10.1098/rspa.2002.1096

THE ROYAL SOCIETY

Fisher information and model selection for projective transformations of the line

OK?

By Stephen J. Maybank

Department of Computer Science, University of Reading, Whiteknights,
PO Box 225, Reading, Berkshire RG6 6AY, UK
(s.j.maybank@reading.ac.uk)

Received 7 November 2001; revised 29 July 2002; accepted 12 November 2002

The Fisher information and the Rao measure are obtained in closed form for a family of probability density functions parametrized by the manifold $\mathrm{PSL}(2,\mathbb{R})$ of projective transformations of the real projective line. In addition, the Fisher information and the Rao measure are obtained for the sub-manifold of affine transformations. An application of these results to computer vision is described. The Rao measure is used to obtain a closed form approximation to the probability of misclassifying a projective transformation of the line as an affine transformation. The approximation is a function of the number of pairs of points that correspond under the projective transformation and the standard deviation of the error in locating a point.

Keywords: Fisher information; Rao metric; Rao measure; projective transformation; homography; affine transformation; misclassification probability; computer vision

Author: between three and six keywords/terms needed before this paper can be published: please delete two.

1. Introduction

The Fisher information (Cover & Thomas 1991; Lindsey 1996) is used to define the Rao metric for pairs of probability density functions in a family of densities parametrized by the points of a smooth manifold (Amari 1985; Kotz & Johnson 1992; Rao 1945). The Rao metric has a statistical meaning when a pair of probability density functions have parameters given by points which are close together in the manifold: suppose that independent samples are drawn from one of the densities. Then the second density is also likely to be good candidate for the source of the sample (Balasubramanian 1996, 1997; Myung *et al.* 2000). The Rao metric transforms under reparametrizations of the manifold in such a way that the distance between the two probability density functions is invariant (Jost 1995). This invariance is necessary if the choice of parametrization is not to affect statistical estimation (Fisher 1922; Jeffreys 1961, ch. III, § 3.10).

The Rao metric defines a canonical measure, or Rao measure, on the parameter manifold. If the manifold has a finite volume under the Rao measure, then the measure can be normalized to give a prior density suitable for the Bayesian estimation of parameter values from data samples. The prior density favours regions of the parameter space where the probability densities change rapidly. It is precisely these regions that offer the best chance of finding a density which fits closely to the data samples.

The pinhole camera provides a good model for image formation (Faugeras 1993). The essential components of a pinhole camera are the pinhole itself and an image plane chosen such that it does not contain the pinhole. The pinhole is usually called the optical centre of the camera. Each space point p distinct from the optical centre O defines a line $\langle \mathrm{O}, p\rangle$. The image of p is formed by taking the intersection of $\langle \mathrm{O}, p\rangle$ with the image plane. The totality of lines through O comprises a projective plane $\mathbb{P}^2$. If p is restricted to lie on a line k not containing O, then the totality of lines $\langle \mathrm{O}, p\rangle$ for $p \in k$ forms a projective line $\mathbb{P}^1$.

OK?

The following two examples show how projective transformations of the line arise in computer vision. Suppose that two images of the same scene are taken. A point q in the first image is said to correspond to a point r in the second image, $q \leftrightarrow r$, if q and r are both projections of the same scene point. For the first example, let m_1, m_2 be the two images of the line k. Then m_1, m_2 are both projective lines and the function $m_1 \mapsto m_2$ defined by the pairs of corresponding points is a projective transformation of the line.

For the second example, suppose that the two images are each projections of a plane, Π, such as the side of a building or a flat area of ground. Let q_i, $0 \leqslant i \leqslant n$, be a set of points in the first image of Π, and let r_i, $0 \leqslant i \leqslant n$, be the set of corresponding points in the second image. The set of image lines through q_0 is called the pencil of lines with centre q_0 (Semple & Kneebone 1952). It comprises a projective line $\mathbb{P}^1$. The correspondences $\langle q_0, q_i\rangle \mapsto \langle r_0, r_i\rangle$, $1 \leqslant i \leqslant n$, between image lines are part of a projective transformation from the pencil of lines with centre q_0 to the pencil of lines with centre r_0. A numerical version of this example is described in § 4.

The geometrical properties of the projective transformations of the line are described by Semple & Kneebone (1952) but note that they call these transformations 'homographies'. Hartley & Zisserman (2000) discuss many estimation problems in computer vision, including the estimation of projective transformations of the plane. They use the term homography to mean an projective transformation of the plane.

A projective transformation of the line can be estimated using data in the form of measurements of point correspondences. Algorithms for estimating projective transformations usually involve a search through a range of models, where each model has a geometric part, namely the projective transformation, and a probabilistic part which quantifies the extent to which the measurements are compatible with the projective transformation. The Fisher information and the Rao metric are the basis of estimation algorithms that are unaffected by changes in the parametrization of the models. Invariance under changes in the parametrization is desirable because the parametrization is chosen independently of the true projective transformation and so should not affect any estimate of it.

