{q} \in \mathcal{R}^n$ given:

- a set of p implicit non-linear *constraints* written $\mathbf{c}_0(\mathbf{q}) = 0_p$, so that the parameter may belong to some specific space $\mathcal{C}$, defined by these equations,

- an *approximate initial estimate* $\mathbf{q}_0$,
  i.e. we consider that $\mathbf{q}$ is close to $\mathbf{q}_0$ for a given distance $||\mathbf{q} - \mathbf{q}_0||^2_{\mathbf{Q}_0}$,
- a set of M *approximate measures* $(\mathbf{m}_1, \cdots, \mathbf{m}_i, \cdots, \mathbf{m}_M)$ with $\mathbf{m}_i \in \mathcal{R}^{n_i}$, i.e.
  - we consider that the *true* measures $\bar{\mathbf{m}}_i$ are close to the observed measures $\mathbf{m}_i$ for a given distance $||\bar{\mathbf{m}}_i - \mathbf{m}_i||^2_{\mathbf{Q}_i}$
  - while, for each measure, a set of p_i *measurement equations* $\mathbf{c}_i(\mathbf{m}_i, \mathbf{q}) = 0_{p_i}$ defines the relation between measure and parameter,

  such approximate measures have to be *corrected* by the algorithm.

as schematised in Fig. 1.

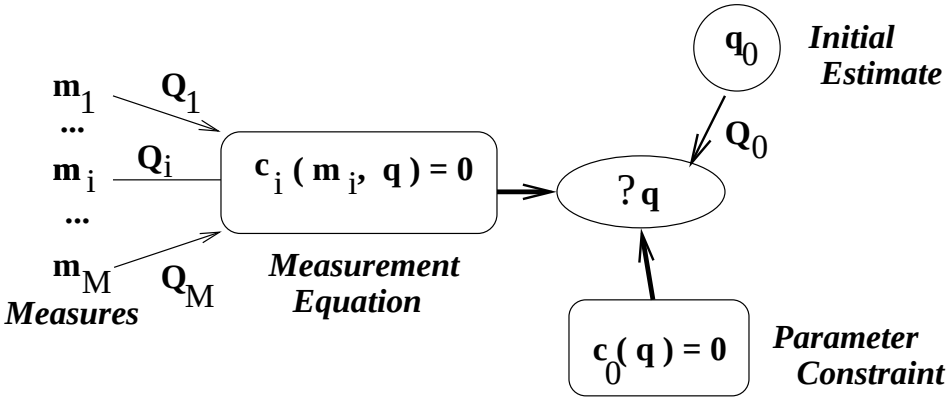


Fig. 1. Description of the estimation problem.

The notion of "approximate" data is formalised here, using *quadratic semi-distances*, , i.e. :

$$||\mathbf{x}-\mathbf{y}||^2_{\mathbf{Q}} = (\mathbf{x}-\mathbf{y})^T\,\mathbf{Q}\,(\mathbf{x}-\mathbf{y}) \tag{1}$$

for $\mathbf{x} \in \mathcal{R}^n$ and $\mathbf{y} \in \mathcal{R}^n$ where $\mathbf{Q}$ is a positive semi-definite symmetric matrix. The matrix $\mathbf{Q}$ may be called the *quadratic information matrix*. In a statistical framework (see appendix A.2 for a discussion) this corresponds to the inverse of a covariance and the distance corresponds to a Mahanalobis distance. However, we will not follow this track here, since we cannot guaranty that the assumptions required to develop such a formalism are verified in our case.

If no initial estimate is available, one can simply write $\mathbf{q}_0 = \mathbf{Q}_0 = 0$, this part of the specification is thus not a constraint.

.2.1.1. Defining estimation as a minimisation problem. Therefore, we can formalise the problem as an *optimisation* problem, i.e. estimate the parameter $\tilde{\mathbf{q}}$ *and* the measures $(\tilde{\mathbf{m}}_1, \cdots, \tilde{\mathbf{m}}_i, \cdots)^T$ which :
(i) minimise the *sum* of the defined distances :

$$\min_{(\tilde{\mathbf{q}},\tilde{\mathbf{m}}_1,\cdots,\tilde{\mathbf{m}}_i,\cdots)} \mathcal{L}^2 = \frac{1}{2}||\tilde{\mathbf{q}}-\mathbf{q}_0||^2_{\mathbf{Q}_0} + \sum_{i=1}^{M} \frac{1}{2}||\tilde{\mathbf{m}}_i-\mathbf{m}_i||^2_{\mathbf{Q}_i} \tag{2}$$

(ii) given the different equations :

$$\mathbf{c}_0(\tilde{\mathbf{q}}) = 0 \text{ and } \forall i \in \{1..M\}\ \mathbf{c}_i(\tilde{\mathbf{q}}, \tilde{\mathbf{m}}_i) = 0 \tag{3}$$

Equivalently, this estimation problem can be formalised as a composite criterion with Lagrangian multiplicators $\lambda = (\lambda_0, \cdots, \lambda_i, \cdots)^T$:

$$\min_{(\tilde{\mathbf{q}},\tilde{\mathbf{m}}_1,\cdots,\tilde{\mathbf{m}}_i,\cdots)} \max_{\lambda} \mathcal{L}^2_\lambda = \frac{1}{2}||\tilde{\mathbf{q}}-\mathbf{q}_0||^2_{\mathbf{Q}_0} + \lambda_0^T\,\mathbf{c}_0(\tilde{\mathbf{q}}) + \sum_{i=1}^{M}\left[\frac{1}{2}||\tilde{\mathbf{m}}_i-\mathbf{m}_i||^2_{\mathbf{Q}_i} + \lambda_i^T\,\mathbf{c}_i(\tilde{\mathbf{q}},\tilde{\mathbf{m}}_i)\right] \tag{4}$$

At the optimum, we can always write, for a given matrix $\tilde{\mathbf{Q}}$:

$$\mathcal{L}^2 = \mathcal{L}^2(\tilde{\mathbf{q}}) + \frac{1}{2}||\mathbf{q}-\tilde{\mathbf{q}}||^2_{\tilde{\mathbf{Q}}} + o(||\mathbf{q}-\tilde{\mathbf{q}}||^2) \tag{5}$$

.2.1.2. Quantifying the precision of the estimate. In this last equation, we not only define the parameter estimate $\tilde{\mathbf{q}}$ but also a *quadratic distance to the parameter estimate* parameterised by $\tilde{\mathbf{Q}}$. This allows to evaluate the precision of the estimate, again as a quadratic distance.

It is straight-forward, although rather painful, to derive:

$$\tilde{\mathbf{Q}} = \mathbf{Q}_0 + \sum_{i=1}^{M} \left.\frac{\partial \mathbf{c}_i(\mathbf{q},\mathbf{m}_i)}{\partial \mathbf{q}}\right|_{(\tilde{\mathbf{q}},\tilde{\mathbf{m}}_i)}^{T} \left[\left.\frac{\partial \mathbf{c}_i(\mathbf{q},\mathbf{m}_i)}{\partial \mathbf{m}_i}\right|_{(\tilde{\mathbf{q}},\tilde{\mathbf{m}}_i)} \mathbf{Q}_i^{+} \left.\frac{\partial \mathbf{c}_i(\mathbf{q},\mathbf{m}_i)}{\partial \mathbf{m}_i}\right|_{(\tilde{\mathbf{q}},\tilde{\mathbf{m}}_i)}^{T} \right]^{-} \left.\frac{\partial \mathbf{c}_i(\mathbf{q},\mathbf{m}_i)}{\partial \mathbf{q}}\right|_{(\tilde{\mathbf{q}},\tilde{\mathbf{m}}_i)} \tag{6}$$

where the notations $\mathbf{M}^{+}$ and $\mathbf{M}^{-}$ denotes pseudo-inverses and will be defined in the next section, this formula being derived in appendix A.1.

.2.1.3. Normalising the estimated criterion. Considering the estimation of $\mathbf{q}$, this problem has n unknowns and p equations, irrespectively of the measures. This means that among the n "variables", p of them are "fixed" by the equations. The remainder $n-p$ variables are "free" and may vary to maintain the unknowns close to the default value $\mathbf{q}_0$. We thus may call $n-p$ the number of *degrees of freedom.*

In addition to this, for each measure estimate $\tilde{\mathbf{m}}_i$, p_i equations constraint it to differ from its approximate value $\mathbf{m}_i$, so that each measurement *bias* $\upsilon_i = \tilde{\mathbf{m}}_i - \mathbf{m}_i$ is governed by a p_i dimensional quantity, i.e. has p_i degrees of freedom.

As a consequence, a natural "normalised" value of the criterion is :

$$\frac{\tilde{\mathcal{L}}^2}{d} \quad \text{with} \quad d = n - p + \sum_i p_i \tag{7}$$

In other words, the error criterion is divided by the total number of degree of freedom.

.2.2. Solving as a local projection problem.

.2.2.1. This estimation is a projection problem. Since we have to minimise this criterion with respect to both parameter and measure estimations, the previous problem can thus be rewritten, using the previous notations, in a more compact form :

$$\min_{\mathbf{x}} \max_{\lambda} \mathcal{L}_\lambda^2 = \frac{1}{2}||\mathbf{x}-\mathbf{x}_0||_{\mathbf{Q}}^2 + \lambda^T \mathbf{c}(\mathbf{x}) \tag{8}$$

with $\mathbf{x}_0 = (\mathbf{q}_0, \mathbf{m}_1, \cdots, \mathbf{m}_i, \cdots, \mathbf{m}_M)$ and $\mathbf{x} = (\tilde{\mathbf{q}}, \tilde{\mathbf{m}}_1, \cdots, \tilde{\mathbf{m}}_i, \cdots, \tilde{\mathbf{m}}_M)$
so that $\mathbf{c}(\mathbf{x}) = (\mathbf{c}_0(\mathbf{q}), \cdots, \mathbf{c}_i(\mathbf{q},\mathbf{m}_i), \cdots)$ assuming, for technical reasons, that these equations are twice differentiable,
while $\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_0 & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{Q}_1 & \cdots \\ \cdots & \cdots & \cdots \end{pmatrix}$ is a block diagonal matrix.

As such, the problem is a simple *projection* problem, i.e. the criterion given in (8) means finding the quantity $\mathbf{x}$
(i) closest to $\mathbf{x}_0$ for the quadratic distance parameterised by $\mathbf{Q}$ and
(ii) in the set $\mathcal{C}$ defined by $\mathbf{c}(\mathbf{x}) = 0$, as schematised in Fig. 2.

It is well known (e.g. [24, 13]) that :
$\mathcal{P}_1$: *This problem has an unique solution if (a) $\mathcal{C}$ is a convex or linear set, else (b) it has a local solution if $\mathcal{C}$ is, in some sense, regular, for instance if the function $\mathbf{x} \to \mathbf{c}(\mathbf{x})$ is twice differentiable with bounded second-order derivatives in a neighbourhood of $\mathbf{x}_0$ containing its projection.*

.2.2.2. Resolution up to the first order. In a more constructive way, i.e. in order to obtain an effective algorithm, we consider the *linear approximation* of the non-linear equations around a point $\mathbf{x}^\bullet$, which may be written :

$$0 = \mathbf{C}\,\mathbf{x} - \mathbf{d} + o(\kappa||\mathbf{x}-\mathbf{x}^\bullet||) \text{ with } \mathbf{C} = \left.\frac{\partial \mathbf{c}(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}^\bullet} \text{ and } \mathbf{d} = \mathbf{C}\,\mathbf{x}^\bullet - \mathbf{c}(\mathbf{x}^\bullet) \tag{9}$$

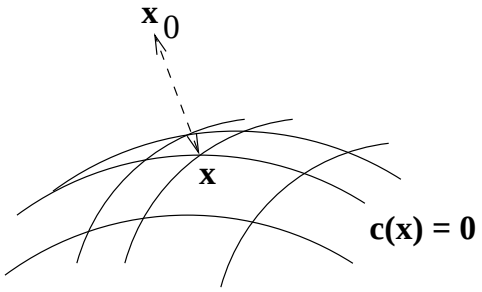


Fig. 2. Non-linear estimation as a projection problem.

where κ is the magnitude of the second-order derivative of $\mathbf{c}(\mathbf{x})$, combined with the normal equation of the criterion at $\mathbf{x}^\bullet$:

$$0 = \frac{\partial \mathcal{L}_\lambda^2}{\partial \mathbf{x}}^T = \mathbf{Q}\,(\mathbf{x} - \mathbf{x}_0) + \mathbf{C}^T\,\lambda \tag{10}$$

which allows to compute $\mathbf{x}$ as the iterative solution of the approximate linear system :

$$\begin{pmatrix} \mathbf{Q} & \mathbf{C}^T \\ \mathbf{C} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \lambda \end{pmatrix} \simeq \begin{pmatrix} \mathbf{Q}\,\mathbf{x}_0 \\ \mathbf{d} \end{pmatrix} \tag{11}$$

as proposed in the literature (e.g. [32] or more recently [45]). Here, we not only revisit this method, but introduce a few improvements:
(a) managing quadratic semi-definite information matrices, i.e. partially defined quantities,
(b) dealing with redundant or singular sets of equations,
(c) allowing, in any cases, the convergence of the method, not necessarily to the optimal value, but at least to a realistic sub-optimal estimate.

However, to attain this double goal, we must also revisit a very standard numerical algorithm.

.2.2.3. Square-root decomposition of positive semi-definite symmetric matrices. The Cholesky or "square-root" decomposition of a symmetric positive definite matrix $\mathbf{S}$ (e.g. [36]) is a lower triangular matrix $\mathbf{L}$ such that $\mathbf{S} = \mathbf{L}\,\mathbf{L}^T$.

In fact, among all algorithms available in linear equations system resolution, this very simple algorithm is the *fastest* (fix number of operations, no pivoting mechanism required for instance) and the more stable, from a numerical point of view (see for instance [33]). This is because it fully makes use of the fact the matrix is symmetric and positive, at it is the case here.

