J Math Imaging Vis (2009) 33: 338–346
DOI 10.1007/s10851-008-0116-z

A Probabilistic Method for Point Matching in the Presence of Noise and Degeneracy

Daniel Keren

Published online: 26 September 2008

Abstract The Bayesian method is widely used in image processing and computer vision to solve ill-posed problems. This is commonly achieved by introducing a prior which, together with the data constraints, determines a unique and hopefully stable solution. Choosing a "correct" prior is however a well-known obstacle.

This paper demonstrates that in a certain class of motion estimation problems, the Bayesian technique of integrating out the "nuisance parameters" yields stable solutions even if a flat prior on the motion parameters is used. The advantage of the suggested method is more noticeable when the domain points approach a degenerate configuration, and/or when the noise is relatively large with respect to the size of the point configuration.

Keywords Motion estimation · Bayesian analysis · Nuisance parameters

1 Introduction

Numerous restoration and parameter estimation algorithms overcome ill-posed problems by assuming a *prior probability* over the space of admissible solutions. It is well-known in the computer vision community that a suitable prior—for example, one which assigns higher probability to smoother surfaces or slowly changing optical flow fields—helps to "stabilize" the problem and often yields a unique solution. But it is also well-known that there is no perfect prior. Instability also hinders feature-based motion recovery, even in relatively simple cases. The problem is quite simple to define: given noisy measurements of "domain points" $\{p_i\}$ and "range points" $\{q_i\}$, one seeks a motion (or equivalently a transformation) T from a certain parametric family such that on the average, $T(p_i)$ is close to q_i (there exist more sophisticated methods, to be discussed shortly). This problem is highly ill-posed in some cases, and it is desirable to stabilize the solutions. That may be achieved by choosing a prior which, for example, penalizes large motions—alas, that will throw us back to the problem of choosing a "good" prior.

The goal of this paper is to show how some simple motion estimation problems can be stabilized not by a prior, but by integrating out "nuisance parameters" in the Bayesian spirit. 1–3 below informally summarize the main features of the suggested method:

1. The probability densities of the parameters are obtainable in closed form, but some of the expressions are involved.
2. The suggested approach yields a distribution with a unique maximum likelihood (ML hereafter) estimate in some cases in which there exist an infinite number of solutions, all of which are equivalent under the so-called "algebraic" and "geometric" motion recovery methods.
3. A transformation is assigned a high likelihood when it not only brings the domain points close to the corresponding range points, but when it has a large "support", in the sense that it *brings a large measure of points which are close to the domain points, to points which are close to the range points.* If the range points are in a degenerate configuration, the ML estimate is also degenerate, and it collapses the entire domain space to a linear affine variety (subspace or translated subspace) spanned by the range points. As the points approach degeneracy, the ML

D. Keren (✉)
Computer Science Department, Haifa University, Haifa 31905, Israel
e-mail: dkeren@cs.haifa.ac.il

estimate does not oscillate, but also smoothly approaches degeneracy.

1.1 The Structure of the Paper

The main goal is to calculate the probability density $f(T|p,q)$ for a transformation T, where p resp. q is the noisy measurement of a domain resp. range point. If this problem is solved, and if measurement noise is independent for distinct domain-range pairs, then the density $f(T|\{p_i,q_i\}_{i=1}^n)$ can be calculated by multiplying the densities for the individual pairs. It turns out that in some cases the calculation for a single domain-range pair cannot be carried out directly, but that can be overcome by first calculating the density when the number of degrees of freedom (d.o.f.) of the domain points is equal to the d.o.f. of the transformation, and then using this result to calculate for a single pair. Thus for example in Sect. 3 the density of a, b for the transformation $x \to ax + b$ is calculated for two domain-range pairs, and this result is used in Sect. 4 to calculate the density for a single pair.

Linear and affine transformations from the line to itself are discussed in Sects. 2, 3, 4. In Sect. 5 linear transformations from the plane to itself are analyzed, and concluding remarks are offered in Sect. 6. Appendix carries some technical points related to Sect. 2.

1.2 Related Previous Work on Motion Estimation

Much of the previous work concerning motion estimation given matching features concentrated on what are often called the "algebraic method" and "geometric method" (e.g. [7], pp. 76–78). Given domain $\{p_i\}$ and range $\{q_i\}$ points, and an admissible family of transformation T, the algebraic method searches for $T \in \mathsf{T}$ which minimizes $\sum_i \|T(p_i) - q_i\|^2$. The geometric method—which is usually superior to the algebraic—seeks two "legal" point sets, $\{\bar{p}_i\}$ and $\{\bar{q}_i\}$, such that there exists a $T \in \mathsf{T}$ with $T(\bar{p}_i) = \bar{q}_i$ (i.e. $\{\bar{p}_i\}$ and $\{\bar{q}_i\}$ are "legal" with respect to T) and $\sum_i(\|\bar{p}_i - p_i\|^2 + \|\bar{q}_i - q_i\|^2)$ is minimal. A rigorous study of these problems and others in 2D and 3D is given in [3, 7, 10, 11, 14, 19, 20], which also describe and reference a great deal of previous work. Whenever the algebraic and geometric methods yield identical results—as in the case when the d.o.f. of the domain points equals the transformation d.o.f.—I'll refer to them as "classical methods". Torr [19], who addresses 3D motion, offers an illuminating treatment of Bayesian recovery of geometry and motion, and also discusses the possibility of integrating out the nuisance parameters which is pursued in this paper, but he eventually assumes "local linearity" of a certain manifold (which is directly related to the motion parameters), and does not carry out the explicit integration. Most other Bayesian studies on motion estimation have not touched on the nuisance parameters. Some notable exceptions are [15, 17, 18]. In [16], integration over the nuisance parameters is applied to curve fitting.