Parameter estimation becomes more difficult if the functional form of the model is allowed to vary, especially if this includes a variation in the number of parameters. In the general case there is a set of candidate manifolds, $M_1, M_2, \ldots$, with varying dimension. The term 'model selection' refers to the selection of the best manifold amongst the different candidates. It is not appropriate simply to choose the manifold containing the parameter value which best fits the measurements, because models with large numbers of parameters will almost always be favoured over models with small numbers of parameters (Torr *et al.* 2000; Torr & Zisserman 1998). An example of model selection for projective transformations is analysed in § 6. The example contains two manifolds. The first manifold parametrizes the projective transforma-

OK?

tions. The second manifold parametrizes the affine transformations and forms a co-dimensional 1 sub-manifold of the first manifold.

Section 2 describes a model for estimating projective transformations of the line. The Fisher information and the Rao measure of the model are obtained in closed form in § 3. A numerical example is described in § 4. The sub-manifold of affine transformations is described in § 5 and the probability of misclassification is estimated in § 6. Some concluding remarks are made in § 7.

2. Geometrical and probabilistic models

All coordinates will be real numbers and all transformations between spaces will be defined over the real numbers. Let the projective line $\mathbb{P}^1$ have coordinates $(x, y)^{\mathrm{T}}$. As usual, if $y \neq 0$, then (x, y) and $(xy^{-1}, 1)$ refer to the same point of $\mathbb{P}^1$ and xy^{-1} is a point of $\mathbb{R} \subset \mathbb{P}^1$. Let H be the invertible matrix

$$H = \begin{pmatrix} a & b \\ c & d \end{pmatrix}. \tag{2.1}$$

The matrix H defines a projective transformation as follows (Thurston 1997):

$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto H \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ax + by \\ cx + dy \end{pmatrix}. \tag{2.2}$$

Two invertible 2×2 matrices H, K define the same projective transformation if and only if there exists a non-zero scalar λ such that $H = \lambda K$. The projective transformations (2.2) such that $ad - bc > 0$ form a Lie group $\mathrm{PSL}(2, \mathbb{R})$. It is assumed from now on that $ad - bc > 0$. If this condition does not hold in an application, then it can be imposed by reversing the coordinate either in the domain or in the range of the projective transformation.

The matrix H defines an affine transformation if $c = 0$. The affine transformations form a subgroup $\mathrm{A}(2, \mathbb{R})$ of $\mathrm{PSL}(2, \mathbb{R})$.

(*a*) *Parametrization of* PSL(2, ℝ)

The scale factor of the matrix H in (2.1) is fixed by requiring firstly that $\det(H) \equiv ad - bc = 1$, and secondly that the vector (a, c) satisfies either $c > 0$ or $c = 0$, $a > 0$.

There exist $\mu, \tilde{\mu} \in [0, \infty)$, $\alpha \in [0, \pi)$, $\beta \in [-\alpha, 2\pi - \alpha)$ such that

$$(a, c) = \tilde{\mu}(\cos\alpha, \sin\alpha), \tag{2.3}$$

$$(b, d) = \mu(\cos(\alpha + \beta), \sin(\alpha + \beta)). \tag{2.4}$$

The values of μ, $\tilde{\mu}$, α, β are uniquely determined by (2.3) and (2.4). It follows from (2.3), (2.4) and the constraint $\det(H) = 1$ that

$$\tilde{\mu}\mu(\cos\alpha \sin(\alpha + \beta) - \sin\alpha \cos(\alpha + \beta)) = 1;$$

thus $\underline{\sin\beta > 0}$ and

$$\tilde{\mu}^{-1} = \mu \sin\beta. \tag{2.5}$$

Author: unnecessary brackets around β removed here and throughout OK?

It follows from (2.4) and (2.5) that

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \cos\alpha/(\mu\sin\beta) & \mu\cos(\alpha + \beta) \\ \sin\alpha/(\mu\sin\beta) & \mu\sin(\alpha + \beta) \end{pmatrix}. \tag{2.6}$$

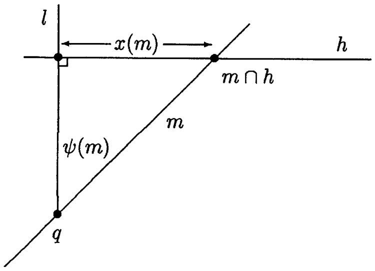

Figure 1. Parametrization of a pencil of lines.