It is *much faster* than a *singular value decomposition* (e.g. [36]) which is of common use in such a context, since the former has a fixed small polynomial complexity, while the latter requires more operation and must be iterated a few times until convergence.

Willing to use this fast method we have defined two *generalisations* of the standard square-root decomposition, if the matrix is not definite:

The "closest" square-root decomposition This is the square-root decomposition of a matrix $\mathbf{S}^> = \mathbf{S} + v\,\mathbf{e}_k\,\mathbf{e}_k^T$ for some small minimal v (here $\mathbf{e}_k$ is the k-th basic vector).
This is simply implemented by enforcing diagonal terms of the square-root matrix to be equal to a small positive quantity (we use 10 times the machine precision) if not yet strictly positive.
When inverting such a matrix though its square-root, the inverse of these small values are large but not huge, thus still manageable quantities. We write $\mathbf{S}^+$ the inverse of $\mathbf{S}^>$, it is a "pseudo-inverse" of $\mathbf{S}$.

It is thus guaranty to compute the square-root decomposition of a positive definite matrix $\mathbf{S}^{>}$ "close" to $\mathbf{S}$, since $||\mathbf{S}^{>} - \mathbf{S}|| = o(\upsilon)$. If the matrix is positive but not definite this distance is infinitesimal, in practice of the order of magnitude of the machine precision.
In the extreme case where $\mathbf{Q} = 0$ the closest positive matrix is $\upsilon\mathbf{I}$, where υ has the order of magnitude of the machine precision.

The "reduced" square-root decomposition. This is the square-root decomposition of the sub-matrix of $\mathbf{S}$ from which *rows and columns whose diagonal elements vanish are removed.*
This is simply implemented by deleting these elements if a diagonal term of the square-root matrix is lower than an ϵ (ϵ being, here, 10 times the machine precision) and resume the calculation with the corresponding sub-matrix.
It is also guaranteed to compute the square-root decomposition of a positive definite matrix $\mathbf{S}^{<}$ but in a (eventually empty !) sub-space generated by a sub-set of the basic vectors.
With this mechanism, if $\mathbf{S} = \mathbf{C}\,\mathbf{C}^T$ is of rank r, the 1st r rows of $\mathbf{C}$ which are independent are selected.
More precisely, if r lines are independent, the 1st $min(n, r)$ equations are selected[3].
We write $\mathbf{S}^{-}$ the inverse of $\mathbf{S}^{<}$, it is a "pseudo-inverse" of $\mathbf{S}$.

These definitions are different from the classical pseudo-inverse $\mathbf{M}^{\dagger}$ (e.g. [36]) of a matrix $\mathbf{M}$, obtained from the singular value decomposition, for instance.

This mechanism is very useful in our case because it allows to consider the cases where :

- the *information matrices matrix are only semi-definite*, here the definite matrix "as close to $\mathbf{Q}$ as possible" is automatically used,
- the *equations are not independent* (in the sense that their linear parts are not independent at $\mathbf{x}^{\bullet}$, i.e. $\mathbf{C}$ is not of full rank), here redundant equations are eliminated,
- we have *more equations than unknowns*, since, in that case $\mathbf{C}$ can only be of rank at most n, thus no more than n equations are taken into account,
- furthermore, if an *equation is "singular"* in the sense that its gradient vanishes, then the algorithm disregards the equation at this point.

.2.2.4. Computation of the local projector. Now, using pseudo-inverses defined previously, we can efficiently solve (11), in order to obtain the 1st order solution.

Since (10) yields $\mathbf{x} = \mathbf{x}_0 - \mathbf{Q}^{+}\,\mathbf{C}^T\,\lambda$, which, combined with (9), leads to a linear equation in λ : $(\mathbf{C}\,\mathbf{Q}^{+}\,\mathbf{C}^T)\,\lambda = \mathbf{C}\,\mathbf{x}_0 - \mathbf{d} + o(\kappa||\mathbf{x} - \mathbf{x}^{\bullet}||)$ we obtain the explicit form :

$$\mathbf{x} = \mathbf{P}_{\mathbf{x}_0}(\mathbf{x}^{\bullet}) + o(\kappa||\mathbf{x} - \mathbf{x}^{\bullet}||) \text{ with } \mathbf{P}_{\mathbf{x}_0}(\mathbf{x}^{\bullet}) = \mathbf{x}_0 - \mathbf{Q}^{+}\,\mathbf{C}^T\,(\mathbf{C}\,\mathbf{Q}^{+}\,\mathbf{C}^T)^{-}\,(\mathbf{C}\,\mathbf{x}_0 - \mathbf{d}) \tag{12}$$

A step further, we can estimate the error up to the first order, since a few algebra yields :

$$\mathcal{E}^2 = ||\mathbf{x} - \bar{\mathbf{x}}||^2_{\mathbf{Q}} = ||c(\mathbf{x})||^2_{(\mathbf{C}\,\mathbf{Q}^{+}\,\mathbf{C}^T)^{-}} + o(\kappa||\mathbf{x} - \bar{\mathbf{x}}||) + o(\kappa||\mathbf{x} - \mathbf{x}^{\bullet}||) \tag{13}$$

$\bar{\mathbf{x}}$ being an unbiased estimation of $\mathbf{x}$, i.e. with $c(\bar{\mathbf{x}}) = 0$.

At the algorithmic level, we simply consider $\mathbf{L}$, the "closest" square-root decomposition of $\mathbf{Q}$, which is a lower triangular matrix :

$$\mathbf{Q} = \mathbf{L}\,\mathbf{L}^T \text{ with } \mathbf{y}_0 = \mathbf{L}^T\,\mathbf{x}_0 \text{ and } \mathbf{B} = \mathbf{L}^{-T}\,\mathbf{C}^T \tag{14}$$

and allows to have equation (12) simplified as :

$$\mathbf{L}^T\,\mathbf{x} = \mathbf{y}_0 - \mathbf{B}^T\,(\mathbf{B}\,\mathbf{B}^T)^{-1}\,(\mathbf{B}\,\mathbf{y}_0 - \mathbf{d}) \tag{15}$$

thus easily computed using the "reduced" square-root decomposition $\mathbf{M}$ of $\mathbf{B}_2 = (\mathbf{B}\,\mathbf{B}^T) = \mathbf{M}\,\mathbf{M}^T$ from :

$$\mathbf{M}\,\mathbf{M}^T\,\lambda = \mathbf{B}\,\mathbf{y}_0 - \mathbf{d} \text{ with } \mathbf{L}^T\,\mathbf{x} = \mathbf{y}_0 - \mathbf{B}^T\,\lambda \tag{16}$$

while, from (13), we have :

$$\mathcal{E}^2 = ||\mathbf{e}||^2 \text{ with } \mathbf{e} = \mathbf{M}^{-1}\, c(\mathbf{x}) \tag{17}$$

In our case, we can even fasten this computation because as defined in (8) the matrix $\mathbf{Q}$ is block diagonal. The derivation is given in appendix A.1. Several other improvements are also present in the implementation, for instance when a matrix $\mathbf{Q}$ is diagonal.

.2.2.5. Non-linear iteration and convergence. As already mentioned, the algorithm defined by the series $\mathbf{x}_{n+1} = \mathbf{P}_{\mathbf{x}_0}(\mathbf{x}_n)$ converges, on the conditions of $\mathcal{P}_1$, with a quadratic rate of convergence, to a fixed point $\mathbf{x}_\infty = \mathbf{P}_{\mathbf{x}_0}(\mathbf{x}_\infty)$ which is likely[4] a solution of (8).

Fair enough, but in practice, we cannot be sure to be "on the conditions of $\mathcal{P}_1$", but we still NEED the algorithm to always converge, hopefully to the optimal value, but a least and last, to a sub-optimal estimation.

With the simple idea that the algorithm may:
(a) compute the series $\mathbf{x}_{n+1} = \mathbf{P}_{\mathbf{x}_0}(\mathbf{x}_n)$ while this converges, whereas
(b) look for a point closer, to "smooth" the estimation, if the previous estimation becomes unstable, we propose the following algorithm :

$$\begin{array}{ll}
\textbf{Input :} & \mathbf{x}_0, \mathbf{Q}, \mathbf{c}() \\
\text{Init :} & n = 0 \\
\text{Loop :} & \delta_n = ||\mathbf{c}(\mathbf{x}_n)||_\infty \\
\quad \text{If} & \quad \delta_n < \delta_{n-1} + \epsilon_\bullet \\
\quad \text{Then} & \quad \mathbf{x}_{n+1} = \mathbf{P}_{\mathbf{x}_0}(\mathbf{x}_n);\, n{+}{+} \\
\quad \text{Else} & \quad \mathbf{x}_n = [\mathbf{x}_n + \mathbf{x}_{n-1}]\,/2 \\
\text{Until} & ||\mathbf{x}_n - \mathbf{x}_{n-1}|| < \epsilon_\bullet \\
\textbf{Return :} & \mathbf{x}_n, \mathcal{E}^2
\end{array} \tag{18}$$

We have chosen the norm $||\mathbf{x}||_\infty = max(|x^1|, |x^2|, \cdots)$ to evaluate δ_n in order to be sure that *all* equations vanish.

With this mechanism, we easily see that we compute $\mathbf{x}_n = (1-\alpha)\,\mathbf{x}_{n-1} + \alpha\,\mathbf{P}_{\mathbf{x}_0}(\mathbf{x}_{n-1})$ for some $\alpha = 1, 1/2, 1/4, \cdots$. Furthermore, the linearisation of $\mathbf{c}(\mathbf{x})$ being performed at $\mathbf{x}_{n-1}$, since $\mathbf{C}\,\mathbf{P}_{\mathbf{x}_0}(\mathbf{x}_{n-1}) - \mathbf{d} = 0$ from (12) and $\mathbf{c}(\mathbf{x}_{n-1}) = \mathbf{C}\,\mathbf{x}_{n-1} - \mathbf{d}$ from (9), writing formally:

$$h_{n-1} = (\mathbf{P}_{\mathbf{x}_0}(\mathbf{x}_{n-1}) - \mathbf{x}_{n-1})^T \left[\left. \frac{\partial^2 \mathbf{c}(\mathbf{x})}{\partial \mathbf{x}^2} \right|_{\mathbf{x}_{n-1}} \right] (\mathbf{P}_{\mathbf{x}_0}(\mathbf{x}_{n-1}) - \mathbf{x}_{n-1}) \tag{19}$$

it appears that :

$$\begin{aligned}
\mathbf{c}(\mathbf{x}_n) &= \mathbf{c}(\mathbf{x}_{n-1}) + \alpha\,\mathbf{C}\,[\mathbf{P}_{\mathbf{x}_0}(\mathbf{x}_{n-1}) - \mathbf{x}_{n-1}] + \alpha^2\, h_{n-1} + o(\alpha^2) \\
&= \mathbf{c}(\mathbf{x}_{n-1}) - \alpha\,[\mathbf{C}\,\mathbf{x}_{n-1} - \mathbf{d}] + \alpha^2\, h_{n-1} + o(\alpha^2) \\
&= (1-\alpha)\,\mathbf{c}(\mathbf{x}_{n-1}) + \alpha^2\, h_{n-1} + o(\alpha^2)
\end{aligned}$$

$\Rightarrow$

$$||c(\mathbf{x}_n)|| < (1-\alpha)\,||\mathbf{c}(\mathbf{x}_{n-1})|| + \alpha^2\,||h_{n-1}|| + o(\alpha^2) \tag{20}$$

If we choose α with $\alpha\,||h_{n-1}|| < ||\mathbf{c}(\mathbf{x}_{n-1})||$ and sufficiently small for higher order terms to be negligible, we obtain $||c(\mathbf{x}_n)|| < ||\mathbf{c}(\mathbf{x}_{n-1})||$ as desired.

In practice, this condition cannot be reached if the required α is smaller than numerical errors, for instance if $||\mathbf{c}(\mathbf{x}_{n-1})||$ becomes negligible, but we thus have converged.

This condition cannot be reached also if the criterion has strong irregularities (corresponding here to the fact that higher order terms may be preponderant) but this means that the present algorithm is not adapted to such a situation and it also must stop.

As consequence either :

- $\alpha > 0$ and $||c(\mathbf{x}_n)||$ is strictly decreasing, so that the algorithm converges
  - for $\alpha = 1$, i.e. when the computation of $\mathbf{P}_{\mathbf{x}_0}(\mathbf{x}_n)$ is stable, we cancel the $\mathbf{c}(\mathbf{x}_{n-1})$ up to the first order, as for a Newton algorithm and the convergence is quadratic,
  - for smaller $\alpha > 0$, but with $||\mathbf{c}(\mathbf{x}_n)|| < ||\mathbf{c}(\mathbf{x}_{n-1})||$, we obtain a linear convergence, the algorithm behaving as a gradient descent method,

  or
- $\alpha \to 0$ and $||\mathbf{x}_n - \mathbf{x}_{n-1}|| = \alpha\, ||\mathbf{P}_{\mathbf{x}_0}(\mathbf{x}_{n-1}) - \mathbf{x}_{n-1}||$ converges towards zero, with an exponential rate and the algorithm quickly stops.

.2.2.6. A few properties of the minimisation method. In order to better understand the behaviour of the method, let us take a look at some interesting particular cases:

Invariance with respect to linear combination of equations. If we consider $\mathbf{c}'(\mathbf{x}) = \mathbf{G}\,\mathbf{c}(\mathbf{x})$ for a general invertible matrix $\mathbf{G}$, from (12), a few algebra allows to verify that $\mathbf{P}_{\mathbf{x}_0}(\mathbf{x})$ is left unchanged. As a consequence, linear combinations or permutations of equations are meaningless.
A step further, if $\mathbf{G}$ is any rectangular matrix, using the reduced square-root allows to deal with a minimal set of equations. However, although faster than a canonical decomposition, our method does not guaranty that $\mathbf{P}_{\mathbf{x}_0}(\mathbf{x})$ is left unchanged, since it may depend on the equations ordering. This seems not to be a limitation in practice. If it would, using the singular value decomposition instead of the square-root decomposition for this part of the calculation cleans the point.