The method presented here (at least for the problems studied) offers a stable solution in the case of degeneracy or near degeneracy (Sects. 4, 5). It is different from previous work on degeneracy in that it does not pursue any model selection [19, 21, 22, 24], but "automatically" recovers a degenerate/near-degenerate transformation when appropriate.

The papers [6, 8, 9, 13] share some common grounds with this paper—mainly the idea of integrating out the nuisance parameters—but they proceed differently. Also, they address only straight line fitting.

1.3 Previous Related Work in which I Cooperated

Some of the ideas in this work were presented in the Dagstuhl workshop on "Theoretical Foundations of Computer Vision: Geometry, Morphology, and Computational Imaging", 2002 [12]. Integrating out the nuisance parameters was applied to the recovery of epipolar geometry in work which appeared in the 2003 International Conference on Computer Vision [4] (of which an extended version appeared in [5]) and to shape fitting in [23]. Some related work was also presented in the workshop on Bayesian inference and maximum entropy methods, Garching 2004.

2 A Toy(?) Problem: Linear Transformation from $\Re$ to Itself

Recovery of motion may typically be viewed as the recovery of a transformation between two real vector spaces. Start with the simplest case—a linear transformation from the line to itself. This is a surprisingly subtle problem. The advantages of tackling it are twofold: first, it captures the method's spirit. Second, recovering one dimensional transformations between sets of epipolar lines can assist in the recovery of 3D motion (e.g. [2], p. 219). Although these transformations are non-linear, they can be approached using the method presented in Sects. 2–4.

Assume then that a linear transformation from the line to itself, $x \to ax$, is given, as well as noisy measurements of a domain point p and range point q. As [6, 13] insightfully point out, the problem of estimating a is *different* than estimating a line which passes through (or, in the general case, close to) the measurement points. Clearly, both the algebraic and geometric methods will yield q/p as the optimal estimate of a. However, it is easy to devise examples in which this estimate is not a good one. Assume for example that $p = 0.001, q = 0.1$, and the measurement noise is

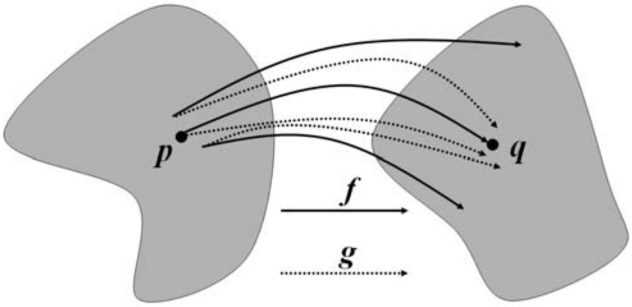

Fig. 1 If noisy estimates of a domain point p and range point q are given, the classical approach will rank the transformation f as better than g, since f maps p to q, while g does not. However, if the noise variance is large with respect to the size of the point configuration, or when the domain points approach a degenerate configuration (see Sect. 5), the paradigm described here will rank g as better than f, since it has a wider "support"—i.e. it moves a larger measure of points which are close to p to locations close to q

$\sigma = 1$. The resulting estimate is $a = 100$, and obviously it is not satisfactory; for example, it is extremely sensitive to changes in p. Intuition tells us that, since the noise is so large relative to the size of the points, a much smaller a should probably be chosen (the so-called "shrinkage", see [6, 13] and references therein). Following the regularization paradigm, we can impose a prior which penalizes large a's. But that annoying question rears its head again: what prior to use?

The approach suggested in this paper tries to minimize the effect of the prior defined on the space of transformations. Instead, it integrates out the "nuisance parameters" corresponding to the "real" domain and range points. Before going into the formalities, some intuition. In the previous example—$p = 0.001$, $q = 0.1$—why is $a = 100$ such a poor estimate? After all, it does bring p exactly to q! The problem is, of course, with the noise; we don't know what the "real" domain and range points are: we only have a distribution on their locations—in a "probability cloud" around p and q respectively. We should therefore find a transformation which brings a large measure of the "p cloud" into the "q cloud". But multiplying by 100 moves most of the "p cloud" way too far from the center of the "q cloud". This is in essence the well known overfitting problem: if the fitted model (in this case, the single parameter a) tries too hard to accommodate the noisy data, it may well give bad results away from the data. A schematic description of this idea is provided in Fig. 1.

The framework presented in this paper proposes to overcome the overfitting problem not by restraining the fit via the application of a prior, but by integrating out the nuisance parameters, i.e. the "real" domain and range points. Denoting these nuisance parameters by $\hat{p}$, $\hat{q}$ respectively, the joint probability density of a, $\hat{p}$ is

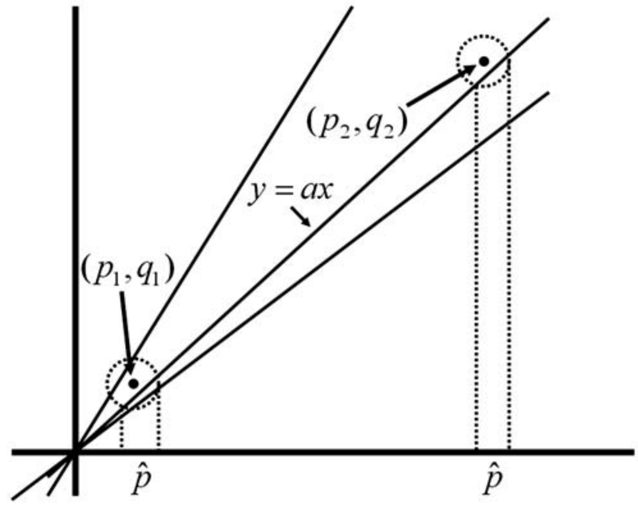

Fig. 2 The measurement (p_2, q_2) carries a "concentrated" support for the parameter *a,* while (p_1, q_1) supports a large interval of slope values

$$f(a, \hat{p} \mid p, q) = \left| \frac{\partial(\hat{p}, a\hat{q})}{\partial(a, \hat{p})} \right| f(\hat{p}, a\hat{p} \mid p, q)$$
$$= |\hat{p}| \frac{f(p, q \mid \hat{p}, a\hat{p}) f(\hat{p}, a\hat{p})}{f(p, q)},$$

where Bayes' rule was invoked in the second equality. At this stage assume a "flat prior" on $\hat{p}, a\hat{p} = \hat{q}$.[1]