The triple $\theta(H) = (\mu, \alpha, \beta)$ is determined by H, and conversely, (μ, α, β) determines H up to scale. It follows that $\theta = (\mu, \alpha, \beta)$ parametrizes $\mathrm{PSL}(2, \mathbb{R})$. The constraints $\alpha \in [0, \pi)$, $\beta \in [-\alpha, 2\pi - \alpha)$, $\sin \beta > 0$ yield $\beta \in (0, \pi)$. Let D be defined by

$$D = \{(\alpha, \beta) : 0 \leqslant \alpha < \pi,\ 0 < \beta < \pi\}.$$

The set $(0, \infty) \times D$ is mapped bijectively to $\mathrm{PSL}(2, \mathbb{R})$ by $(\mu, \alpha, \beta) \mapsto H$.

(*b*) *Parametrizations of a pencil of lines*

To parametrize a pencil of lines in the image, pick a fixed reference line l through the centre q of the pencil. Each line m in the pencil defines a unique angle $\psi(m) \in [-\pi/2, \pi/2)$ between m and l, as shown in figure 1. The angle $\psi(m)$ is the angular coordinate of m. The reference line l has angular coordinate $\psi(l) = 0$. The argument m may be omitted from $\psi(m)$ if the meaning is clear from the context.

A second parametrization of the pencil is obtained. Let h be a fixed line normal to l such that $\|l \bigcap h - q\| = 1$. The coordinate $x(m)$ of m is defined by $x(m) = \tan(\psi(m))$, i.e. $x(m)$ is the signed distance from $l \bigcap h$ to $m \bigcap h$. The corresponding projective coordinates, $(x(m), 1)$, are given in terms of $\psi(m)$ by $(\sin(\psi(m)), \cos(\psi(m)))$, after multiplying $(x(m), 1)$ by the scale factor $\cos(\psi(m))$. If m is parallel to h, then $x(m)$ is not defined, but $\psi(m) = -\pi/2$, and the projective coordinates of m are $(-1, 0)$ or, equivalently, $(1, 0)$.

Recall the example from § 1 in which a plane Π in space contains a pencil of lines and two images of the pencil are taken by different cameras. Let m_1, m_2 be corresponding lines from the first and second images, respectively, and let $x_1 = x_1(m_1)$, $x_2 = x_2(m_2)$ be the coordinates of the lines. Then x_1, x_2 are related by a projective transformation

$$x_2 = \frac{ax_1 + b}{cx_1 + d}, \tag{2.7}$$

where the coefficients a, b, c, d depend on the relative positions and orientations of the two cameras and Π.

Let ψ_1, ψ_2 be the angular coordinates of m_1, m_2. The function $F(\psi_1, \theta)$ is defined such that $\psi_1 \mapsto F(\psi_1, \theta)$ is the function between angular coordinates equivalent to (2.7),

$$\psi_2 = \tan^{-1}\left(\frac{a \tan \psi_1 + b}{c \tan \psi_1 + d}\right) \equiv F(\psi_1, \theta), \tag{2.8}$$

where $\theta = \theta(a, b, c, d)$ is as defined in § 2 *a*, Suppose that the projective transformation is required to map a specified line m_1 in the first pencil to a specified line m_2 in the second pencil. Then coordinates can be chosen in each pencil such that $m_1 = (1,0)^{\mathrm{T}}$, $m_2 = (1,0)^{\mathrm{T}}$. The projective transformations which map m_1 to m_2 are then exactly the transformations for which $c = 0$ in (2.7), i.e. they are the affine transformations.

(*c*) *Probabilistic model*

Let ϕ_1, ϕ_2 be the measured values of the angular coordinates of the corresponding lines m_1, m_2. The measurement ϕ_1 is given a uniform prior density on $[-\pi/2, \pi/2)$, to reflect the belief that there is no preferred direction for the observed lines. In general, $\phi_2 \neq F(\phi_1, \theta)$, because of the effects of measurement errors.

It is assumed that the difference $\phi_2 - F(\phi_1, \theta)$ has a Gaussian density with standard deviation $\sigma \ll \pi/2$. The pair (ϕ_1, ϕ_2) of measurements has the density

$$p(\phi_1, \phi_2 \mid \theta) = \frac{1}{\sqrt{2\pi^3\sigma^2}} \exp\biggl(-\frac{1}{2\sigma^2}(\phi_2 - F(\phi_1, \theta))^2\biggr). \tag{2.9}$$

In numerical work the difference $\phi_2 - F(\phi_1, \theta)$ must be normalized to the range $[-\pi/2, \pi/2)$. The advantage of (2.9) is that it is simple enough to allow the calculation of closed form expressions for the Fisher information and the Rao measure, but at the same time it is similar to the probability density functions used in practice, for example by Hartley & Zisserman (2000) in their § 3.3.

The above assumption that ϕ_1 has a uniform density on $[-\pi/2, \pi/2)$ differs from the usual formulation in which ϕ_1 is the sum of a true measurement ψ_1 and an error ε_1. If ψ_1 has a uniform prior density on $[-\pi/2, \pi/2)$ and ε_1 is Gaussian and with a small standard deviation, then the density of $\psi_1 + \varepsilon_1$ is closely approximated by the uniform density.