Dealing with linear constraints or measurement equations. If some of the equations in $\mathbf{c}(\mathbf{x})$ are linear, they are *directly solved in one step* by the proposed method. More explicitly, if we write $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)^T$ so that $\mathbf{c}(\mathbf{x}) = (\mathbf{G}\,\mathbf{x}_1 + \mathbf{f}, \mathbf{h}(\mathbf{x}_1, \mathbf{x}_2))$ by the virtue of (11), the equation $\mathbf{G}\,\mathbf{x}_1 + \mathbf{f} = 0$ is solved.
For instance, if $\mathbf{c}(\mathbf{x})$ is entirely linear, $\mathbf{P}_{\mathbf{x}_0}(\mathbf{x}^\bullet)$ provides directly an explicit solution (this is the well-known "QL" problem i.e. quadratic criterion with linear constraints, e.g. [14]).
A step further, if measurement equations are linear and there is no constraint on $\mathbf{q}$, the algorithm behaves as *a simple weighted least-square estimator* as easily verified from derivations given in appendix A.1, since (A5) corresponds to the normal equation of a such a criterion (e.g. [45]).

Relation with the Newton algorithm. In the particular case where (1) $n = p$, (ii) $\mathbf{C}$ is invertible, with (iii) $\mathbf{Q} = \mathbf{I}$, equation (12) simplifies to $\mathbf{x} = \mathbf{x}^\bullet - \mathbf{C}^{-1}\,\mathbf{c}(\mathbf{x}^\bullet) + o(\kappa||\mathbf{x} - \mathbf{x}^\bullet||)$ which corresponds to the classical Newton's method.
This shows that Newton's like methods can also be interpreted as "looking for the closest solution" of a set of regular equations. i.e. : $min_{\mathbf{x}}||\mathbf{x} - \mathbf{x}_0||^2 + \lambda^T\,\mathbf{c}(\mathbf{x})$.

Explicit measurement equation. If a measurement equation is explicit, i.e. $\mathbf{m_i} = \mathbf{f}(\mathbf{q})$ so that we can write $\mathbf{c}_i(\mathbf{q}, \mathbf{m_i}) = \mathbf{m_i} - \mathbf{f}(\mathbf{q})$, the algorithm minimises $||\tilde{\mathbf{m}}_i - \mathbf{f}(\mathbf{q})||^2_{\mathbf{Q}_i}$ up to the first order.
This is coherent with the fact that, in this case, we indeed want to minimise the "measurement error", as for *non-linear least-square problem.* However, we minimise the criterion proposed by [21] which has been shown to be unbiased by this author, contrary to other formulations.
In this case, each measure add $p_i = n_i$ degrees of freedom to $\mathcal{L}^2$ as discussed when deriving the definition given in (7).

In the sequel we are going to use the estimation process develope in this section for two specific purposes:

Minimal resolution without initial conditions

In the case where $\mathbf{Q}_0 = 0$ (i.e. initial conditions are not to be taken into account), while we have a "minimal set of coherent equations" (i.e. we assume that *there exists a solution in* $\mathbf{q}$ *to the equations* $\mathbf{c}_{(\tilde{\mathbf{m}}_1, \cdots, \tilde{\mathbf{m}}_i, \cdots)}(\mathbf{q}) = (\mathbf{c}_0(\mathbf{q}), \cdots, \mathbf{c}_i(\mathbf{q}, \tilde{\mathbf{m}}_i)) = 0$ *in a neighbourhood of the observed measure*), the criterion simply minimises:

$$\min_{(\bar{\mathbf{q}},\bar{\tilde{\mathbf{m}}}_1,\cdots,\bar{\tilde{\mathbf{m}}}_i,\cdots)} \sum_{i=1}^{M} \frac{1}{2}||\tilde{\mathbf{m}}_i - \mathbf{m}_i||^2$$

which is indeed minimal for $\tilde{\mathbf{m}}_i = \mathbf{m}_i$ irrespectively of the constraints.
This means that the system is solved only with respect to $\mathbf{q}$ and not with respect to $\mathbf{m}_i$.
It thus corresponds to the projection problem :

$$min_{\mathbf{q}} max_{\lambda} ||\mathbf{q} - \mathbf{q}_0||^2_{\mathbf{I}} + \lambda^T \mathbf{c}(\mathbf{q}) \text{ with } \mathbf{c}(\mathbf{q}) = (\mathbf{c}_0(\mathbf{q}), \cdots, \mathbf{c}_i(\mathbf{q}, \tilde{\mathbf{m}}_i), \cdots)^T \tag{21}$$

In this case, each measure add no degree of freedom to $\mathcal{L}^2$, again in coherence with what has been discussed for (7).

Estimating the precision with respect to a measure

Let us consider *another measure* $\mathbf{m}_\bullet$ which has not been used to obtain a given parameter estimate $\tilde{\mathbf{q}}$.
We may want to estimate how such a measure "matches" this parameter estimate. A coherent way of solving this problem is to determine:

$$min_{\tilde{\mathbf{m}}_\bullet} max_{\lambda} ||\tilde{\mathbf{m}}_\bullet - \mathbf{m}_\bullet||^2_{\mathbf{Q}_\bullet} + \lambda^T \mathbf{c}_\bullet(\tilde{\mathbf{q}}, \tilde{\mathbf{m}}_\bullet) \tag{22}$$

i.e. to find the "corrected measure" $\tilde{\mathbf{m}}_\bullet$ given the parameter estimate.
If we apply the relation (13) to this criterion we have, up to the first order, an evaluation of the distance between the corrected measure and the true (unknown) measure, i.e. the "bias" related to this measure:

$$\mathcal{E}^2_\bullet = ||\tilde{\mathbf{m}}_\bullet - \bar{\mathbf{m}}_\bullet||^2_{\mathbf{Q}_\bullet} \simeq ||\mathbf{c}_\bullet(\tilde{\mathbf{q}}, \tilde{\mathbf{m}}_\bullet)||^2_{(\mathbf{C}_\bullet \mathbf{Q}^+_\bullet \mathbf{C}^T_\bullet)^-} \tag{23}$$

with $\mathbf{C}_\bullet = \left.\frac{\partial \mathbf{c}_\bullet(\mathbf{q},\mathbf{m}_\bullet)}{\partial \mathbf{q}}\right|_{(\tilde{\mathbf{q}},\tilde{\mathbf{m}}_\bullet)}$, while $\mathcal{D}^2_\bullet = ||\tilde{\mathbf{m}}_\bullet - \mathbf{m}_\bullet||^2_{\mathbf{Q}_\bullet}$ evaluates the distance between the observed and corrected measure, i.e. its "imprecision".
Both errors $\mathcal{E}^2_\bullet$ and $\mathcal{D}^2_\bullet$ have to be taken into account.
Since from (7) it appears that we have $p_\bullet$ degrees of freedom, we define as "measurement error":

$$\frac{\mathcal{L}^2_\bullet}{p_\bullet} = \frac{||\tilde{\mathbf{m}}_\bullet - \mathbf{m}_\bullet||^2_{\mathbf{Q}_\bullet} + ||\mathbf{c}_\bullet(\tilde{\mathbf{q}}, \tilde{\mathbf{m}}_\bullet)||^2_{(\mathbf{C}_\bullet \mathbf{Q}^+_\bullet \mathbf{C}^T_\bullet)^-}}{p_\bullet} = \frac{\mathcal{E}^2_\bullet + \mathcal{D}^2_\bullet}{p_\bullet} \tag{24}$$

.3. Dealing with outliers while using a hierarchy of models.

.3.1. Solving as a robust local estimation problem

Considering "realistic" estimation problems, we also have to deal with the problem of outliers, i.e. the fact we have measures not corresponding to the model under estimation, but to other "objects".

In order to be *robust* with respect to such artifacts we have implemented a classical (see for instance [31, 19, 34, 35]) *randomised estimation method*, i.e. we repeatedly solve the estimation problem, selecting randomly a set of measures, with the hope that at least one of them will not contain outliers. A "good" sample should be detected by the fact that its estimation looks more coherent than for other ones.

Implementing a randomised estimation method.

This is implemented here as follows:

1. We randomly select a *minimal* number of measures M, so that $n = p + \cdots + p_i + \cdots$ without any initial information, i.e. $\mathbf{Q}_0 = 0$. According to the previous discussion, this induces $\mathbf{m}_i = \tilde{\mathbf{m}}_i$ and we simply have to solve the projection problem given in (21).

This provides an estimate $\tilde{\mathbf{q}}$ of the parameter, compatible with this random set of measures. If ν is the percentage of relevant measures, the probability of having selecting a correct set of measure (i.e. a set of measures without outliers) after T sampling is easy to estimate :

$$P = \left[1 - \left[1 - \left(1 - \frac{\nu}{100}\right)^M\right]^T\right] \tag{25}$$

It is thus obvious that, the smallest the sub-set of measures, the more chance to detect a unique object. This is why we choose a minimal set of measures. However the numerical estimation is not expected to be very precise, since we take a small number of measures into account. It thus must be refined, as discussed in the sequel.

2. Before that, we must compute, for each measure, an indicator of its coherence with respect to the estimated parameter. This is done using the criterion proposed in (22) and the related error computed in (24).
   The expected histogram of such an error distribution is schematised in Fig. 3.
   It is expected that small errors correspond to true approximate measures, whereas higher errors correspond to outliers. This will be discussed in the next section.
3. From such a distribution, in order to estimate the validity of the estimate $\tilde{\mathbf{q}}$, two main strategies (see for instance [31] for a review) are used, either :
   $\mathcal{S}_\mathcal{A}$ **finding a sufficient number of "good" measures** counting the percentage of measures ν which error is below a fixed threshold $\mathcal{L}^2_\bullet$ (e.g. [5] this being known as "RANSAC-like" methods), finding the random estimate which allows to model the maximal number of measures; or
   $\mathcal{S}_\mathcal{B}$ **finding a sufficiently small error** considering the maximal error $\mathcal{L}^2_\bullet$ of the $\nu\%$ first measures, i.e. those with a smaller error (e.g. the median error if $\nu = 50\%$ [48], this being known as "trimmed least median of squares" methods), finding the random estimate which has a minimal error at this percentage.
   Here, we will combine these two ideas in the next section, defining the "relevance" of the estimate.
   In both cases, we may either :
   (a) choose a fixed number T_{max} of iterations, based on a chosen probability of error, as given in (25) and take the best measure or
   (b) repeat until a *relevant* estimate is found.
   We finally have to refine the obtained estimation.

Defining the relevance of an estimate.

As studied in details in [37], robust methods may easily reject random outliers but may fail if several set of inliers (i.e. several set of measures corresponding to a given parameter estimate) are present, i.e. if we have a multi-modal distribution.

According to this author, the observed bias is due to the fact that the used criteria are only based on the residual error. Here we try to limit this problem assuming that *if we have a multi-modal distribution, and an estimation which combine more than one distribution or includes outliers, the error histogram will be "flatter" around zero* whereas if a given parameter estimate fits with a unique set of inliers the error histogram will be sharper around zero.

From this, we may define the model "relevance" by analysing qualitatively the error distribution, as illustrated in Fig. 3.

From general experimental observations (e.g. [22, 31, 19, 34, 35]), it seems that we can consider :
(a) the distribution of the true approximate measures is "flat" at the origin,
(b) the distribution of the outliers, if randomly distributed, is almost constant, at the origin.

Such a distribution may thus be characterised by :
(1) α : the distribution amplitude, at zero ; the highest α, the more "good" measures,
(2) γ : the distribution convexity, at zero : the highest γ, the smaller the average error for these "good" measures,
$\rightarrow$ the histogram distribution for inliers being of the form:

$$N_i(\mathcal{L}^2) = \alpha \left(1 - \gamma \frac{\mathcal{L}^2}{2}\right) + o((\mathcal{L}^2)^2) \tag{26}$$

(3) β : the bias introduced by outliers, at zero,
$\rightarrow$ the histogram distribution for outliers being of the form:

$$N_o(\mathcal{L}^2) = \beta + o((\mathcal{L}^2)^2) \tag{27}$$

With these general parameters, we can define the model relevance as an indicator maximising *both* quantities α and γ, together. One classical trick, to (i) maximise two quantities together, in such a way that (ii) none of them is negligible, is to maximise their product. We will follow this track here.

In our context, we *intentionally* do not want to refer to any particular model, e.g. statistical distribution may not be Gaussian. On the contrary, we only make use of the rather generic properties of the error distribution, introduced here.

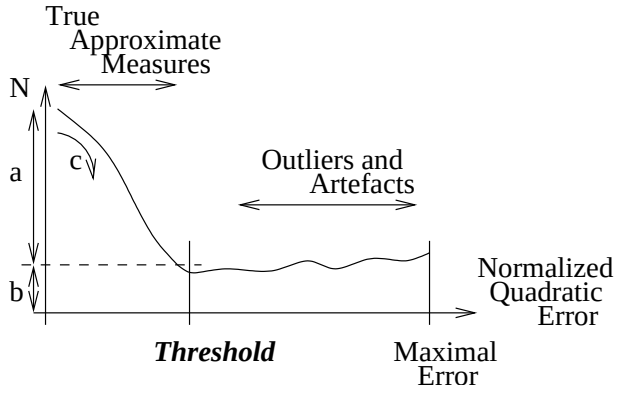


Fig. 3. The form of error distribution in the presence of outliers.

We thus easily can relate the distribution to its momentum around zero, i.e. :

$$\mu_n = \int_0^{\mathcal{L}^2_\bullet} N(\mathcal{L}^2)\, d\,\mathcal{L}^2 = (\alpha + \beta)\, \frac{(\mathcal{L}^2_\bullet)^{n+1}}{n+1} - \frac{\alpha\,\gamma}{2}\, \frac{(\mathcal{L}^2_\bullet)^{n+3}}{n+3} + o((\mathcal{L}^2_\bullet)^{n+4}) \tag{28}$$

so that we obtain : $\alpha\,\gamma = \frac{12}{(\mathcal{L}^2_\bullet)^3} \left[\mu_0 - 2\,\frac{\mu_1}{\mathcal{L}^2_\bullet}\right] + o(\mathcal{L}^2_\bullet)$.