(The denominator $f(p, q)$ can be assumed to be a constant, as the measurements were taken already.) Therefore, assuming Gaussian noise with variance σ^2, the density is proportional to $|\hat{p}| \exp(-\frac{(p-\hat{p})^2 + (a\hat{p}-q)^2}{2\sigma^2})$, and (up to a normalizing factor), the density of a is obtained by integrating out $\hat{p}$:

$$f(a|p, q) = \int_{-\infty}^{\infty} |\hat{p}| \exp\left(-\frac{(p - \hat{p}) + (a\hat{p} - q)^2}{2\sigma^2}\right) d\hat{p} \tag{1}$$

A very short discussion of this integral is deferred to Appendix. Minka [13] also suggests to integrate out the nuisance parameter, but his approach—and results—are different. Note that no range nuisance parameter has to be integrated out—it is determined by a and the domain nuisance parameter.

When is the probability density of a large? For the integral in (1) to obtain a large value, there should be a large measure of points $\hat{p}$ satisfying: (1) $\hat{p}$ is close to p (where

[1] This assumption is tricky in that it indirectly imposes a prior on a. For example, a prior on $\hat{p}, \hat{q}$ which is uniform in a circle centered at the origin (of any radius) and zero outside it, induces the prior $1/\pi(1 + a^2)$. Such priors—which may be dictated by the real world constraints of the problem—should in general be considered, but the thrust here is on how the integration over the nuisance parameters affects $f(a)$.

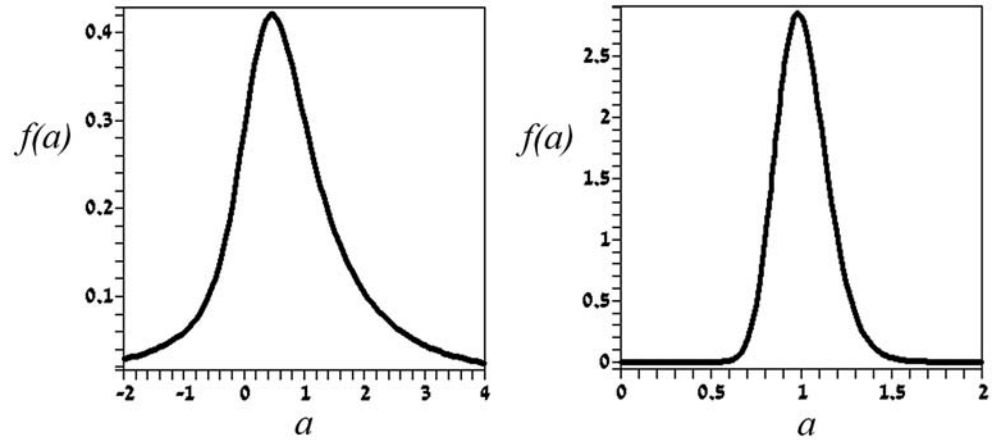

Fig. 3 Probability density plots for $a, p = q = \sigma = 1$ (*left*) and $p = q = 10, \sigma = 1$ (*right*). The optimal value for a in the classical methods is of course 1. In the example on the *left*, where $|p|$ and σ are of the same order of magnitude, the distribution is centered around the ML value of 0.463. As the noise becomes smaller with respect to $|p|$, the distribution approaches the classical one: the distribution on the *right* is highly peaked at 0.98, which is close to the classical estimate of 1. Note that in both cases, the ML transformation a_{ML} avoids the overfitting problem: $a_{ML}p$ is not equal to q but smaller than it

"close" is with respect to the magnitude of the noise), (2) $a\hat{p}$ is close to q, (3) $|\hat{p}|$ is large. The intuition behind conditions (1) and (2) is clear—if a is to be trusted, then there should be a large measure of $(\hat{p}, \hat{q})$ pairs which are close to (p, q) and such that $\frac{\hat{q}}{\hat{p}} = a$. What about (3)? There is a simple intuitive explanation for this condition: if $|\hat{p}|$ is small for a large measure of $\hat{p}$ which are close to p, there will also be a large measure of pairs close to (p, q) "supporting" values a' which are quite different from a—this because for noise n_i, $\frac{y+n_1}{x+n_2}$ is, on the average, closer to $\frac{y}{x}$ when $|x|$ is larger. Thus $|\hat{p}|$ can be viewed as a "trust factor" assigned to the integration variable $\hat{p}$, which measures how strongly it supports the "naïve" estimate $\frac{\hat{q}}{\hat{p}}$. Such "trust factors" will appear in all the cases addressed in this paper. They have the effect of further reducing the likelihood of "large" transformations: in the case of $x \to ax$, for example, large values of a are "supported" by small $\hat{p}$, but these small $\hat{p}$ are ascribed less importance in the integrand, due to the $|\hat{p}|$ factor. This is just intuition—formally, the "trust factors" are there because of the appropriate Jacobians.

2.1 Discussion and Some Examples

There are a few ways to proceed, once the density of a is given. One can for example follow the ML paradigm (choose the most likely a), or average over all a. I do not touch on these questions here (although the ML estimate will often be mentioned, it being an indicator of the distribution's properties). Some examples are provided, to give the feeling of what the density looks like. Further questions—such as the nature of the density for one and more domain-range pairs—(is it unimodal?) are to be further investigated.