In many previous applications of projective transformations to computer vision the coordinates x_1, x_2, rather than ϕ_1, ϕ_2, are used. The problem with x_1 (and x_2) is that the weighting of the measurement errors in x_1 should decrease as $|x_1|$ increases: if $|x_1|$ is large, then a large error in the measurement of x_1 may be equivalent to only a small error in the orientation of the line m_1 specified by x_1. The strategy adopted here is to use the more complicated angular coordinates, ϕ_1, ϕ_2, and to make the probability density function for the errors depend on ϕ_1 and ϕ_2 only through the difference $\phi_2 - F(\phi_1, \theta)$.

3. Properties of the Fisher information for PSL$(2, \mathbb{R})$

In § 3 *a*–*e* closed form expressions are obtained for the Fisher information and the Rao measure. The final subsection, § 3 *f*, gives a justification for using the Rao metric to compare probability density functions and for using the Rao measure to define a prior density on the parameter manifold.

(*a*) *Fisher information*

The Fisher information, $J(\theta)$, for the parametrized family of densities $\theta \mapsto p(\phi_1, \phi_2 \mid \theta)$ is the symmetric 3×3 matrix defined by (Amari 1985; Balasubramanian

1996, 1997; Lindsey 1996)

$$J_{ij}(\theta) = -E_\phi\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln(p(\phi_1,\phi_2 \mid \theta))\right), \quad 1 \leqslant i,j \leqslant 3,$$

where E_ϕ is the expected value with respect to the density (2.9) for ϕ. If $\hat{\theta}$ is the maximum-likelihood estimation of θ for data consisting of independent measurements of the coordinates of N pairs of corresponding points, then as $N \to \infty$ the probability density of $\hat{\theta} - \theta$ tends to the Gaussian density $\mathcal{N}(0, N^{-1}J(\theta)^{-1})$ (Lindsey 1996).

A short calculation shows that

$$J_{ij}(\theta) = E_\phi\left(\left(\frac{\partial}{\partial\theta_i}\ln(p(\phi_1,\phi_2 \mid \theta))\right)\left(\frac{\partial}{\partial\theta_j}\ln(p(\phi_1,\phi_2 \mid \theta))\right)\right), \quad 1 \leqslant i,j \leqslant 3. \quad (3.1)$$

It follows from (2.9) that

$$\frac{\partial}{\partial\theta_i}\ln p(\phi_1,\phi_2 \mid \theta) = \sigma^{-2}(\phi_2 - F(\phi_1,\theta))\frac{\partial F}{\partial\theta_i}, \quad 1 \leqslant i \leqslant 3;$$

thus

$$J_{ij}(\theta) = \sigma^{-4}E_\phi\left((\phi_2 - F(\phi_1,\theta))^2\frac{\partial F}{\partial\theta_i}\frac{\partial F}{\partial\theta_j}\right) = \frac{1}{\pi\sigma^2}\int_{-\pi/2}^{\pi/2}\frac{\partial F}{\partial\theta_i}\frac{\partial F}{\partial\theta_j}\,\mathrm{d}\phi_1, \quad 1 \leqslant i,j \leqslant 3. \quad (3.2)$$

Author: rearrangement of equation OK? Also please distinguish between variable '*d*' and differential 'd' throughout.

The Fisher information, $J(\theta)$, defines on $\mathrm{PSL}(2,\mathbb{R})$ a Riemannian metric known in statistics as the Rao metric for the family of densities (2.9). If θ_1, θ_2 are nearby points of $\mathrm{PSL}(2,\mathbb{R})$, then the square of the distance between θ_1 and θ_2 is

$$(\theta_1 - \theta_2)^{\mathrm{T}}J(\theta_1)(\theta_1 - \theta_2) + O(\|\theta_1 - \theta_2\|^3).$$

Under the Rao metric, $\mathrm{PSL}(2,\mathbb{R})$ has a canonical measure $\tau(\theta)\,\mathrm{d}\theta$ defined by (Gallot *et al.* 1990)

$$\tau(\theta) = \sqrt{|\det(J(\theta))|}. \quad (3.3)$$

In applications to statistics the canonical measure $\tau(\theta)\,\mathrm{d}\theta$ is called the Rao measure.