As a consequence, in coherence with the previous discussion, M being the total number of measures, the model relevance can be defined as :

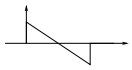
$$\mathcal{R} = \frac{1}{M} \left[\mu_0 - 2\,\frac{\mu_1}{\mathcal{L}^2_\bullet}\right] < 1 \tag{29}$$

Here, the value $\mathcal{L}^2_\bullet$ is the value under which we expect the error distribution to be close to its second-order expansion. In practice, $\mathcal{L}^2_\bullet$ *is the value under which we expect errors to correspond only to uncertainty on inliers, not outliers.* It is user-defined. In fact, this value is not highly significant, since it does not act as a threshold but only as an order of magnitude.

This is very easily computed on the data, much faster than the distribution median for instance.

In fact, this corresponds to a convolution of the error distribution, i.e :

$$R = \int_0^{\mathcal{L}^2_\bullet} r(\mathcal{L}^2)\, N(\mathcal{L}^2)\, d\,\mathcal{L}^2 \text{ with } r(\mathcal{L}^2) = \left[1 - 2\,\frac{\mathcal{L}^2}{\mathcal{L}^2_\bullet}\right] \text{ i.e. of the form } \quad (30)$$



easily calculated without any explicit analysis of the distribution.

In comparison, "RANSAC-like" methods correspond to a convolution with $r(\mathcal{L}^2) = 1$. i.e. only consider μ_0, whereas the present methods does not only "count" the samples but attempt also to evaluate the error shape.

This quantity can also be related to the average slope of the distribution. More precisely, if we write $N(\mathcal{L}^2) = N_0 - N_1\,\mathcal{L}^2 + o((\mathcal{L}^2)^2)$ we obtain $N_1 = \frac{6}{(\mathcal{L}^2_\bullet)^2}\,\mathcal{R} + o(\mathcal{L}^2_\bullet)$. This means that our "relevance" also describes the "thickness" of the distribution.

A step ahead, we may better understand the role of this quantity by looking at the relevance for some characteristic examples of distribution :

- If we consider a *uniform* distribution of the error, as illustrated in Fig. 4.A, the relevance is :

$$R = \textbf{ if } \mathcal{L}^2_0 < \mathcal{L}^2_\bullet \textbf{ then } \frac{M_0}{M}\,\left[1 - \frac{\mathcal{L}^2_0}{\mathcal{L}^2_\bullet}\right] \textbf{ else } 0 \quad (31)$$

  where M_0 is the total number of good measures and $\mathcal{L}^2_0$ the maximal quadratic error for these measures. We thus verify that the relevance increases with both (i) the number of good measures and (ii) the precision on these measures.
  It also shows that :
  (a) the relevance is *positive if and only if the quadratic error for good measures is below the threshold* $\mathcal{L}^2_\bullet$,
  (b) constant components of the distribution have no influence on the relevance,
  (c) the relevance is maximal (i.e. equal to 1) for a distribution without outliers (i.e. $M = M_0$) and with an infinite precision (i.e. $\mathcal{L}^2_0 \to 0$).
- If we consider an *exponential* distribution of the error, as illustrated in Fig. 4.B, the relevance is still of the form : $R = \frac{M_0}{M}\,\left[1 - \frac{\mathcal{L}^2_0}{\mathcal{L}^2_\bullet}\right] + o((exp(\frac{\mathcal{L}^2_0}{\mathcal{L}^2_\bullet}))^2)$ where M_0 is again the total number of good measures, while $\mathcal{L}^2_0$ is also in relation with the precision of the measures.

  Here, we have chosen $\mathcal{L}^2_0$ so that $N(\mathcal{L}^2_0) = N(0)/e^2$, with $N(\mathcal{L}^2) = \frac{M_0}{\mathcal{L}^2_0/2}\,e^{-\frac{\mathcal{L}^2}{\mathcal{L}^2_0/2}} + b$ to be in coherence with the previous formula.
  This thus shows that the defined relevance is qualitatively not dependent upon the form of these two distributions, as expected.

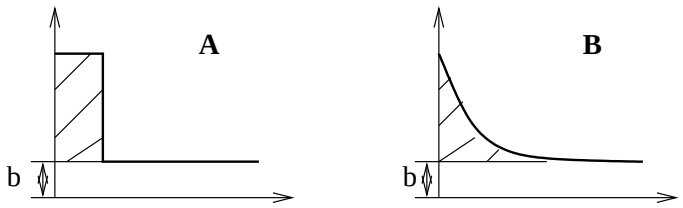


Fig. 4. Histogram of the quadratic error, for uniform **A** or exponential **B** distributions. The dashed part of the histogram corresponds to the "good" measures, the rest being outliers.

To complete the discussion, let us note that if our assumption that the distribution is flat at the origin is wrong, i.e. if we have $N(\mathcal{L}^2) = N(0) + N'(0)\,\mathcal{L}^2 + N''(0)\,\frac{(\mathcal{L}^2)^2}{2} + o((\mathcal{L}^2)^3)$, we still obtain :

$R = \frac{(\mathcal{L}^2_\bullet)^3}{12}\,(N''(0) + 2\,N'(0)/\mathcal{L}^2_\bullet)$ also related to both the distribution slope and convexity at the origin, in coherence with our requirement.

If, finally, we compare our approach with the two classes of methods formalised in robust statistics, that is ($\mathcal{S}_\mathcal{A}$) is counting samples under a given (somehow arbitrary) threshold (i.e. considering μ_0 in our case) or ($\mathcal{S}_\mathcal{B}$) measuring the precision as the maximal error, for a percentage of the best measures, it appears that we indeed compute value also related to the "precision" of the estimates, for measures with small errors, as in ($\mathcal{S}_\mathcal{B}$). As such we have indeed an indicator which is a synthesis of both points of view.

Evaluating the relevance indicator.

In order to verify the efficency of this indicator, we have considered the paradigm proposed by [37]. Let us denote $\mathcal{U}(a,b)$ the uniform distribution in the $[a,b]$ interval and $\mathcal{G}(\mu,\sigma)$ the normal distribution of mean μ and standard-deviation σ. Following [37] we choose a distribution with "good" and "bad" data of the form

$$p = (1-\epsilon_0)[\epsilon_1 \underbrace{\mathcal{G}(\mu_1,\sigma_1)}_{\text{principal inliers}} + (1-\epsilon_1) \underbrace{\mathcal{G}(\mu_2,\sigma_2)}_{\text{secondary inliers}}] + \epsilon_0 \underbrace{\mathcal{U}(0,m_0)}_{\text{outliers}} \tag{32}$$

where ϵ_0 is the proportion of outliers, $\epsilon_1 > 0.5$ the proportion of inliers in the main distribution, while m_0, (μ_1,σ_1) and (μ_2,σ_2) describe the outlier, principal and secondary inliers distributions, respectively. See [44] for an example of simulation and method details.

In our case we have set $m_0 = 100$, $\mu_1 = 10, \mu_2 = \mu_1 + \delta$, $\sigma_1 = \sigma_2 = 5$ and varied the proportions ϵ_0 and ϵ_1 of inliers and outliers and the proximity δ between both inlier distributions. For estimated values $\tilde{v} \in \{0..\mu_2\}$ we detect the estimation corresponding to the best estimation and analyse the bias β of such an estimation, for the three methods discussed here. In order to have the three methods working in their best conditions we have consider $\mathcal{S}_\mathcal{A}$ with a threshold equal to the standard deviation of the inliers distribution and $\mathcal{S}_\mathcal{B}$ with a trimmed least median of squares $\nu = \frac{(1-\epsilon_0)\,\epsilon_1}{2}$ [31].

ϵ_0	0.1	0.2	0.5	0.8	0.2	0.2	0.2	0.2	0.2	0.2
ϵ_1	1	1	1	1	0.8	0.6	0.8	0.6	0.8	0.6
δ	0	0	0	0	50	50	20	20	10	10
$\mathcal{R}$	0	-1	-1	1	1	1	1	1	0	1
$\mathcal{S}_\mathcal{A}$	0	0	-1	1	2	3	1	2	0	2
$\mathcal{S}_\mathcal{B}$	0	1	1	1	5	6	4	3	2	2

Fig. 5. Bias estimation in our simulation, using relevance $\mathcal{R}$, "RANSAC-like" $\mathcal{S}_\mathcal{A}$ or "trimmed least median of squares" $\mathcal{S}_\mathcal{B}$ methods. See text for details.

This leads to the results given in Fig. 5. In the presence of outliers all three methods are very robust since the estimation bias is 0 or 1 in any cases. When a second set of inliers appears the $\mathcal{S}_\mathcal{B}$ method becomes unstable and tends to provide an average result between both modalities. This is the reason why all bias are positive (i.e. in the direction of the second set of inliers) when considering a bi-modal data set.

Surprisingly perhaps, the distance between both modalities have no significant influence on the bias. On one hand, we might have assumed that if the higher the distance between modalities, the less the influence on the bias. But, on the other hand, the higher the distance between modalities, the higher the average value of both estimations, which tends to be what the estimators choose.

As analysed by [37], the "RANSAC-like" $\mathcal{S}_{\mathcal{A}}$ allows to obtain better results, while our method appears as a small but significant improvement of this class of method. Other run of the simulations may tend to show that our method performs better when there are large distances between modalities, whereas this advantage with respect to "RANSAC-like" methods seems to disappear for closer distributions.

.3.2. Using a hierarchy of models to estimate parameters.

Minimising the previous non-linear criterion may not be sufficient to obtain a relevant estimation of a parameter for two reasons:

1. estimating a parameter does not only mean calculating a numerical value but choosing which model best fits the data,
2. since we find only a local estimate of the parameter, the initial condition is determinant, otherwise the previous minimisation process may not converge.

In the latter case a relevant initial value may be found by a simpler model.

Defining a hierarchy of models.

To face these two problems we propose that *a lattice of models* is to be specified for the parameter estimation, as follows:

- each estimation problem must have a "null-model" (most constrained model) as *reference* ,
- each model is a "generalisation" of another models (its parents) relaxing or changing some constraints. Since there is also a "general model" with no equations (and no interest !) this forms a lattice as shown in Fig. 6.

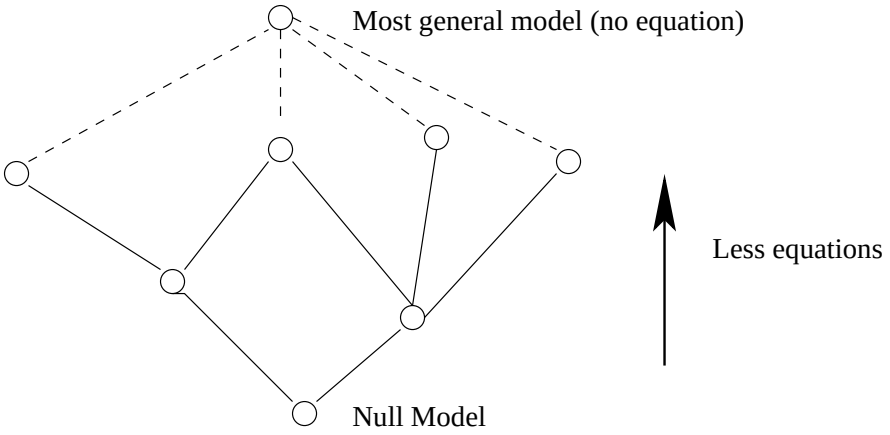


Fig. 6. Representing a lattice of models.

In order to integrate this general idea in our estimation framework, we consider that :

- for all models[5], we have to estimate (i) a common parameter $\mathbf{q}$ with (ii) the same measurement equations $\mathbf{c}_i(\mathbf{q}, \mathbf{m}_i)$,
- two models differ by their constraints $\mathbf{c}_0(\mathbf{q})$ on the parameter i.e. by the number of equation p, so that :
  - the model "complexity", for a given set of measure, is the number of degrees of freedom $d = \underbrace{\left[n + \sum_i p_i\right]}_{d_{max}} - p$ defined in (7) while
  - the model "cost", used to compare two models, is the normalised criterion defined in (7). If the criterion $\mathcal{L}^2$ decreases regularly with the number of constraints, as expected, we obtain a profile as schematised in Fig. 7.

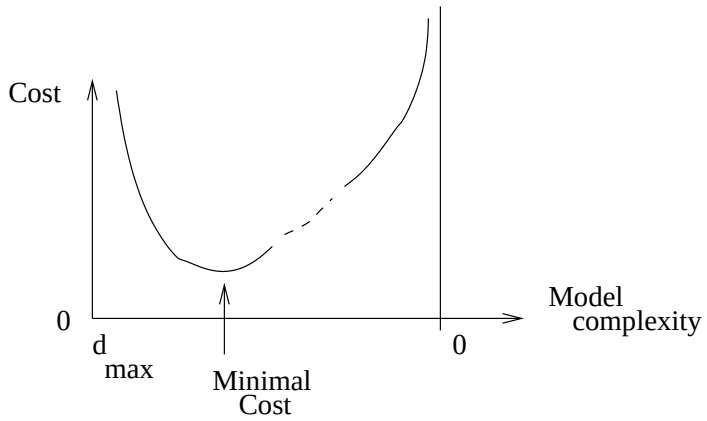


Fig. 7. The expected form of the model cost.

Within this framework the problem is formalised as follows: *finding the most specific model of optimal cost*, i.e. :
(i) for a given model, we choose a more general (less specific) alternative only if the cost is slightly lower,
(ii) we do not estimate a model unless its parents (i.e. more specific models) have been estimated.

From one model to another.