3 Affine Transformation from $\Re$ to Itself: Two Pairs

An affine transformation from the line to itself is defined by $x \to ax + b$.Given two domain/range pairs, (p_1, q_1) and (p_2, q_2), classical methods yield the solution

$$a = \frac{q_1 - q_2}{p_1 - p_2}, \qquad b = \frac{p_1 q_2 - p_2 q_1}{p_1 - p_2}.$$

Clearly this solution will become more unstable as the noise grows larger relative to $|p_1 - p_2|$.

The approach used in Sect. 2 can be immediately extended to this case, yielding the following expression for the probability density of a, b, with $\hat{p}_1$, $\hat{p}_2$ standing for the nuisance parameters corresponding to the "real" domain points:

$$f(a, b|p_1, q_1, p_2, q_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |\hat{p}_1 - \hat{p}_2| \times \exp\left(-\frac{(p_1-\hat{p}_1)^2+(p_2-\hat{p}_2)^2+(ap_1-\hat{q}_1)^2+(ap_2-\hat{q}_2)^2}{2\sigma^2}\right) d\hat{p}_1 d\hat{p}_2 \quad (2)$$

Where the source of the $|p_1 - p_2|$ factor is the Jacobian of the transformation. The intuition behind this equation is similar to that behind (1). Note that higher weights are assigned to "stable" configurations, i.e. those in which $\hat{p}_1, \hat{p}_2$ are further apart—in accordance with the observation that such pairs yield a more accurate estimate, as the denominator in the classical solution for a, b is more resistant to noise. Thus $|p_1 - p_2|$ is a "trust factor" equivalent to $|\hat{p}|$ in (1).

The explicit form of the integral in (2) is rather cumbersome, and will not be included here. Like the p.d.f. in (1), it can be normalized to sum to 1, but the details of the normalization will be left out hereafter. Two examples of the p.d.f. (see Fig. 4).

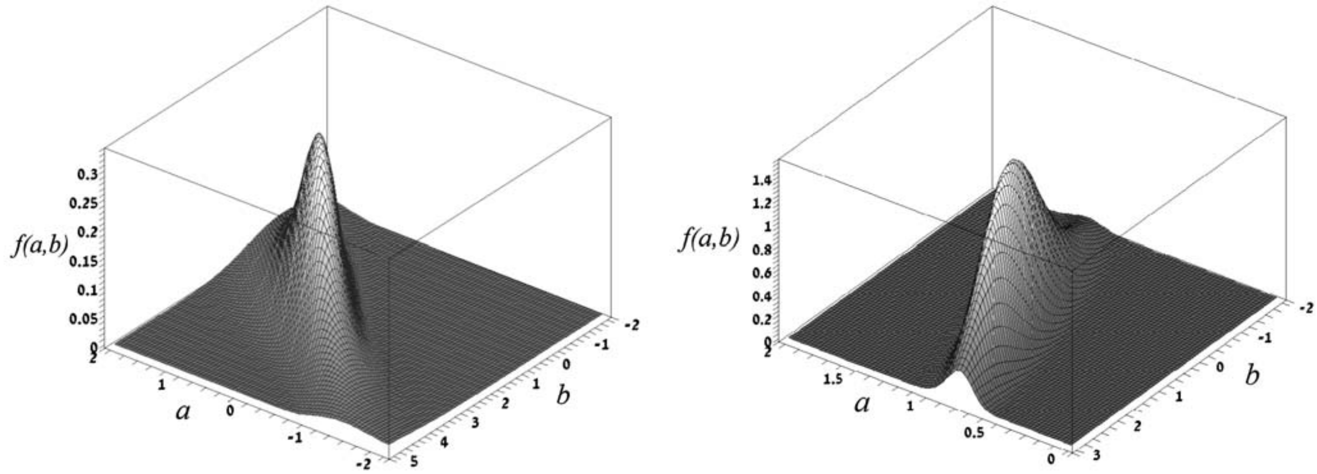

Fig. 4 *Left*: probability density plot for a, b when $p_1 = 1, q_1 = 2$, $p_2 = 2$, $q_2 = 3$, $\sigma = 1$. The ML is obtained at $a = 0.37$, $b = 1.95$, while the classical methods choose $a = b = 1$ as optimal. *Right*: probability density plot for a, b when $p_1 = 1, q_1 = 2$, $p_2 = 10$, $q_2 = 11$, $\sigma = 1$. The ML is obtained at $a = 0.96$, $b = 1.19$. When the magnitude of the difference between p_1 and p_2 grows with respect to the noise, the ML estimate converges to the classical solution $a = b = 1$

4 Affine Transformation from $\Re$ to Itself: One Pair

How to recover $x \to ax + b$ when only one domain-range pair (p_1, q_1) is given? In the classical approaches, there are an infinite number of solutions, all equally good, which correspond to the equations of the straight lines going through (p_1, q_1). What happens when one tries to integrate out the nuisance parameters in order to evaluate $f(a, b|p_1, q_1)$? It would be nice to proceed as in Sect. 2, where $f(a|p, q)$ was computed by integrating over $f(a, \hat{p}|p, q)$, which was computed by Bayes' theorem and by computing $f(a\hat{p}, \hat{p}|p, q)$. Alas here it is impossible: $f(a, b, \hat{p}_1|p_1, q_1)$ cannot be reduced to $f(a\hat{p}_1 + b, \hat{p}_1|p_1, q_1)$, because there is no Jacobian to transform from three variables to two (the opposite is possible, by using the square root of $|JJ^t|$, where J is a 2×3 Jacobian). One way to overcome this problem is to add further nuisance parameters (p_2, q_2), and reduce the problem to that of Sect. 3. A "flat" prior on (p_2, q_2) is defined by the limit $c \to 0$ of $\frac{c}{\pi} \exp[-c(p_2^2 + q_2^2)]$. So, up to a constant factor, the probability density $f(a, b|p_1, q_1)$ equals the limit at $c \to 0$ of

$$c \iiiint \exp[-c(p_2^2 + q_2^2)]|\hat{p}_1 - \hat{p}_2| \times \exp\left(-\frac{(p_1-\hat{p}_1)^2+(p_2-\hat{p}_2)^2+(a\hat{p}_1-q_1)^2+(a\hat{p}_2-q_2)^2}{2\sigma^2}\right) d\hat{p}_1 d\hat{p}_2 dp_2 dq_2$$

where the integral is over $\Re^4$. Up to a multiplicative factor depending on π and σ, this integral equals

$$\frac{\exp[-\frac{(ap_1+b-q_1)^2}{s\sigma^2(a^2+1)}]}{(a^2+1)^{\frac{3}{2}}} \tag{3}$$