(*b*) *Invariance of the Rao metric under rotations*

The Fisher information and the Rao metric are independent of the coordinate α introduced in § 2 *a*. To see this, let $R(\gamma)$ be the 2×2 rotation matrix

$$R(\gamma) = \begin{pmatrix}\cos\gamma & -\sin\gamma\\ \sin\gamma & \cos\gamma\end{pmatrix}.$$

The matrix $R(\gamma)$ is an element of the special orthogonal group SO(1). The group SO(1) acts on $\mathrm{PSL}(2,\mathbb{R})$ by matrix multiplication on the left, $H \mapsto R(\gamma)H$. Let $\theta(H) = (\mu,\alpha,\beta)$, with the action of $R(\gamma)$ on the parameter vector θ written as $\underline{\theta \mapsto R(\gamma)\cdot\theta}$. A short calculation shows that

Change to centred dot here OK? Is this multiplication or scalar product?

$$R(\gamma)\begin{pmatrix}\cos\alpha/(\mu\sin\beta) & \mu\cos(\alpha+\beta)\\ \sin\alpha/(\mu\sin\beta) & \mu\sin(\alpha+\beta)\end{pmatrix} = \begin{pmatrix}\cos(\alpha+\gamma)/(\mu\sin\beta) & \mu\cos(\alpha+\beta+\gamma)\\ \sin(\alpha+\gamma)/(\mu\sin\beta) & \mu\sin(\alpha+\beta+\gamma)\end{pmatrix}. \quad (3.4)$$

Let the function $\gamma \mapsto s(\gamma)$ be defined as follows. If there exists an integer n such that $\alpha+\gamma+2n\pi=0$, then set $s(\gamma)=1$. Otherwise, choose $s(\gamma)\in\{-1,1\}$ and choose the integer n such that $0<s(\gamma)(\alpha+\gamma+2n\pi)<\pi$. It follows that

$$R(\gamma)\cdot\theta=\begin{cases}(\mu,\alpha+\gamma+2n\pi,\beta) & \text{if } s(\gamma)=1,\\ (\mu,\alpha+\gamma+(2n+1)\pi,\beta) & \text{if } s(\gamma)=-1.\end{cases}$$

Note that if $s(\gamma)=-1$, then the matrix $R(\gamma)H$ is scaled by -1 before obtaining the components of $\theta(R(\gamma)H)$.

Let $(u,v)^{\mathrm{T}}$ be defined by

$$\begin{pmatrix}u\\ v\end{pmatrix}=R(\gamma)H\begin{pmatrix}\tan\phi_1\\ 1\end{pmatrix}.$$

The vector $(u,v)^{\mathrm{T}}$ is parallel to

$$R(\gamma)\begin{pmatrix}\sin(F(\phi_1,\theta(H)))\\ \cos(F(\phi_1,\theta(H)))\end{pmatrix}. \tag{3.5}$$

It follows from (3.4) and (3.5) that

$$F(\phi_1,\theta(R(\gamma)H))=\tan^{-1}(u/v)=F(\phi_1,\theta(H))-\gamma;$$

thus

$$J_{ij}(R(\gamma)\cdot\theta)=\sigma^{-4}E_\phi\biggl((\phi_2-F(\phi_1,\theta)+\gamma)^2\frac{\partial F}{\partial\theta_i}\frac{\partial F}{\partial\theta_j}\biggr)=J_{ij}(\theta),\quad 1\leqslant i,j\leqslant 3.$$

The Fisher information and the Rao metric are invariant under the action of $SO(1)$ on $\mathrm{PSL}(2,\mathbb{R})$. In particular, $J(\mu,\alpha,\beta)=J(\mu,0,\beta)$.

(*c*) *Closed form expressions for $J(\theta)$ and $\tau(\theta)$*

It follows from (2.6) and (2.8) that

$$F(\phi,\theta)=\tan^{-1}\biggl(\frac{\cos\alpha(\mu\sin\beta)^{-1}\tan\phi+\mu\cos(\alpha+\beta)}{\sin\alpha(\mu\sin\beta)^{-1}\tan\phi+\mu\sin(\alpha+\beta)}\biggr).$$

The six different integrals in (3.2) can all be evaluated in closed form using Mathematica (Wolfram 1999). Certain substitutions are necessary to speed up the calculations. For example, before submitting to Mathematica, the right-hand side of

$$J_{11}(\theta)=\frac{1}{\pi\sigma^2}\int_{-\pi/2}^{\pi/2}\frac{\mu^2\sin^2(2\phi_1)}{(\operatorname{cosec}^2\beta\cos^2(\phi_1)+\mu^4\sin^2(\phi_1)+\mu^2\cot\beta\sin(2\phi_1))^2}\,\mathrm{d}\phi_1.$$

Here $\cos^2\phi$ is replaced by $\frac{1}{2}(1+\cos(2\phi))$, $\sin^2\phi$ is replaced by $\frac{1}{2}(1-\cos(2\phi))$, $\cot\beta$ is replaced by a symbol such as w and $\operatorname{cosec}^2\beta$ is replaced by $1+w^2$. With these replacements the integral is evaluated quickly in closed form.

Author: change to text and punctuation added OK? Please clarify sentence(s)!