The relationship between the estimated parameter $\tilde{\mathbf{q}}$ of a more specific model $\mathcal{M}$ and the estimated parameter $\tilde{\mathbf{q}}'$ of a more general model $\mathcal{M}'$ may be summarised in the following two equations:

$$\begin{array}{rcl} \mathcal{L}^2 & = & ||\tilde{\mathbf{q}} - \mathbf{q}_0||^2_{\mathbf{Q}_0} \; + \; ||\tilde{\mathbf{m}} - \mathbf{m}||^2_{\mathbf{Q}_m} \; + \; \lambda^T \mathbf{c}(\tilde{\mathbf{q}}, \tilde{\mathbf{m}}) \\ \mathcal{L}^{2'} & = & ||\tilde{\mathbf{q}}' - \tilde{\mathbf{q}}||^2_{\tilde{\mathbf{Q}}} \; + \; ||\tilde{\mathbf{m}}' - \mathbf{m}'||^2_{\mathbf{Q}'_m} \; + \; \lambda^T \mathbf{c}'(\tilde{\mathbf{q}}', \tilde{\mathbf{m}}') \end{array} \tag{33}$$

providing that we have rewritten (4), in both cases, in a compact form as in (8), i.e.:
- $\mathbf{q}_0$ is an initial estimate for $\mathcal{M}$, while $\tilde{\mathbf{q}}$ is the initial/default estimate for $\mathcal{M}'$,
- $\mathbf{m} = (\mathbf{m}_1 \cdots)$, with the corresponding matrix $\mathbf{Q}_m$ stacks the measures taken into account in estimating $\mathcal{M}$, and $\mathbf{m}'$, with $\mathbf{Q}'_m$, the same for $\mathcal{M}$', these two sets differ as discussed here,
- $\mathbf{c}(\tilde{\mathbf{q}}, \tilde{\mathbf{m}})$ represents all measurements equations used for $\mathcal{M}$ plus the constraints on the parameter $\tilde{\mathbf{q}}$, and $\mathbf{c}'(\tilde{\mathbf{q}}', \tilde{\mathbf{m}}')$ the same for $\mathcal{M}'$,

Here, the key point is the fact that we have computed from (4) for measures taken into account for the estimation of $\tilde{\mathbf{q}}$ or (22) for other measures a "corrected" estimation $\tilde{\mathbf{m}}_i$ such that $\mathbf{c}_i(\tilde{\mathbf{q}}, \tilde{\mathbf{m}}_i) = 0$.

As a consequence, if we initiate the non-linear minimisation process with $(\tilde{\mathbf{q}}'_0, \tilde{\mathbf{m}}'_0) \leftarrow (\tilde{\mathbf{q}}, \tilde{\mathbf{m}})$ we thus have $\mathbf{c}'(\tilde{\mathbf{q}}', \tilde{\mathbf{m}}') = 0$.

Therefore, from (12) and using the same derivation as for (20) we obtain:

$$||\mathbf{c}'(\tilde{\mathbf{q}}'_n, \tilde{\mathbf{m}}'_n)|| = o(||(\tilde{\mathbf{q}}'_n, \tilde{\mathbf{m}}'_n) - (\tilde{\mathbf{q}}'_{n-1}, \tilde{\mathbf{m}}'_{n-1}||)) \tag{34}$$

i.e. during the algorithm minimisation, starting at a point for which the constraints are verified, we maintain this property at each step (see section .2.2 for a discussion on convergence). This means that we indeed stay in the conditions of convergence reviewed in $\mathcal{P}_1$ and we also increase the speed of convergence.

Furthermore, we also obtain:

$$||(\tilde{\mathbf{q}}'_n, \tilde{\mathbf{m}}'_n) - (\tilde{\mathbf{q}}, \tilde{\mathbf{m}})||^2_{\mathbf{Q}} = ||(\tilde{\mathbf{q}}'_{n-1}, \tilde{\mathbf{m}}'_{n-1}) - (\tilde{\mathbf{q}}, \tilde{\mathbf{m}})||^2_{\mathbf{C}^T (\mathbf{C}\mathbf{Q}^+ \mathbf{C}^T) \mathbf{C}} + o(||(\tilde{\mathbf{q}}'_n, \tilde{\mathbf{m}}'_n) - (\tilde{\mathbf{q}}'_{n-1}, \tilde{\mathbf{m}}'_{n-1}||)) \tag{35}$$

i.e. the new estimate is "as close as possible" to $(\tilde{\mathbf{q}}, \tilde{\mathbf{m}})$ is a direction tangent to constraints, as expected.

In practice, these two properties allow the multi-model algorithm to efficiently converge from one solution to another.

Deciding between two models.

In order to be able to "tune" this process of comparison, we add the feature that a more general model $\mathcal{M}'$ of cost $\mathcal{L}^{2'}/d'$ is chosen with respect to a more specific model $\mathcal{M}$ of cost $\mathcal{L}^2/d$, *if and only if it decreases the cost by a given factor* $0 < \Phi(d, d') \leq 1$ so that the comparison criterion is finally :

$$\frac{\mathcal{L}^{2'}}{d'} < \Phi(d, d') \frac{\mathcal{L}^2}{d} \tag{36}$$

Tuning this parameter allows to deal with more or less "conservative" estimation, the lower $\Phi(d, d')$, the more specific model will be preferred by the system.

At this level of specification, the function $\Phi(d, d')$ is user defined. However, this mechanism may be related to a rigorous statistical test, considering specific hypotheses, as detailed in appendix A.2.

More general formalisms to specify $\Phi(d, d')$ may be designed, for instance learning $\Phi(d, d')$ from a set of reference data. However, the proposed method is robust enough to allow us to consider, for the experimentations reported here .... $\Phi(d, d') = 1$! For more tricky situations, developments given in appendix A.2 suggest that the functions :

$$\Phi(d, d') = e^{-\kappa (d'-d)/d'} \text{ with } \kappa \in \{1..10\} \text{ or } \Phi(d, d') = \frac{d}{d'} \tag{37}$$

should be rather efficient, because they are relevant approximations of well formalised statistical thresholds.

Integrating robust estimation in multi-modelling.

In order to implement such a method in cooperation with robust estimation, we make use of the following assumption : *given a relevant (i.e. estimated without outliers) model for a minimal set of measure, a more general model, thus requiring more measures in its minimal set, can always consider the measures of its parent as not outliers*, because they indeed also verify a sub-set of the model equations they are coherent with. Therefore when looking for a more general model we only have to randomly select the *additional* measures. This remark dramatically reduces the chance to randomly select outliers, since in (25) the number M is only the number of additional measures.

A step further, we must remark that given a relevant model, all measures fitting this model are coherent with its constraints and thus will not help estimating *new constraints*. As a consequence, *given a relevant model and one generalisation of it, additional measures sampled to estimate the general model must be*

taken outside the set of measures fitting the more original model. This again restrains the number of measures to sample and thus increases the chance to randomly select a relevant estimate.

But, much more important is the fact that this may *avoid selecting singular configurations of points* for a given model. The user just has to know which are the singular configurations for a given model and put in the lattice structure of the model hierarchy more restrictive models which correspond to such a singular configuration. As a consequence, because new measures selected will not verify less restrictive models, as required previously, they will not be singular and this will make the job.

Implementing robust multi-modelling.

In order to implement these ideas,

1. a model "state" is thus represented by :
   (i) the estimated parameter $\tilde{\mathbf{q}}$ and its related quadratic precision $\tilde{\mathbf{Q}}$,
   (ii) the indexes of the points sampled to estimate the state,
   (iii) the indexes of the points not coherent, i.e. the outliers for this model, while
2. the "model" is specified through :
   (i) its name,
   (ii) its constraints and intrinsic cost,
   (iii) a list of "alternatives" i.e. models less specific, with less constraints,
   (iv) with their related cost factor Φ,
   (v) a list of models which are "parents" of these.

Using this data structure, the previous ideas are implemented by the following algorithm :

Initialisation Put the null-model, with a user provided initial parameter value $\mathbf{q}_0$, in a *candidate list.*
Iteration

- For each *model* of the candidate list :
  - randomly select a set of "additional" measures in the set of points not coherent with the parent model (if any),
  - estimate :
    1. the model parameter, using the parent parameter as initial value $\mathbf{q}_0$, solving the projection problem in (21) using the algorithm in (18);
    2. the coherence of each measure, solving the projection problem in (22) using the algorithm in (18);
    3. the model relevance, as formalised in (29)
  - If the model is more relevant than previously estimated parameter for this model :
    * delete previously estimations of this model,
    * then :
      1. threshold the outliers set, as illustrated in Fig. 3 and discussed in section .3.1;
      2. refine the model parameter estimation, applying the algorithm in (18);
      3. repeat step 1 and 2, to stabilise the estimation as discussed in section .3.1;
      4. evaluate its cost, from (7).
    * If the cost is lower than his parent its alternatives are put in the candidate list.
    * the model is removed from the candidate list.
  - Else repeat the iteration selecting randomly a model in the candidate list.

Termination and Output The model, with minimal cost below a given threshold, is chosen as "best" model.

Algorithm states. Analysing this algorithm, we easily see that it can be in three states : *(I) initialisation*, *(A) model-available* (when the candidate list is no more empty), *(T) termination*.

Current output. As a consequence, the present algorithm can always output the model of minimal cost as the "best current model". In state (I), the "best" model may be the null-model, as default value.

Adding/Deleting measures. Another nice property is the fact we easily can add or delete measures to the set of measures input to the algorithm, without having to reinitialise the whole estimation process.

Of course, if in state (I) we just have to add or remove the measure. Since nothing has been estimated, this will have no influence on the output.

If in state (A) or (T), we have to estimate the potential influence of the measure on the already estimated admissible models, with three cases :
- if the added/deleted measure belongs to the outliers (this being tested by comparing its error in the sense of (24) to the model threshold estimated, as illustrated in Fig. 3) then nothing is to be done,
- else an added measure may leads to a *more complex* model and the admissible model as thus to be refined with this new measure and then put again in the candidate list,
- while a deleted measure may leads to a *more specific* model so that *parents* of the admissible model are to be reconsidered and have thus to be put again in the candidate list, the admissible model being removed. In the worst case, if the measure has p_i degrees of freedom, the chosen parents must are those who have just at least $p + p_i$ constraints, i.e. p_i less degrees of freedom, in the hierarchy.

As a consequence, adding or deleting a measure makes the algorithm switch back to state (A) but without having to restart from state (I) which is an obvious gain of performance.

However, contrary to "incremental" algorithms such as the Extended Kalman Filter (see for instance [45] for discussion), the criterion itself is always entirely refined and reconsidered for each new measure, in order to avoid to accumulate bias in the estimation. Otherwise, the estimation result would have depend upon the order of arrival of the measures.

.4. Software integration and experimentation.

.4.1. The estimation module architecture.

At the *integration* level, in order to be usable in an effective software system, the estimation algorithm has to be embedded in an input/output module, as described in Fig. 8.

Based on the previous specifications, the *module interface* should contain at the *data flow* level :

data input , i.e. the measures with their quadratic precision,
  – which may be set on several "input channels", defined by different measurement equations,
state input , i.e. additional "constants" in the model and measurement equations,
  – which allows upper-layers of the system (e.g. a user interface) to tune the module[6],
data output , i.e. the estimated parameter and its quadratic precision, plus the indexes of the measures not considered as outliers,
status output , i.e. the chosen model and the related normalised criterion value.

Considering these data ports, the module interface must be able to :

- *get* the data and status output,
  – which involve the property of being able to provide the information "at any time".
  This specification is easily achieved because, as discussed before, a sub-optimal or default model estimation is always available.
- *set/unset* a data input $\mathbf{m}_i, \mathbf{Q}_i$ on a given input channel,
  – which involve a mechanism of measure addition/deletion as discussed in the previous section.

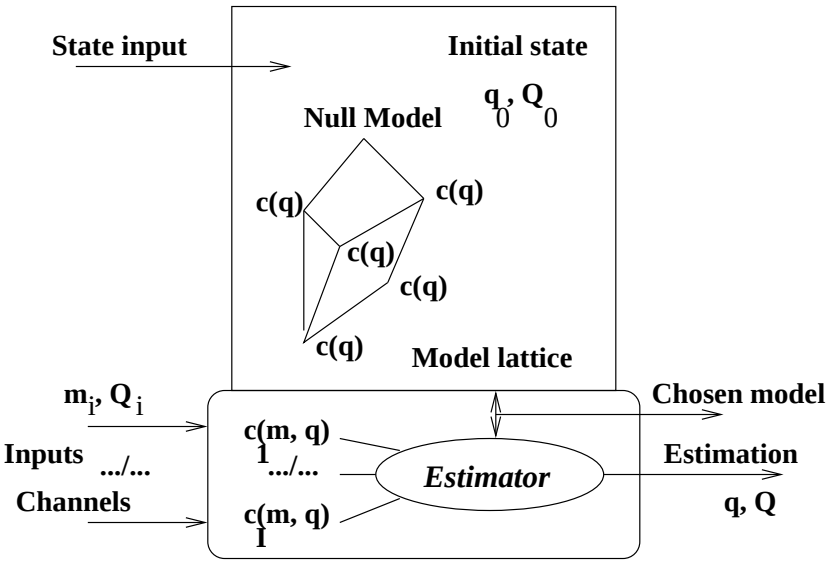


Fig. 8. Architecture of the estimation module.

- *set/modify* the state input,
  – which involve the action of "restarting" the estimation, since the estimation criterion has been changed.
  This specification is also easily achieved with the present algorithm, by cleaning the candidate list of models and reintroducing the null-model in it.

Similarly, at the *control level*, this *module interface* must be able to :

send *start/stop* or *suspend/resume* signals to the estimation process,
– which involve the property of being able to halt properly all computations. This is yet another easily implementable feature in our context, because the algorithm is based on two simple loops :
(i) the iteration mechanism of the algorithm itself and
(ii) the calculation loop of the algorithm given in (18).
In order to react to such an event without any particular threading mechanism, it is very easy to `check`, at the end of each iteration if a "suspend/stop flag" has been raised (and then react properly to it), applying the schema :

```
init();
 while(iter())
  check();
finalize();
```

which is to be implemented for each loop in the code.
Therefore the calculation is guaranteed to be suspended and eventually restarted very easily, without any need of throwing exception.
In other words, we have *decomposed the code in terms of loops and "straight-line programs"* [15] so that, given the data size, we precisely know the amount of operation of each step, especially the `iter()` step. This allows to be sure it will stop and to calculate approximately when. As a consequence, as discussed in works like [2], the execution of such a program is fully controllable in a real-time constrained environment.

receive a signal when :

(a) a new admissible *model has been found* and, at last,

(b) the estimation is *terminated*,

– while, here, the key point is that only "good-news" signals have to be received from the module, whereas other exceptions are not expected. This is due to the fact, that the general algorithm avoid any kind of exception.