It is easy to verify that (3) obtains its maximum when $a = 0$, $b - q_1$: the power in the exponent, which is always ≤ 0, can be zero only when $ap_1 + b = q_1$, and among all such combinations, the one with $a = 0$ minimizes the denominator. So the numerator obtains its maximum and the denominator its minimum when $a = 0$, $b - q_1$.

Is this a surprise? Hardly: as before, a higher probability density is assigned to "small" and "stable" transformations. Clearly, the constant transformation $x \to q_1$ is the "most stable" which takes p_1 to q_1. It happens also to be degenerate—is that bad? I believe not; there is no a-priori preference among the transformations, and there is therefore no real reason for the ML estimate to be non-degenerate. Is (3) important? Hopefully so, since one is not only interested in the ML estimate, but in the entire distribution over a and b. Suppose that a few domain-range pairs are present—one may then consider multiplying the corresponding exemplars of (3), which would yield a more interesting distribution and ML estimate.

What about the simple (classical) solution which, given (p_1, q_1), assigns to a, b the density $\frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(ap_1+b-q_1)^2}{2\sigma^2})$? It does not look good, as it does not penalize transformations with a large a: as long as $ap_1 + b = q_1$, the transformations represented by a, b have an equal probability density (which means that the distribution cannot be directly normalized). However, adding the nuisance parameters (p_2, q_2) radically changes the density. The calculation is quite similar to the one carried in the beginning of this section: calculate the density when (p_2, q_2) are added, integrate out (p_2, q_2), and take the limit when the prior on them approaches a flat prior.

Up to a multiplicative constant, the result is

$$\frac{\exp[-\frac{(ap_1+b-q_1)^2}{2\sigma^2}]}{(a^2+1)^{\frac{1}{2}}} \tag{4}$$

4.1 Interpretation of Results as Priors on a

Although a flat prior was assumed on a, one may try and view the results as a combination of classical fitting (e.g. trying to minimize $\sum_{i=1}^{n}(ap_i + b - q_i)^2$) and a prior on a. For example, (4) can be viewed as a prior that penalizes large values of a: the logarithm of the density is

$$-\left[\frac{n}{2}\log(1+a^2) + \frac{\sum_{i=1}^{n}(ap_i+b-q_i)^2}{2\sigma^2}\right]$$

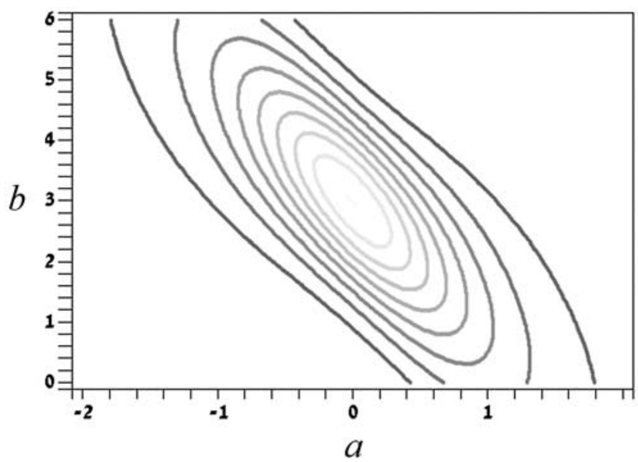

Fig. 5 Contour plot for probability density in (3). There is a unique peak at $a = 0$, $b = 3$. The contours depict equi-density (*brighter* is *higher*)

In addition to the penalty defined by the "data term" $-\frac{\sum_{i=1}^{n}(ap_i+b-q_i)^2}{2\sigma^2}$, there is a $-\frac{n}{2}\log(1 + a^2)$ summand which penalizes large a's. This summand somewhat resembles Jaynes' prior [8] $-\frac{3n}{2}\log(1 + a^2)$, which was derived using a transformation group argument. Gull [6] suggests that the rotational symmetry used by Jaynes is not relevant to the line fitting problem, and derives a different, more complicated prior—which includes a $-\frac{3n}{2}\log(|a|)$ term. The prior induced by (4) is different from both Gull's and Jaynes'.

Lastly, the prior induced by (3) has a different shape from Gull's and Jaynes' prior, as well as the prior induced by (4): the density's logarithm is

$$-\left[\frac{3n}{2}\log(1+a^2) + \frac{\sum_{i=1}^{n}(ap_i+b-q_i)^2}{2\sigma^2(1+a^2)}\right]$$

As opposed to all the other cases herewith, a appears not only in a summand which is disjoint from the "data term", but is also mixed with the data term.

4.2 Examples

Plots of the three densities for $p_1 = 2$, $q_1 = 3$, $\sigma = 1$ (see Figs. 5, 6).

Remark One may be misled to assume that the ML transformations are always "trivial". That is not the case. Given more points, in a configuration which is far from degeneracy, and if the noise is small relative to the total size of the configuration, the distribution is "reasonable" in the sense that its ML estimate converges to the ML estimate of the classical methods.