Let r be defined by

$$r=(1+\mu^2)^2+\cot^2\beta.$$

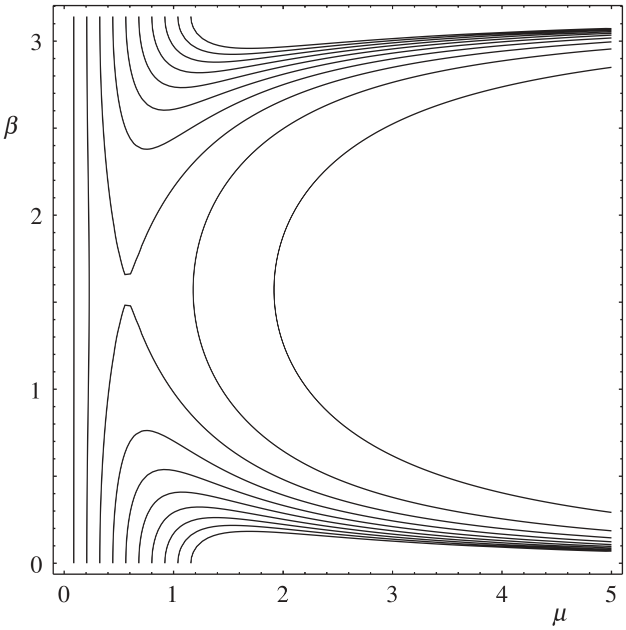

Figure 2. Contour plot of $(\mu, \beta) \mapsto \tau(\mu, 0, \beta)$.

The following expressions for $J(\theta)$ and $\tau(\theta)$ are obtained:

$$\left.\begin{aligned}
J_{11} &= 2\sigma^{-2}r^{-2}((1+\mu^2)^2 + (2+4\mu^2+\mu^4)\cot^2\beta + \cot^4\beta),\\
J_{12} &= J_{21} = -2\sigma^{-2}r^{-1}\mu\cot\beta,\\
J_{13} &= J_{31} = \sigma^{-2}r^{-2}\mu\cot\beta(1-\mu^4+(2+4\mu^2+\mu^4)\cot^2\beta+\cot^4\beta),\\
J_{22} &= \sigma^{-2},\\
J_{23} &= J_{32} = \sigma^{-2}r^{-1}\mu^2(1+\mu^2-\cot^2\beta),\\
J_{33} &= \frac{\mu^2}{2\sigma^2r^2}(2\mu^6+4\mu^2(1+\cot^4\beta)+\mu^4(5-2\cot^2\beta+\cot^4\beta)+\operatorname{cosec}^6\beta)
\end{aligned}\right\} \tag{3.6}$$

and

$$\tau(\theta) = \frac{\mu}{\sigma^3(\cos^2\beta + \sin^2\beta(\mu^2+1)^2)}. \tag{3.7}$$

A contour plot of $\tau(\theta)$ is shown in figure 2. The function $(\mu, \beta) \mapsto \tau(\mu, 0, \beta)$ has a saddle point at $\mu = 1/\sqrt{3}$, $\beta = \frac{1}{2}\pi$.

(*d*) *Ricci tensor and scalar curvature*

Various curvature properties of $\mathrm{PSL}(2,\mathbb{R})$ under the Rao metric can be calculated. The Ricci curvature and the scalar curvature (Gallot *et al.* 1990; Jost 1995) are obtained in this subsection, but note that neither is required for the estimation in § 6 of the probability of misclassification.

The definitions of the Ricci tensor and the scalar curvature are given by Misner *et al.* (1973) together with the formulae for calculating them. Let the partial derivatives of a function f with respect to μ, α or β be denoted by $f_{,i}$, with the appropriate i. For example, $f_{,\beta} = \partial f/\partial\beta$. Define the connection coefficients $\mathit{\Gamma}_{ijk}$ by

$$\mathit{\Gamma}_{ijk} = \tfrac{1}{2}(J_{ij,k} + J_{ik,j} - J_{jk,i}).$$

Let J^{ij} be the i,jth entry of the inverse matrix J^{-1}. Indices are raised using J^{-1}, $\mathit{\Gamma}^{i}_{jk} = J^{im}\mathit{\Gamma}_{mjk}$, where the Einstein summation convention applies to the repeated index m in the usual way.

The Ricci tensor $\underline{\mathsf{R}}$ and the scalar curvature κ are defined by

Author: change to notation for clarity OK here and below?

$$\begin{aligned} \mathsf{R}_{ij} &= \tau^{-1}(\tau\mathit{\Gamma}^{k}_{ij})_{,k} - (\ln(\tau))_{,ij} - \mathit{\Gamma}^{k}_{li}\mathit{\Gamma}^{l}_{jk}, \\ \kappa &= R^{i}{}_{i} \equiv J^{ij}R_{ji}. \end{aligned}$$

Calculations with Mathematica yield the covariant components of R,

$$\mathsf{R} = \frac{1}{\mu^2\sin^4\beta}\begin{pmatrix} -2\sin^2\beta & 0 & -\mu\sin\beta\cos\beta \\ 0 & 0 & 0 \\ -\mu\sin\beta\cos\beta & 0 & -\tfrac{1}{2}\mu^2 \end{pmatrix}$$

and the scalar curvature

$$\kappa = -2\sigma^2\left(2 + \mu^2 + \frac{1}{\mu^2\sin^2\beta}\right). \tag{3.8}$$

It follows from (3.8) that κ has a global maximum of $-8\sigma^2$, attained at all points on the codimension-2 sub-manifold defined by $\mu = 1$, $\beta = \tfrac{1}{2}\pi$.