From these basic signals other more specific signals may be derived, e.g. an alert when a "reasonable" model but not necessarily "optimal" has been found, for instance, say, a model which cost is below a required threshold.

Towards symbolic computations over the estimation module.

A step ahead, the software architecture of the estimation module has been discussed with the goal of being integrated in reactive software environments or within applications with real-time constraints. Beside what has been given on code properties and architecture, let us discuss what concerns optimisation of the code. In fact, most of the computation time is *spent in computing the projectors* given in (12) and appendix A.1 build out of simple fixed-size loops of polynomial computations. However, as soon as the dimensions of the parameter and/or measure are known, those loops can be expanded, while in many cases (e.g. explicit measurement equations) the expression to be computed may be simplified. Aside the actual "demonstration" code which is optimised only at the compiler level (e.g. using in-line methods) the function itself is easy to optimise, performing *partial evaluation* such as *constant propagations* in this part of the code [8].

Furthermore, it has been shown [27] that such multi-model estimation module does not only require numeric but also symbolic derivations, because : (i) some parameter components may be eliminated using the constraints which are linear, so that evaluation is only to be performed on a reduced set of equations and variables, (ii) for some huge model hierarchy it is necessary to generate "at execution time" a model, given its parent in the hierarchy. However, redundant models may be generated and it is necessary to (iii) obtain a *canonical form* for the model constraints, which is not a trivial problem. This is why we have limited the present mechanism to a pre-defined static hierarchy of models.

.4.2. Experimenting the estimation module.

Considering retinal displacements between two views.

We still consider the well known problem of the "fundamental" matrix estimation.

Following [43], we have designed a model hierarchy for a realistic subset of specific Euclidean displacements. A much larger experimentation over such models has been conducted in [27], applying a restrained form of the present formalism.

We show in Fig. 9 and Fig. 10 two experimental results, from a set of experimentations on several indoors and outdoors scenes, in order to test our model inference.

The model cost has been given here in pixel, i.e. it is the square root of the least-square average distance between the measured point locations and those predicted by the estimated model. This is a very common way to estimate the estimation precision [28, 39, 49].

In both cases, the chosen model corresponds to the expected displacement. In [27] other results have been obtained with manual camera displacements, qualitatively realized as a specific displacement and estimated, using model comparison, with a model coherent with the displacement realized.

In a complementary set of experiment [11] using a small hierarchy of models, as illustrated in Fig. 11, the estimation method has been able to detect displacements corresponding to planar structures. This

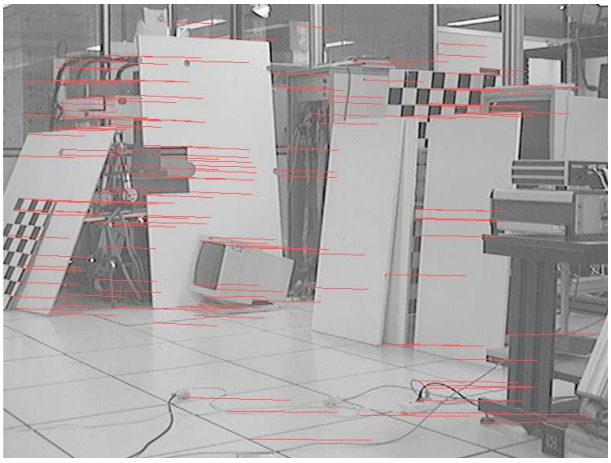


Fig. 9. A partial view of the model hierarchy for a specific displacement estimation. See text for details. The displacement was a pure translation in the X-direction.

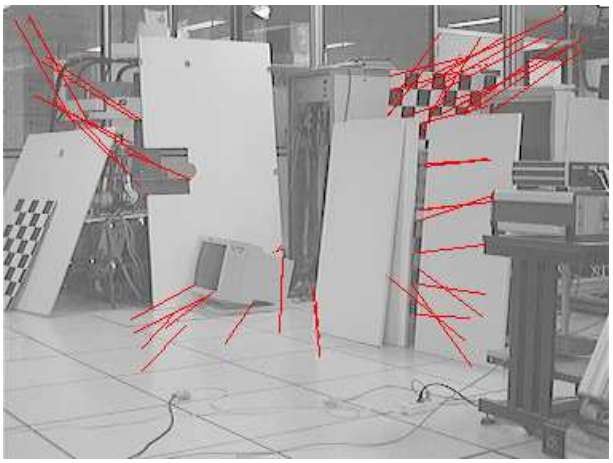


Fig. 10. A partial view of the model hierarchy for a specific displacement estimation. See text for details. The displacement was a zoom of the camera.

allows to "segment" them in the scene. The method formalised in this paper had also already been used in a restrained form to evaluate different models of planar rigid displacements in [46].

In the left part of Fig. 11 we see that it has been possible to identify planar structures of the scene, including the "horizon", i.e. points at infinity which rigid displacement only correspond to the rotational part of the displacement. In the right part of Fig. 11 we see that, due to relatively small amount of data points, the estimation process has estimated the two moving objects of the scene as "shallows", i.e. planar objects, because it was numerically more stable than estimating the parameters of full rigid objects.

In order to quantify these results we have analysed the residual error for different displacements as shown in Fig. 12. This again illustrates the efficency of the method.

Although we provide here numerical results for future comparisons, while more data is available in [27], it is rather difficult to compare with available results of the literature such as [28, 41, 39] because we *do not estimate the same quantities*, as discussed in the introduction.

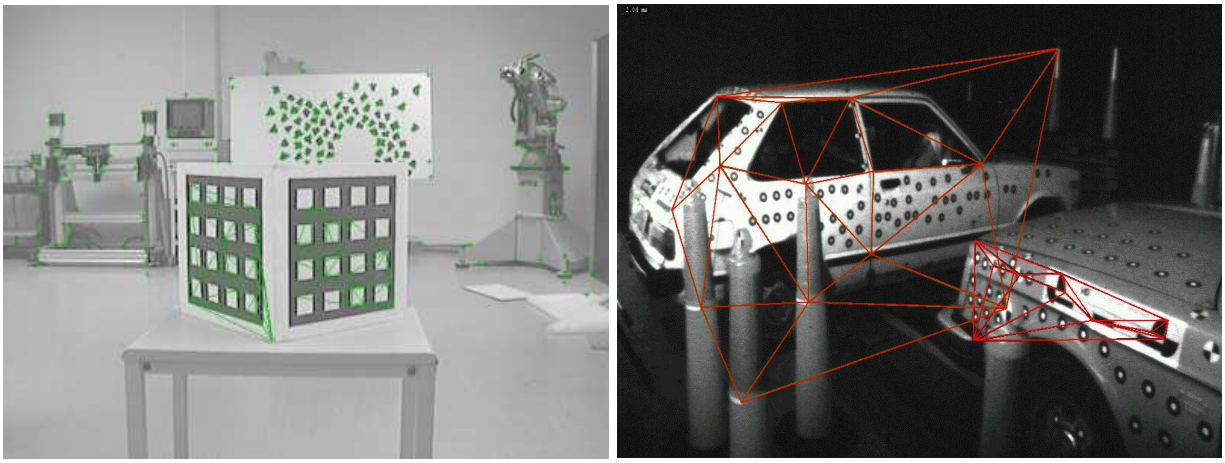


Fig. 11. Detecting planar structures : estimation of a model in a clustered environment is possible with the method.

Effective displacement	Estimated displacement	Residual error (pixel) for this displacement	Residual error (pixel) for a general displacement
No displacement	No displacement	0.056	0.078
Pure retinal translation	Pure retinal translation	0.456	0.879
Retinal displacement	Retinal displacement	0.766	0.947
Planar displacement	Planar displacement	1.766	2.947
Pure translation	Pure translation	0.342	1.023
Zoom	Zoom	0.342	1.023

Fig. 12. Illustrating the method numerical robustness, comparing the residual error obtained for the specific displacement with respect to a general one. Here "retinal" displacement means a displacement which does not move the retinal plane.

.5. Conclusion.

We have revisited the problem of parametric estimation considering non-linear implicit measurement equations and parameter constraints, plus robust estimation in the presence of outliers and multi-model comparisons.

More specifically, a projection algorithm based on generalisations of square-root decompositions has been proposed to allow an efficient and numerically stable local resolution of a set of non-linear equations, while a robust estimation module of a hierarchy of non-linear models has been designed and validated.

The non trivial discussion on the software implementation shows that there is a non negligible gap between an "algorithm" and a software "module", the former being unusable without the latter.

This method has been designed with the perspective of being used as a basic module in parameter adjustment routines [2]. Such a general parametric learning capability is mandatory when considering adaptive property of a system [18]. In [44] , its application to general system modelling is discussed in details.

Appendix

A.1. Computing the local projector.

Considering the criterion in the form of (4) it is clear that the method developed for the "compact" form (8) is not optimal because it does not make use of the fact $\mathbf{Q}$ and $\mathbf{C}$ are block diagonal matrices. It is however trivial, although rather painful, to explicit it and obtain a faster calculation.

The linearisation of the non-linear equations at a point $(\mathbf{q}^\bullet, \cdots, \mathbf{m}_i^\bullet, \cdots)$ may be written:

$$\begin{array}{rcll} \mathbf{c}_0(\mathbf{q}) & = & \mathbf{C}_0\,\mathbf{q} - \mathbf{d}_0 & \text{with } \mathbf{C}_0 = \frac{\partial \mathbf{c}_0(\mathbf{q})}{\partial \mathbf{q}}\Big|_{\mathbf{q}^\bullet} \text{ and } \mathbf{d}_0 = \mathbf{C}_0\,\mathbf{q}^\bullet - \mathbf{c}_0(\mathbf{q}^\bullet) \\ & + & o(||\mathbf{q}-\mathbf{q}^\bullet||) & \\ \mathbf{c}_i(\mathbf{q},\mathbf{m}_i) & = & \mathbf{C}_i\,\mathbf{q} + \mathbf{D}_i\,\mathbf{m}_i - \mathbf{d}_i & \text{with } \mathbf{C}_i = \frac{\partial \mathbf{c}_i(\mathbf{q},\mathbf{m}_i)}{\partial \mathbf{q}}\Big|_{(\mathbf{q}^\bullet,\mathbf{m}_i^\bullet)},\ \mathbf{D}_i = \frac{\partial \mathbf{c}_i(\mathbf{q},\mathbf{m}_i)}{\partial \mathbf{m}_i}\Big|_{(\mathbf{q}^\bullet,\mathbf{m}_i^\bullet)} \\ & + & o(||\mathbf{q}-\mathbf{q}^\bullet||) + o(||\mathbf{m}_i-\mathbf{m}_i^\bullet||) & \text{and } \mathbf{d}_i = \mathbf{C}_i\,\mathbf{q}^\bullet + \mathbf{D}_i\,\mathbf{m}_i^\bullet - \mathbf{c}_i(\mathbf{q}^\bullet,\mathbf{m}_i{}^\bullet) \end{array} \tag{A1}$$

while the corresponding normal equations are :

$$\begin{array}{rcll} 0 & = & \frac{\partial \mathcal{L}_\lambda^2}{\partial \mathbf{q}}^T & = \mathbf{Q}_0\,(\tilde{\mathbf{q}}-\mathbf{q}_0) + \mathbf{C}_0^T\,\lambda_0 + \sum_i \mathbf{C}_i^T\,\lambda_i \\ 0 & = & \frac{\partial \mathcal{L}_\lambda^2}{\partial \mathbf{m}_i}^T & = \mathbf{Q}_i\,(\tilde{\mathbf{m}}_i-\mathbf{m}_i) + \mathbf{D}_i^T\,\lambda_i \end{array} \tag{A2}$$

so that the same algebra used to derived (12) leads to (up to the first order) :

$$\mathbf{C}_i\,\tilde{\mathbf{q}} + \mathbf{D}_i\,\tilde{\mathbf{m}}_i - \mathbf{d}_i = \mathbf{S}_i\,\lambda_i \text{ with } \mathbf{S}_i = \mathbf{D}_i\,\mathbf{Q}_i^{-1}\,\mathbf{D}_i^T \tag{A3}$$

used to obtain an estimation $\tilde{\mathbf{q}}$ of the parameter from :

$$\mathbf{Q}_0\,(\tilde{\mathbf{q}}-\mathbf{q}_0) + \mathbf{C}_0^T\,\lambda_0 + \mathbf{A}\,\tilde{\mathbf{q}} - \mathbf{b} = 0 \text{ with } \mathbf{A} = \sum_i \mathbf{C}_i^T\,\mathbf{S}_i^{-1}\,\mathbf{C}_i \text{ and } \mathbf{b} = -\sum_i \mathbf{C}_i^T\,\mathbf{S}_i^{-1}\,(\mathbf{D}_i\,\mathbf{m}_i - \mathbf{d}_i) \tag{A4}$$

which may also be written:

$$\left[\mathbf{Q}_0 + \sum_i \mathbf{C}_i^T\,\mathbf{S}_i^{-1}\,\mathbf{C}_i\right]\,\tilde{\mathbf{q}} = \mathbf{Q}_0\,\mathbf{q}_0 + \sum_i \mathbf{C}_i^T\,\mathbf{S}_i^{-1}\,(\mathbf{D}_i\,\mathbf{m}_i - \mathbf{d}_i) - \mathbf{C}_0^T\,\lambda_0 \tag{A5}$$

so that we have to solve the linear system :

$$\begin{pmatrix} \mathbf{Q}_0 + \mathbf{A} & \mathbf{C}_0^T \\ \mathbf{C}_0 & \mathbf{0} \end{pmatrix} \begin{pmatrix} \tilde{\mathbf{q}} \\ \lambda_0 \end{pmatrix} = \begin{pmatrix} \mathbf{Q}_0\,\mathbf{q}_0 + \mathbf{b} \\ \mathbf{d}_0 \end{pmatrix} \tag{A6}$$

which allows to estimate $(\tilde{\mathbf{q}}, \cdots, \tilde{\mathbf{m}}_i, \cdots) = \mathbf{P}_{(\mathbf{q}_0,\cdots,\mathbf{m}_i,\cdots)}(\mathbf{q}^\bullet, \cdots, \mathbf{m}_i^\bullet, \cdots)$, because the corrected measures, from the previous equations, are given by :

$$\tilde{\mathbf{m}}_i = \left[\mathbf{m}_i - \mathbf{Q}_i^{-1}\,\mathbf{D}_i^T\,\mathbf{S}_i^{-1}\,(\mathbf{D}_i\,\mathbf{m}_i - \mathbf{d}_i)\right] - \left[\mathbf{Q}_i^{-1}\,\mathbf{D}_i^T\,\mathbf{S}_i^{-1}\,\mathbf{C}_i\right]\,\tilde{\mathbf{q}} \tag{A7}$$