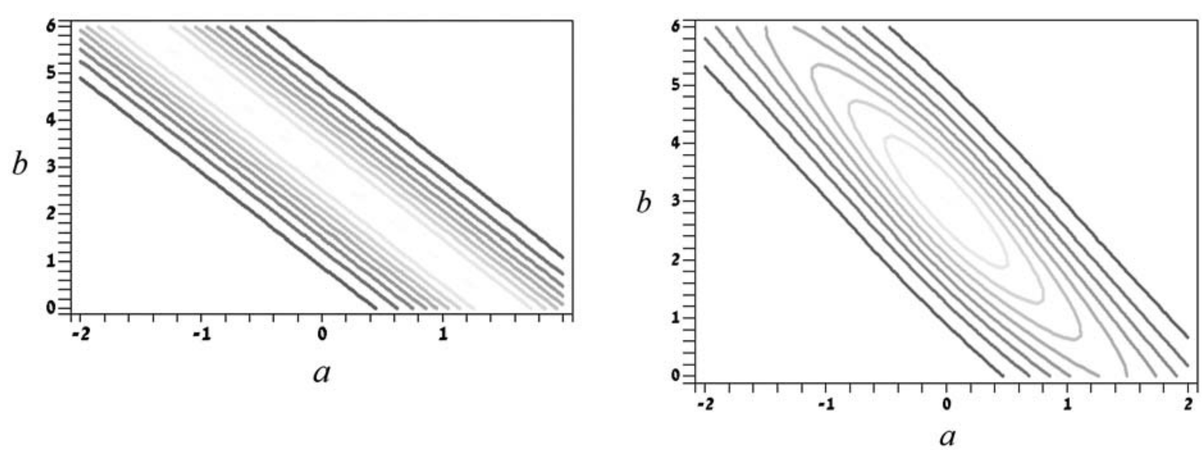

Fig. 6 *Left*: probability density plot for a, b, classical methods (density given by $\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(ap_1+b-q_1)^2}{2\sigma^2})$). The density is peaked along the line satisfying $2a + b = 3$. Right: contour plot for probability density in (4). There is a unique peak at $a = 0$, $b = 3$, but the overall shape of the distribution is different from that in Fig. 5

5 Linear Transformation from $\Re^2$ to Itself: Two Pairs

Computing the probability density for a linear transformation $T : \Re^2 \to \Re^2$ proceeds very much like the previous cases. Two domain pairs have a total of four d.o.f., hence they suffice to calculate the density in a straightforward fashion. Assume that noisy measurements of two domain-range pairs are given: (p_1, q_1), (p_2, q_2), and denote the nuisance parameters by $\hat{p}_i = (\hat{p}_i^x, \hat{p}_i^y)$, $i = 1, 2$. The density is

$$f(T|p_1, p_2, q_1, q_2) = \iiiint \underbrace{(\hat{p}_1^x \hat{p}_2^y - \hat{p}_1^y \hat{p}_2^x)^2}_{\text{“trust factor”}} \times \exp\left[-\frac{\|\hat{p}_1 - p_1\|^2 + \|\hat{p}_2 - p_2\|^2 + \|T(\hat{p}_1) - q_1\|^2 + \|T(\hat{p}_2) - q_2\|^2}{2\sigma^2}\right] d\hat{p}_1^x d\hat{p}_1^y d\hat{p}_2^x d\hat{p}_2^y \quad (5)$$

where the integration is carried over $\Re^4$. The integral can be computed in closed-form. As for the 1D cases, T has a high density if it is supported by a large measure in domain-range space of points which are close to the measured domain-range pairs, but this support is again weighted by a certain “trust factor”, which is an objective quality of the domain points. Here the “trust factor” reflects the fact that in order to obtain a reliable estimate to a linear transformation (in the plane) under noise, we should have two domain points which are not only far from the origin but also far from being linearly dependent. The “trust factor” in (5) is the determinant squared of the two domain points, and for it to be large, the points have to be in a stably non-degenerate position, i.e. far from the origin and also far from being collinear with the origin.

5.1 Examples

The simplest example which emphasizes the difference between the method presented here and the classical methods is the following: let the measurements be given by

$$p_1 = (\alpha, \alpha), \qquad p_2 = (2\alpha, 2\alpha),$$
$$q_1 = (\alpha, \alpha), \qquad q_2 = (2\alpha, 2\alpha)$$

where α is a real number. This example may appear artificial and limited, but the results extend to cases in which the domain points approach degeneracy ([12] and Sect. 5.2 here).

The classical methods have an infinite variety of equally good transformations to choose from—$\left(\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right)$ is fine, but so is $\left(\begin{smallmatrix}101 & -100\\ -200 & 201\end{smallmatrix}\right)$, and they have the same probability density. On the other hand, a direct computation (omitted here) proves that the suggested method has a unique ML estimate, which equals $\left(\begin{smallmatrix}\delta & \delta\\ \delta & \delta\end{smallmatrix}\right)$, where δ is a positive number which approaches $\frac{1}{2}$ as $\frac{\sigma}{\alpha} \to 0$, and σ is the noise variance. This is a degenerate matrix, which takes the entire plane to the line $x = y$. Is this “wrong”? No: as for the one-dimensional case, there is no a-priori assumption on the transformations, and the chosen degenerate transformation performs well, in that it moves a large measure of points which are close to the domain point (α, α), to the vicinity of the range point (α, α)—and same for $(2\alpha, 2\alpha)$. If σ is relatively large with respect to α, the “shrinkage effect” results in smaller values of δ (compare with Fig. 1 and discussion therein).