(*e*) *Finite volume subset of* PSL(2, ℝ)

The Rao measure $\tau(\theta)\,\mathrm{d}\theta$ is a candidate for the prior density on $\mathrm{PSL}(2,\mathbb{R})$ needed for a Bayesian estimation of θ from a sample of independent measurements of the coordinates of pairs of corresponding points. A difficulty arises, because the Rao measure cannot be normalized: the integral of $\tau(\theta)\,\mathrm{d}\theta$ over $\mathrm{PSL}(2,\mathbb{R})$ is infinite.

To overcome the problem of infinite volume, the range of μ is restricted to the interval $[0,\mu_m]$, where μ_m is large and positive. Define $B(\mu_m)$, $V(\mu_m)$ by

$$\left.\begin{aligned} B(\mu_m) &= \{(\mu,\alpha,\beta) \in \mathrm{PSL}(2,\mathbb{R}),\ 0 \leqslant \mu \leqslant \mu_m\}, \\ V(\mu_m) &= \int_{B_{\mu_m}} \tau(\theta)\,\mathrm{d}\theta. \end{aligned}\right\} \tag{3.9}$$

The volume $V(\mu_m)$ is finite and $\tau(\theta)/V(\mu_m)$ is a probability density function on $B(\mu_m)$. It follows from (3.7) and (3.9) that

$$V(\mu_m) = \frac{\pi^2}{2\sigma^3}\ln(1 + \mu_m^2), \quad 0 \leqslant \mu_m < \infty. \tag{3.10}$$

(*f*) *Justification for the Rao metric and the Rao measure*

Any manifold M has an infinite number of Riemannian metrics defined on it. If M parametrizes a family of probability density functions, $\theta \mapsto p(x \mid \theta)$, $\theta \in M$, then why should the Rao metric be used to measure the distance between nearby parameter values θ_1, θ_2?

A partial answer can be given as follows. Let d_m be the dimension of M and let $\Delta\theta = \theta_2 - \theta_1$. If

$$\sum_{i,j=1}^{d_m} J_{ij}\Delta\theta_i\Delta\theta_j \equiv \Delta\theta^{\mathrm{T}} J(\theta_1)\,\Delta\theta$$

is small, then, as noted in the first paragraph of § 1, the probability density functions $p(x \mid \theta_1)$, $p(x \mid \theta_2)$ provide similar descriptions of the data x. There is no reason to select one of θ_1, θ_2 in preference to the other.

Amari (1985) develops the argument further, by showing that a wide range of methods for comparing probability density functions are well approximated by the Rao metric when the densities being compared are close together on the manifold. Suppose that two probability densities $p(x \mid \theta_1)$, $p(x \mid \theta_2)$ are compared using a function $D(\theta_1, \theta_2)$ of the form

$$D(\theta_1, \theta_2) = \int G(p(x \mid \theta_1), p(x \mid \theta_2))p(x \mid \theta_1)\,\mathrm{d}x,$$

where G is a function differentiable up to third order. If $D(\theta_1, \theta_2)$ is required to be invariant under reparametrizations of the *data* x, then there exists a function g such that $G(p(x \mid \theta_1), q(x \mid \theta_2)) = g(p(x \mid \theta_2)/p(x \mid \theta_1))$. It follows that

$$D(\theta_1, \theta_2) = g(1) + \tfrac{1}{2}g''(1)\,\Delta\theta^{\mathrm{T}} J(\theta_1)\,\Delta\theta + \text{third-order terms in } \Delta\theta. \qquad (3.11)$$

The well-known Kullback–Leibler, Bhattacharya–Matusita–Hellinger and Chernoff methods for comparing densities satisfy (3.11).

The Rao measure $\tau(\theta)\,\mathrm{d}\theta = |\det(J(\theta))|^{1/2}\,\mathrm{d}\theta$ is important because it measures the density of models parametrized by M (Balasubramanian 1996; Myung *et al.* 2000). More precisely, if $\tau(\theta)$ is large then $p(x \mid \theta)$ varies rapidly with θ. When $\tau(\theta)\,\mathrm{d}\theta$ is used as a prior density for θ it makes explicit a bias which would otherwise be hidden: those regions where $p(x \mid \theta)$ varies rapidly are more likely to contain a value of θ for which $p(x \mid \theta)$ is a good fit to the data.