From the previous derivations, the criterion may be finally written :

$$\mathcal{L}^2 = ||\mathbf{q}-\tilde{\mathbf{q}}||_{\mathbf{Q}}^2 + \tilde{\mathcal{L}}^2 \tag{A8}$$

with $\tilde{\mathcal{L}}^2 = \tilde{\mathbf{q}}^T\,(\mathbf{Q}_0 + \mathbf{A})\,\tilde{\mathbf{q}} - 2\,\tilde{\mathbf{q}}^T\,(\mathbf{Q}_0\,\mathbf{q}_0 + \mathbf{b}) + (\mathbf{q}_0^T\,\mathbf{Q}_0\,\mathbf{q}_0 + c)$
while $c = \sum_i (\mathbf{D}_i\,\mathbf{m}_i - \mathbf{d}_i)^T\,\mathbf{S}_i^{-1}\,(\mathbf{D}_i\,\mathbf{m}_i - \mathbf{d}_i)$
so that its optimal value equals $\tilde{\mathcal{L}}^2$.

We also verify that $\tilde{\mathbf{Q}} = \mathbf{Q}_0 + \mathbf{A}$ which demonstrates (6) as expected.

Following the same method as for the simple projection problem of section .2.2, a fast calculation of $(\tilde{\mathbf{q}}, \cdots, \tilde{\mathbf{m}}_i, \cdots) = \mathbf{P}_{(\mathbf{q}_0,\cdots,\mathbf{m}_i,\cdots)}(\mathbf{q}^\bullet, \cdots, \mathbf{m}_i^\bullet, \cdots)$ can be derived [44].

With this calculation the major algorithm complexity is of $o(n^3 + \sum_i n_i^3 + p^3 + \sum_i p_i^3)$ instead of $o((n + \sum_i n_i)^3 + (p + \sum_i p_i)^3)$, thus much faster.

A.2. Statistical interpretation of the estimation.

We had defined our estimation problem as minimising a quadratic distance of the form :

$$\mathcal{L}^2 = \tfrac{1}{2}\,(\mathbf{x}-\tilde{\mathbf{x}})^T\,\mathbf{Q}\,(\mathbf{x}-\tilde{\mathbf{x}})$$

under the constraints : $\mathbf{c}(\mathbf{x}) = 0$.

If, now, we consider that $\mathbf{x}$ is a random variable with a Gaussian density of mean $\tilde{\mathbf{x}}$ and covariance $\mathbf{Q}^{-1}$ its density is given by:

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n/det(\mathbf{Q})}}\,e^{-\frac{1}{2}\,\Xi^2} \quad with \quad \Xi^2 = 2\,\mathcal{L}^2 = (\mathbf{x}-\tilde{\mathbf{x}})^T\,\mathbf{Q}\,(\mathbf{x}-\tilde{\mathbf{x}}) \tag{A9}$$

so that minimising this so-called Mahanalobis distance Ξ^2 is equivalent of maximising the probability, i.e. the "likehood" of the estimate. As being a random variable, what is minimised in truth is indeed the expectation $\bar{\mathcal{L}}^2 = E[\mathcal{L}^2]$ of the quantity.

This model is valid for linear systems (i.e. if $\mathbf{c}(\mathbf{x}) = 0$ are p linear equations)and Gaussian distributions.

A step ahead, the Mahanalobis distance follows a chi-square distribution of $r = n-p$ degrees of freedom which probability density function is:

$$\mu_r(\xi^2) = \frac{1}{2^{r/2}\Gamma(r/2)}(\xi^2)^{r/2-1}e^{(\xi^2)/2} \text{ with } \left\{ \begin{array}{rl} E[\xi^2] = r & V\,[(\xi^2)] = 2\,r \\ arg\,max_{\xi^2}\,\left[\gamma_r(\xi^2)\right] & = \; r-2 \end{array} \right. \tag{A10}$$

defined in $[0,\infty[$ where $\Gamma(r) = \int_0^\infty t^{r-1}e^{-t}dt$. For a given value $\xi^2(\mathbf{x})$, $P(\xi_0^2;r) = \int_0^{\xi_0^2}\,\mu_r(\xi^2)\,d\xi^2$ is the probability, for a *correct model* the observed value to be lower that the threshold ξ_0^2 while $1-P(\xi_0^2;n)$ is the probability, even for a correct model, the observed value to be higher than ξ_0^2.

As a consequence, considering an initial estimate $\mathbf{q}_0$ of covariance $\mathbf{Q}_0^{-1}$ and a set of measures $\mathbf{m}_i$ of covariance $\mathbf{Q}_i^{-1}$, with the corresponding equations, minimising the expectation of the criterion given in (4) corresponds exactly to minimise the Mahanalobis about all available information, i.e. maximise the likehood of the estimate.

This statistical interpretation is the one chosen by Kanatani [21].

Presenting the AIC criterion

In order to evaluate the estimation, Kanatani proposes, following Akaike [1], to develop an absolute statistical criterion.

He considers a set of measures $\mathbf{m}_i$, with :
(i) their *true* values $\bar{\mathbf{m}}_i$ so that we measure $\mathbf{m}_i = \bar{\mathbf{m}}_i + \epsilon_i$ where ϵ_i is a Gaussian white noise with zero mean and covariance $\mathbf{Q}_i^{-1}$,
(ii) their *estimated* value $\tilde{\mathbf{m}}_i$, defined by equation (4),
and particularly,
(iii) a set of new *virtual* measures $\mathbf{m}_i^*$ with no relation with the other measures but having the same statistical distribution, i.e. the same covariances $\mathbf{Q}_i^{-1}$.

The main idea is that a "good" parameter is not the one which is optimal for the measures used to estimate it (because it is already tuned for these measures) but optimal for "new" measures, i.e. a parameter which correctly *predicts* the data. So that the chosen statistical criterion is :

$$\bar{\mathcal{L}}_*^2 = E\left[\frac{1}{2}||\mathbf{q}-\mathbf{q}_0||_{\mathbf{Q}_0}^2 + \sum_{i=1}^{M}\,\frac{1}{2}||\mathbf{m}_i^*-\tilde{\mathbf{m}}_i||_{\mathbf{Q}_i}^2\right] \tag{A11}$$

given the related constraints.

If we want to estimate $\bar{\mathcal{L}}_*^2$ from what has been calculated, i.e. $\mathcal{L}^2$, we may write from (4) :

$$\bar{\mathcal{L}}_*^2 = \mathcal{L}^2 + d \quad \text{with} \quad d = E\left[\frac{1}{2}\sum_{i=1}^{M} ||\mathbf{m}_i^* - \tilde{\mathbf{m}}_i||^2_{\mathbf{Q}_i} - ||\mathbf{m}_i - \tilde{\mathbf{m}}_i||^2_{\mathbf{Q}_i}\right] \tag{A12}$$

given the related constraints. This is an unbiased estimator of $\bar{\mathcal{L}}_*^2$ since both sizes of this equation have the same expectation. Here d is the expectation of chi-square random variable, thus equal to its number of degrees of freedom as reviewed in (A10).

As a consequence, a more general model $\mathcal{M}'$ of cost $\mathcal{L}^{2'}$ with $d' > 1$ degrees of freedom is chosen with respect to a more specific model $\mathcal{M}$ of cost $\mathcal{L}^2$ with $d \geq 1$ degrees of freedom if and only if $\mathcal{L}^{2'} + d' < \mathcal{L}^2 + d$ which can be written :

$$\frac{\mathcal{L}^{2'}}{d'} < \Phi(d,d')\frac{\mathcal{L}^2}{d} - \Psi(d,d') \text{ with } \left\{ \begin{array}{rcccl} 0 & \leq & \Phi(d,d') & = & \frac{d}{d'} \quad \leq 1 \\ 0 & \leq & \Psi(d,d') & = & \frac{1}{d} - \frac{1}{d'} \leq 1 << \frac{\mathcal{L}^{2'}}{d'} \end{array} \right. \tag{A13}$$

so that, considering that $\Psi(d,d')$ is negligible in the expression, we see that the formalism is roughly equivalent of choosing the ratio of the number of degrees of freedom in (36) as function $\Phi(d,d')$.

Now we can estimate d from :

$$\begin{aligned}
d &= \tfrac{1}{2}\textstyle\sum_{i=1}^{M} E\left[||\mathbf{m}_i^* - \tilde{\mathbf{m}}_i||^2_{\mathbf{Q}_i}\right] - E\left[||\mathbf{m}_i - \tilde{\mathbf{m}}_i||^2_{\mathbf{Q}_i}\right] \\
&\quad \text{... from the previous equation,} \\
&= \tfrac{1}{2}\textstyle\sum_{i=1}^{M} \left[\underbrace{E\left[||\mathbf{m}_i^* - \bar{\mathbf{m}}_i||^2_{\mathbf{Q}_i}\right]}_{measurement\ error} + \underbrace{E\left[||\bar{\mathbf{m}}_i - \tilde{\mathbf{m}}_i||^2_{\mathbf{Q}_i}\right]}_{estimation\ error}\right] - E\left[||\mathbf{m}_i - \tilde{\mathbf{m}}_i||^2_{\mathbf{Q}_i}\right] \\
&\quad \text{... since both errors are not correlated,} \\
&= \tfrac{1}{2}\textstyle\sum_{i=1}^{M} E\left[||\mathbf{m}_i - \bar{\mathbf{m}}_i||^2_{\mathbf{Q}_i}\right] + E\left[||\bar{\mathbf{m}}_i - \tilde{\mathbf{m}}_i||^2_{\mathbf{Q}_i}\right] - E\left[||\mathbf{m}_i - \tilde{\mathbf{m}}_i||^2_{\mathbf{Q}_i}\right] \\
&\quad \text{... since } \mathbf{m}_i^* \text{ and } \mathbf{m}_i \text{ have the same distribution,} \\
&= \tfrac{1}{2}\textstyle\sum_{i=1}^{M} \left[\underbrace{E\left[||\mathbf{m}_i - \tilde{\mathbf{m}}_i||^2_{\mathbf{Q}_i}\right]}_{measurement\ correction} + \underbrace{E\left[||\tilde{\mathbf{m}}_i - \bar{\mathbf{m}}_i||^2_{\mathbf{Q}_i}\right]}_{estimation\ error}\right] \\
&\quad + E\left[||\bar{\mathbf{m}}_i - \tilde{\mathbf{m}}_i||^2_{\mathbf{Q}_i}\right] - E\left[||\mathbf{m}_i - \tilde{\mathbf{m}}_i||^2_{\mathbf{Q}_i}\right] \\
&\quad \text{... since both quantities are also not correlated,} \\
&= \textstyle\sum_{i=1}^{M} E\left[||\tilde{\mathbf{m}}_i - \bar{\mathbf{m}}_i||^2_{\mathbf{Q}_i}\right]
\end{aligned}$$

thus, from (A10), finally equal to the *estimation error degrees of freedom* since it follows a chi-square distribution. Each measure is defined by $n_i - p_i$ degrees of freedom and also function of the parameter itself defined by $n - p$ degrees of freedom, we thus obtain : $d = n - p + \sum_i n_i - p_i$.

Unfortunately, as discussed for instance in [38, 4] this criterion is usually selecting models with a too many parameters (see [27] for a more complete discussion) whereas other more flexible criteria (e.g. [38] for a review) are always to be tuned by a few non-intuitive parameters. A step further, in our case, the number of degrees of freedom is not counted as for the AIC, since for each measure we consider the dimension of the *measurement correction* $v_i = \mathbf{m}_i - \tilde{\mathbf{m}}_i$ given the measurement equation, i.e. p_i, and not of the *estimation error* number of degrees of freedom, i.e. $n_i - p_i$. It has been discussed all along this paper, and it particular for some important particular cases (see section .2.2) that this is a more relevant point of view.

An alternative to the AIC criterion

Another point of view might be to forget about estimating the "absolute" cost of a given model, but only compare two models, using "relative" costs values.