5.2 Behavior for Nearly Degenerate Configurations

To make matters a little more interesting, let us perturb the points a little. Take $\alpha = 1$. The domain and range points are then

$$p_1 = (1 + n_1, 1 + n_2), \qquad p_2 = (2 + n_3, 2 + n_4),$$
$$q_1 = (1 + n_5, 1 + n_6), \qquad q_2 = (2 + n_7, 2 + n_8) \quad (6)$$

where each n_i is taken to be Gaussian noise with a standard deviation 0.1. Denote $T = \left(\begin{smallmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{smallmatrix}\right)$. T can be solved for explicitly in the classical paradigm, since it then has to

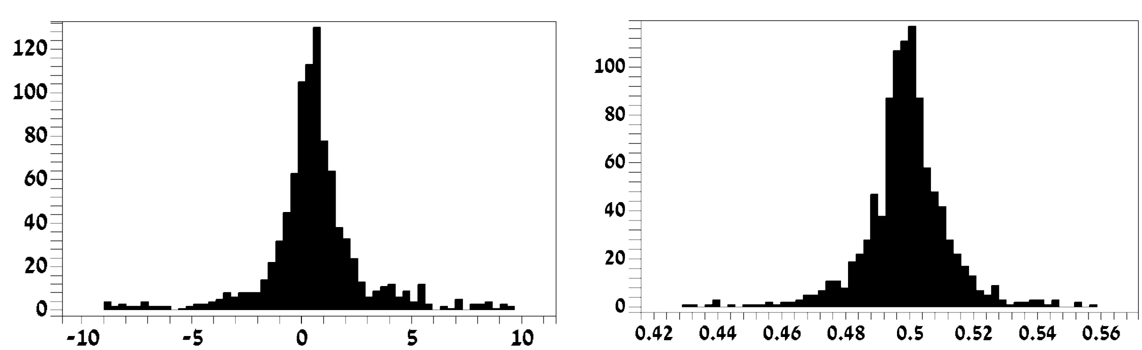

Fig. 7 Histogram of a_{11} values for the input in (6), with the noise variance equal to 0.01. *Left*—classical method, 1,000 inputs. 49 values with absolute values >10 were discarded. *Right*—suggested method, same inputs. The suggested method is far more stable

satisfy $T(p_1) = q_1, T(p_2) = q_2$. The solution for a_{11}, for example, is

$$\frac{n_4 + 2n_5 + n_4 n_5 - 2n_2 - n_7 - n_2 n_7}{n_4 + 2n_1 + n_1 n_4 - 2n_2 - n_3 - n_2 n_3}$$

Clearly, this is a highly sensitive estimate, which oscillates wildly as the noise changes. On the other hand, the suggested estimate is very stable under such perturbations: histograms for the ML value of a_{11} for the two methods are depicted in Fig. 7. As noted before, the ML value is presented since it is a good indicator of the distribution's shape. In the classical methods, it is very unstable, which reflects an undesirable sensitivity to the input, while in the suggested method it is very stable.

Experiments carried out in [12] have demonstrated that when the domain configuration is close to a degenerate one, then this behavior of the classical method, and the suggested method as well, persist when the number of domain-range pairs increases.

5.3 Linear Transformation from $\Re^2$ to Itself: One Pair

Given only one domain-range pair (p_1, q_1), we can proceed exactly as in Sect. 4: assume a prior on (p_2, q_2) which approaches a flat prior in the limit, and integrate (5) over (p_2, q_2). The result is the following octal integral:

$$f(T|p_1, q_1) = \lim_{c \to 0} c^2 \int\!\!\int\!\!\int\!\!\int\!\!\int\!\!\int\!\!\int\!\!\int \exp(-c(\|p_2\|^2 + \|q_2\|^2))(\hat{p}_1^x \hat{p}_2^y - \hat{p}_1^y \hat{p}_2^x)^2 \times \exp\left[-\frac{\|\hat{p}_1 - p_1\|^2 + \|\hat{p}_2 - p_2\|^2 + \|T(\hat{p}_1) - q_1\|^2 + \|T(\hat{p}_2) - q_2\|^2}{2\sigma^2}\right] d\hat{p}_1^x d\hat{p}_1^y d\hat{p}_2^x d\hat{p}_2^y dq_2^x dq_2^y dp_2^x dp_2^y$$

This integral can be computed in closed-form expression, which is not especially illuminating, and is not included in this paper. Then, the density for any number of domain-range points can be computed by multiplying the expressions for the individual pairs.

6 Summary and Future Work

- The Bayesian notion of integrating out nuisance parameters was applied to the recovery of linear and affine transformations between low-dimensional vector spaces. The nuisance parameters vary over all the possible domain and range points. The method assumes a flat prior on the space of transformations.
- The method results in closed-form expressions which are non-trivial, and I could not find a simple method to find the ML estimate, for example (save for some simple cases).
- The method returns results markedly different from the classical methods when:

  1. The magnitude of the domain points is not large with respect to the noise.
  2. The range points approach a degenerate configuration.

  In these cases, the classical methods strongly overfit the noise, and are also highly unstable—a very small perturbation in the input can radically change the results. The results of the suggested approach were better.
- The suggested method generally assigns a higher probability density to transformations with small coefficients in the linear terms (for the 1D case) and to matrices with small eigenvalues (in the 2D case)—the "shrinkage" effect.

6.1 Future Work

Many open questions remain:

- How to use the resulting density (i.e. if one transformation should be chosen, is the ML, or the expectation, a good choice)? Related question—when is the density unimodal?
- How should the performance of transformation recovery methods be evaluated? A very common method is: choose a transformation T_0 and a set of domain-range points (p_i, q_i) such that $T_0(p_i) = q_i$, add noise to (p_i, q_i) to obtain $(\tilde{p}_i, \tilde{q}_i)$, compute the T-density from $(\tilde{p}_i, \tilde{q}_i)$, and check whether T_0 is a good representative of this density. Alas it is clear that if for example the domain points approach a degenerate configuration, this is *not* a good method. The analysis presented in this paper suggests that perhaps the "trust factor" should be used to weigh the transformation corresponding to (p_i, q_i) when conducting such an analysis.
- How can the method presented here be extended to more general transformations (e.g. projective)? A first attempt at this direction was suggested in [4], but there's a lot more to be done.

- Higher dimensions: the idea presented here can be extended to affine transformations in higher dimensions. Since the corresponding integrals will still be Gaussian, they have closed-form solutions.
- Outliers: these can be handled in a standard way, by incorporating a RANSAC technique [1] with the method suggested here. It remains to see how this will work in practice.

Acknowledgements I thank Padmanabhan Anandan, Alexander Goldenshluger, Liran Goshen, Ilan Shimshoni, and Michael Werman for many interesting discussions.

Appendix: Equation (1)

The integral in (1)

$$\int_{-\infty}^{\infty} |\hat{p}| \exp\left(-\frac{(p-\hat{p})^2 + (a\hat{p}-q)^2}{2\sigma^2}\right) d\hat{p}$$

is of the general form

$$\int_{-\infty}^{\infty} |x| \exp[-(Ax^2 + Bx + C)] dx$$

where x substitutes for $\hat{p}$ and the exponent is replaced by a general quadratic (with $A > 0$). The integral equals

$$\frac{\left(2\sqrt{A} + B\sqrt{\pi}\exp\left(\frac{B^2}{4A}\right)\mathrm{erf}\left(\frac{B}{2\sqrt{A}}\right)\right)\exp(-C)}{2A^{3/2}}$$

and the original integral can be recovered by replacing A, B, C with the appropriate expressions in a, p, q, σ.

The integral is easily normalizable. It can be proved using a direct calculation that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\hat{p}| \exp\left(-\frac{(p-\hat{p})^2 + (a\hat{p}-q)^2}{2\sigma^2}\right) da\, d\hat{p} = 2\pi\sigma^2$$

so it is trivial to normalize it so as to define a proper distribution on a (i.e. which sums to 1).

References

1. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (1981)
2. Forsyth, D.A., Ponce, J.: Computer Vision: A Modern Approach. Prentice Hall, New York (2002)
3. Georgescu, B., Meer, P.: Balanced recovery of 3d structure and camera motion from uncalibrated image sequences. 7th Eur. Conf. Comput. Vis. **II**, 294–308 (2002)
4. Goshen, L., Shimshoni, I., Anandan, P., Keren, D.: Recovery of epipolar geometry as a manifold fitting problem. In: Proceedings of the 9th International Conference on Computer Vision, pp. 1321–1328
5. Goshen, L., Shimshoni, I., Anandan, P., Keren, D.: Motion recovery by integrating over the joint image manifold. Int. J. Comput. Vis. **65**(3), 131–145 (2005)
6. Gull, S.F.: Bayesian data analysis: straight-line fitting. In: Maximum Entropy and Bayesian Methods, pp. 511–518 (1989)
7. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge (1998)
8. Jaynes, E.T.: In: Rosenkrantz, R. (ed.) Papers on Probability, Statistics, and Statistical Physics, pp. 190–209. Reidel, Dordrecht (1983)
9. Jaynes, E.T.: Straight line fitting—a Bayesian solution. Unpublished manuscript, item 22 in http://bayes.wustl.edu/etj/node2.html
10. Kanatani, K.: Geometric Computation for Machine Vision. Clarendon Press, Oxford (1993)
11. Kanatani, K.: Statistical analysis of geometric computation. Comput. Vis. Graph. Image Process. **59**(3), 286–306 (1994)
12. Keren, D., Shimshoni, I., Goshen, L., Werman, M.: All points considered a maximum-likelihood method for motion estimation. In: Dagstuhl Workshop on Geometry, Morphology, and Computational Imaging, April (2002). Also in LNCS, vol. 2616, pp. 72–85 (2003)
13. Minka, T.: Linear regression with errors in both variables: a proper Bayesian approach. Unpublished manuscript, http://www.stat.cmu.edu/~minka/papers/eiv.html
14. Matei, B., Meer, P.: Optimal rigid motion estimation and performance evaluation with bootstrap. IEEE Conf. Comput. Vis. Pattern Recognit. **I**, 339–345 (1999)
15. Nestares, O., Fleet, D.J., Heeger, D.J.: Likelihood functions and confidence bounds for total-least-squares problems. IEEE Conf. Comput. Vis. Pattern Recognit. **I**, 523–530 (2000)
16. Newsam, G.N., Redding, N.J.: Fitting the most probable curve to noisy observations. Proc. Int. Conf. Image Process. **II**, 752–755 (1997)
17. Ohta, N.: Motion parameter estimation from optical flow without nuisance parameters. In: 3rd IEEE Workshop on Statistical and Computational Theories of Vision (2003). http://department.stat.ucla.edu/~yuille/meetings/papers/sctv03_24.pdf
18. Okatani, T., Deguchi, K.: Is there room for improving estimation accuracy of the structure and motion problem? In: Suter, D. (ed.) Proceedings of the Statistical Methods in Video Processing Workshop, pp. 25–30 (2002)
19. Torr, P.H.S.: Bayesian model estimation and selection for epipolar geometry and generic manifold fitting. Int. J. Comput. Vis. **50**(1), 27–45 (2002)
20. Torr, P.H.S., Zisserman, A.: Feature based methods for structure and motion estimation. In: Triggs, W., Zisserman, A., Szeliski, R. (eds.) Vision Algorithms: Theory and Practice. LNCS, vol. 1883, pp. 278–295. Springer, Berlin (1999)
21. Torr, P.H.S., Zisserman, A., Maybank, S.J.: Robust detection of degenerate configurations while estimating the fundamental matrix. Comput. Vis. Image Underst. **71**(3), 312–333 (1998)
22. Torr, P.H.S., Zisserman, A., Fitzgibbon, A.W.: The problem of degeneracy in structure and motion recovery from uncalibrated image sequences. Int. J. Comput. Vis. **32**(1), 27–44 (1999)
23. Werman, M., Keren, D.: A Bayesian method for fitting parametric and non-parametric models to noisy data. IEEE Trans. Pattern Anal. Mach. Intell. **23**(5), 528–534 (2001)
24. Zelnik-Manor, L., Irani, M.: Multi-view subspace constraints on homographies. In: IEEE International Conference on Computer Vision (1999)