A well-established methodology, so called "extra sum-of-squares" principle (e.g.[9]), provides such a method for comparing models in a hierarchy. Here, we wish to test whether the extra set of parameters defined in the more general model $\mathcal{M}'$ (of cost $\mathcal{L}^{2'}$ with d' degrees of freedom) is statically significant with respect to the more specific model $\mathcal{M}$ (of cost $\mathcal{L}^2$ with d degrees of freedom) (i.e. if we can *reject* the corresponding null hypothesis $\mathcal{H}_0$ that this extra-parameterisation is negligible). This extra sum-of-squares due to $\mathcal{M}'$ after $\mathcal{M}$ (and in addition to it) is then defined as $\Xi^2(\mathcal{M}'|\mathcal{M}) = \mathcal{L}^2 - \mathcal{L}^{2'}$ which is a chi-square distribution with $p' - p$ degrees of freedom, under $\mathcal{H}_0$. If $\mathcal{H}_0$ is not true then $\Xi^2(\mathcal{M}'|\mathcal{M})$ has a non-central chi-square distribution, but still independent of $\mathcal{M}'$. Therefore, the following F-statistics expresses the evidence against $\mathcal{H}_0$:

$$f = \left[\frac{\bar{\mathcal{L}}^2 - \bar{\mathcal{L}}^{2'}}{d' - d} \right] \Bigg/ \left[\frac{\bar{\mathcal{L}}^{2'}}{d'} \right] \tag{A14}$$

which probability density function is :

$$\nu_{\delta,d'}(f) = \frac{\Gamma(\frac{\delta+d'}{2})}{\Gamma(\frac{\delta}{2})\,\Gamma(\frac{d'}{2})} \frac{(\frac{\delta}{d'})^{\frac{\delta}{2}}\, f^{\frac{\delta-2}{2}}}{(1+\frac{\delta}{d'}\,f)^{\frac{\delta+d'}{2}}} \tag{A15}$$

with $\delta = d' - d$. Significance can the be assessed by comparing the previous statistics with the inverse cumulative density function of (A15).

Coming back to our notations, this is equivalent to compare :

$$\frac{\mathcal{L}^{2'}}{d'} < \Phi_P(d,d')\,\frac{\mathcal{L}^2}{d} \text{ with } \Phi_P(d,d') = \frac{d}{d'}\,\frac{1}{1+f\,(d'-d)/d'} \tag{A16}$$

The $\Phi_P(d,d')$ values, for $d' \in \{1..8\}$ and $d \in \{d'+1..8\}$, given a probability of $P = 0.5$ are shown in the following matrix :

$$\begin{bmatrix} 0.25 & 0.14 & 0.09 & 0.06 & 0.04 & 0.03 & 0.02 \\ & 0.38 & 0.25 & 0.17 & 0.12 & 0.09 & 0.07 \\ & & 0.47 & 0.34 & 0.25 & 0.19 & 0.14 \\ & & & 0.54 & 0.41 & 0.31 & 0.25 \\ & & & & 0.60 & 0.47 & 0.37 \\ & & & & & 0.64 & 0.52 \\ & & & & & & 0.68 \end{bmatrix}$$

allowing to have a look at the order of magnitude of such values.

Their "exponential-like" profiles with respect to the number of degrees of freedom is illustrated in Fig. 13.

Unfortunately, these values are only valid : (i) in the linear case, (ii) for a given probability threshold and (iii) in the case where the measurements errors have a Gaussian distribution. This is why, in our formalism (see (36)) we consider this is application dependent and thus user defined. Generalisation to other modelisation of the errors may be a challenging subject, although the approximate profile given in (A17) seems to be quite efficient for model comparisons such as in [27].

More precisely, we have verified numerically that the function $\Phi_P(d,d')$ derived from this small piece of theory is easily approximated by :

$$\Phi_P(d,d') = e^{-\kappa\,(d'-d)/d'} \; with \; \begin{cases} \kappa = 8.8 \text{ for } P = 0.5 \\ \kappa = 3.1 \text{ for } P = 0.9 \end{cases} \tag{A17}$$

with a precision of about 5%.

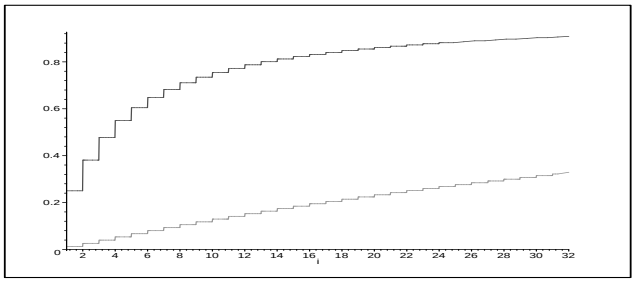


Fig. 13. Profile of $\Phi_P(d'+1, d')$, $d' \in \{1, 32\}$, for $P = 0.5$ (upper curve) and $P = 0.9$ (lower curve).

Notes

1. *http://www.ail.cs.gunma-u.ac.jp/~kanatani/e*
2. **Notations:** We write vectors and matrices in bold letters, matrices being written with capital letters. The duals of vectors are represented as the transpose of a vector and scalars in italic, the dot-product being written as $\mathbf{x}^T\mathbf{y}$ and the cross-product $\mathbf{x} \times \mathbf{y}$ or $[\mathbf{x}]_\times \mathbf{y}$. The identity matrix is written $\mathbf{I}$. We represent the components of a matrix or a vector using superscripts from 0 to 2, e.g.: $\mathbf{x} = (x^0, x^1, x^2)^T$. Here $\mathbf{x} \equiv \mathbf{y}$ means $\lambda\,\mathbf{x} = \mathbf{y}$ for some $\lambda \neq 0$.
3. Although, this is exactly what will be needed in the sequel, we could also have easily select another set of equations, for instance those which errors are maximal. This is easily obtained by sorting the set of equations before applying the reduced square-root decomposition.
4. In fact, since using the normal equations of the criterion, we may -in theory- converge towards a maximum or a saddle point of the criterion. This in fact would be detected by the algorithmic schema described here.
5. This is not a limitation, because (i) if two models do not share the parameter components, it is always possible to concatenate the two parameters and assign default values on $\mathbf{q}_0$ for those components which will not be evaluated for a given model; on the other hand (ii) for a model $\mathcal{M}_1$ with a measurement equation $\mathbf{c}_i^1(\mathbf{q}, ; \mathbf{m}_i) = 0$ and another model $\mathcal{M}_2$ with a measurement equation $; \mathbf{c}_i^2(\mathbf{q}, \mathbf{m}_i) = 0$ we can use the common measurement equation $(q_{n+1} - 2)\,\mathbf{c}_i^1(\mathbf{q}, \mathbf{m}_i) + (q_{n+1} - 1)\,; \mathbf{c}_i^2(\mathbf{q}, \mathbf{m}_i) = 0$ using a new qualitative variable $q_{n+1} \in \{1, 2\}$ with $c_0^{n+1}(\mathbf{q}, q_{n+1}) = q_{n+1} - i$ with $i \in \{1, 2\}$ as additional constraint, depending on the model.
6. These constants have not been made explicit in the previous sections because they are transparent for the estimation process, but are mandatory for a given module to be adapted to different configurations.

References

1. H. Akaike. On entropy maximisation principle. *Applications of Statistics*, pages 27–41, 1977.
2. S. Arias. *Formalisation et Intégration en Vision par Ordinateur*. PhD thesis, University of Nice, 1999.
3. Y. Bard. *Nonlinear Parameter Estimation*. Academic Press, 1974.
4. C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated classification likelihood. Technical Report RR-3521, INRIA Rhône-Alpes, Oct. 1998.
5. R. C. Bolles and M. A. Fischler. A RANSAC-based approach to model fitting and its application to finding cylinders in range data. In *International Joint Conference on Artificial Intelligence*, pages 637–643, Vancouver, Canada, Aug. 1981.
6. F. Chaumette and P. Rives. Modélisation et Calibration d'une caméra. In *AFCET*, pages 527–536, 1989.
7. W. Chojnacki, M. J. Brooks, and A. Hengel. Rationalising Kanatani's method of renormalisation in computer vision. In *Statistical Methods for Image Processing, Uppsala, Sweden, August*, pages 61–63, 1999.
8. O. Danvy, R. Glück, and P. Thiemann, editors. *Partial Evaluation*, volume 1110 of *Lecture Notes in Computer Science*. Springer Verlag, 1996.
9. N. Draper and H. Smith. *Applied Regression Analysis*. John Wiley and Sons, New-York, 1981.
10. R. Enciso and T. Vieville. Experimental self-calibration from four views. In C. Braccini-etal, editor, *8th International Conference Image Analysis and Processing (ICIAP'95)*, volume 974 of *Lecture Notes in Computer Science*, pages 307–312, San remo, Italy, Sept. 1995. Springer.
11. F. Gaspard and T. Viéville. Hierarchical visual perception without calibration. RR 3002, INRIA Sophia-Antipolis, Oct. 1996.

12. F. Gaspard and T. Viéville. Non linear minimization and visual localization of a plane. In *The 6th International Conference on Information Systems, Analysis and Synthesis*, volume VIII, pages 366–371, 2000.
13. P. E. Gill, W. Murray, and M. H. Wright. *Practical Optimization*. Academic Press, 1993.
14. A. Goelb. *Applied Optimal Estimation*. MIT Press, 1974.
15. J. Grimm, L. Pottier, and N. Rostaing-Schmidt. Optimal time and minimum space-time product for reversing a certain class of programs. RR 2794, INRIA, 1996.
16. R. Hartley. In defense on the 8-point algorithm. *PAMI*, 19(6):580–593, 1997.
17. R. I. Hartley. Estimation of relative camera positions for uncalibrated cameras. In G. Sandini, editor, *Proceedings of the 2nd European Conference on Computer Vision*, pages 579–587, Santa Margherita, Italy, May 1992. Springer-Verlag.
18. J. H. Holland. *Adaptation in Naturel and Artificial Systems*. PhD thesis, university of Michigan, 1975.
19. P. Huber. *Robust Statistics*. John Wiley & Sons, New York, 1981.
20. K. Kanatani. *Geometric computation for machine vision*. Oxford university press, 1992.
21. K. Kanatani. Automatic singularity test for motion analysis by an information criterion. In B. Buxton, editor, *Proceedings of the 4th European Conference on Computer Vision*, pages 697–708, Cambridge, UK, Apr. 1996.
22. K. Kanatani. *Statistical Optimization for Geometric Computation: Theory and Practice*. Elsevier Science, 1996.
23. K. Kanatani. Statistical optimization and geometric inference in computer vision. *Phil. Trans. R. Soc. Lond.*, A(356):1303–1320, 1998.
24. R. Lee. *Optimal Estimation, identification and control*. M.I.T. Press, Cambridge, 1964.
25. Y. Leedan. *Statistical analysis of quadratic problems in computer vision*. PhD thesis, Department of Electrical and Computer Engineering, Rutgers University, 1997.
26. Y. Leedan and P. Meer. Heteroscedastic regression in computer vision: problems with bilinear constraint. *The International Journal of Computer Vision*, 37(2):1–24, June 2000.
27. D. Lingrand. *Analyse Adaptative du Mouvement dans des Séquences Monoculaires non Calibrées*. PhD thesis, Université de Nice - Sophia Antipolis, INRIA, Sophia Antipolis, France, July 1999.
28. Q.-T. Luong, R. Deriche, O. Faugeras, and T. Papadopoulo. On determining the fundamental matrix: Analysis of different methods and experimental results. Technical Report 1894, INRIA, 1993.
29. B. Matei and P. Meer. A general method for errors-in-variables problems in computer vision. In *Computer Vision and Pattern Recognition proceedings*, June 2000.
30. S. J. Maybank and O. D. Faugeras. A theory of self-calibration of a moving camera. *The International Journal of Computer Vision*, 8(2):123–152, Aug. 1992.
31. P. Meer, D. Mintz, A. Rosenfeld, and D. Kim. Robust regression methods for computer vision: A review. *The International Journal of Computer Vision*, 6(1):59–70, 1991.
32. M. J. D. Powell. *The convergence of variable metric methods for nonlinearly constrained optimization calculations*, chapter 4. Nonlinear Programming. Academic press, New York, 1978.
33. W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C*. Cambridge University Press, 1988.
34. W. J. Rey. *Introduction to Robust and Quasi-Robust Statistical Methods*. Springer, Berlin, Heidelberg, 1983.
35. P. Rousseeuw and A. Leroy. *Robust Regression and Outlier Detection*. John Wiley & Sons, New York, 1987.
36. H. Schwarz. *Numerical Analysis*. Wiley and Sons, New-York, 1989.
37. C. Stewart. Bias in robust estimation caused by discontinuities and multiple structures. *pami*, 19(8), 1997.
38. P. H. S. Torr. Geometric motion segmentation and model selection. *Phil. Trans. R. Soc. Lond. A*, 356:1321–1340, 1998.
39. P. H. S. Torr and D. W. Murray. The development and comparison of robust methods for estimating the fundamental matrix. *IJCV*, 24(3):271–300, 1997.
40. B. Triggs. Optimal estimation of matching constraints. In R. Koch and L. V. Gool, editors, *Workshop on 3D Structure from Multiple Images of Large-scale Environments SMILE'98*, Lecture Notes in Computer Science, 1998.
41. T. Viéville and O. Faugeras. The first order expansion of motion equations in the uncalibrated case. *CVGIP: Image Understanding*, 64(1):128–146, July 1996.
42. T. Viéville, O. D. Faugeras, and Q.-T. Luong. Motion of points and lines in the uncalibrated case. *The International Journal of Computer Vision*, 17(1):7–42, Jan. 1996.
43. T. Viéville and D. Lingrand. Using specific displacements to analyze motion without calibration. *The International Journal of Computer Vision*, 31(1):5–29, 1999.
44. T. Vieville, D. Lingrand, and F. Gaspard. Implementing a variant of the Kanatani's estimation method. RR 4050, INRIA, Nov. 2000.
45. T. Viéville and P. Sander. Using pseudo Kalman-filters in the presence of constraints. Technical Report RR-1669, INRIA, Sophia, France, 1992.
46. T. Viéville, C. Zeller, and L. Robert. Using collineations to compute motion and structure in an uncalibrated image sequence. *The International Journal of Computer Vision*, 20(3):213–242, 1996.
47. G. Wei and S. Ma. Implicit and explicit camera calibration: Theory and experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):469–480, 1994.
48. Z. Zhang. Parameter estimation techniques: A tutorial with application to conic fitting. *Image and Vision Computing Journal*, 15(1):59–76, 1997.
49. Z. Zhang, R. Deriche, O. Faugeras, and Q.-T. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence Journal*, 78(1-2):87–119, 1994. Appeared in October 1995, also INRIA Research Report No.2273, May 1994.

Acknowledgement : We are thankful to **Olivier Faugeras** for some his powerful ideas at the origin of this work.
We are especially thankful to the IJCV reviewers for their exceptional help in improving and clarifying the 1st draft of this paper.