The results of Balasubramanian (1996) and Myung *et al.* (2000) shed light on a long-standing controversy: what should be done when a prior density for θ cannot be obtained from $\tau(\theta)\,\mathrm{d}\theta$ because the integral of $\tau(\theta)\,\mathrm{d}\theta$ over M is infinite? Suggested solutions (Jeffreys 1961) include (i) modifying the form of $p(x \mid \theta)$; (ii) using $\tau(\theta)\,\mathrm{d}\theta$ as an unnormalized or improper density; or (iii) restricting θ to a subset of M over which the integral of $\tau(\theta)\,\mathrm{d}\theta$ is finite. The root of the problem is that M parametrizes an infinite number of probability density functions, one for each value of θ. In many cases there is a 'lucky cancellation': the number of densities is infinite, but nearby densities are indistinguishable, leaving in effect only a finite number of distinguishable densities. As a consequence, the integral of $\tau(\theta)\,\mathrm{d}\theta$ over M is finite. However, there are cases in which M contains an infinite number of distinguishable densities and the integral of $\tau(\theta)\,\mathrm{d}\theta$ over M is infinite. The problem is that M parametrizes 'too many' densities. This suggests that option (i) above is not appropriate. Option (ii) is ruled out because it contradicts the axioms of probability theory. The preferred option in this work is (iii).

Table 1. *Pixel coordinates of the image points q_i, r_i*

i	0	1	2	3	4	5	6	7
q_i	$392, 217$	$392, 119$	$454, 118$	$502, 116$	$546, 113$	$545, 144$	$546, 187$	$546, 220$
r_i	$522, 220$	$524, 124$	$575, 121$	$615, 118$	$651, 117$	$653, 151$	$650, 193$	$650, 226$

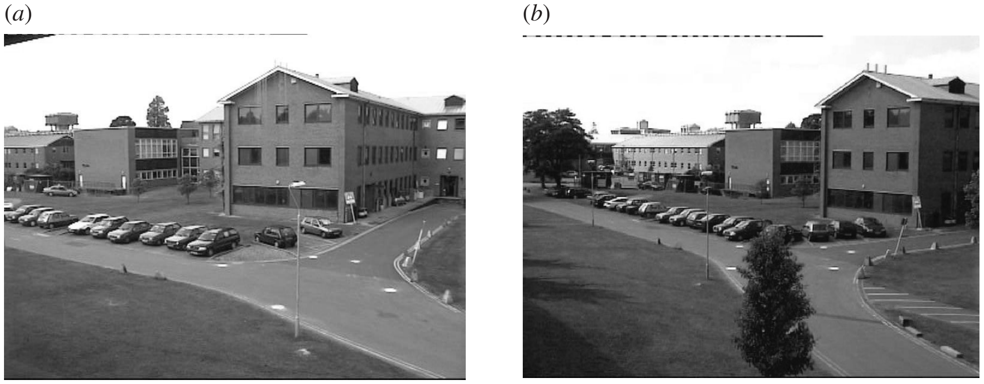

Figure 3. Test images.

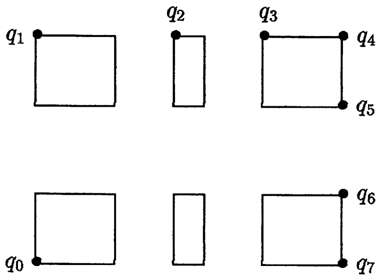

Figure 4. Diagram showing windows and chosen measurements.

4. Numerical example

The example is based on two images of a pencil of lines, as described in § 1. The projective transformation matches corresponding lines in the images. Figure 3 shows two images from the PETS'2001 database (Ferryman 2001). The left-hand image is number 0017.jpg in /DATASET1/TESTING/CAMERA1_JPEGS/ and the right-hand image is number 0017.jpg in /DATASET3/TESTING/CAMERA2_JPEGS/.

Author: these URLs do not appear to be complete! Please provide the domain name, etc. Is this case-dependent (style is to make URLs lower-case)? Author: change from ref citation to footnote OK?

In the left-hand image the front face of the nearest building has a pattern of six windows as illustrated in figure 4. Eight of the corners of the windows are labelled $q_0, \dots, q_7$. The matching points in the right-hand image are $r_0, \dots, r_7$. The pixel coordinates of all 16 points were measured by hand using the `xv` program.† The measurements are shown in table 1. The image origin is at the top left-hand corner.

The angular coordinates $\phi_{q,i}$ of the lines $\langle q_0, q_i \rangle$, $1 \leqslant i \leqslant 7$, were calculated using $\langle q_0, q_0 + (0, 1) \rangle$ as the line with angular coordinate 0. The angular coordinates $\phi_{r,i}$ of the $\langle r_0, r_i \rangle$, $1 \leqslant i \leqslant 7$, were calculated similarly. The estimate $\hat{\theta}$ was obtained by

† Available from John Bradley at http://www.trilon.com/xv/.

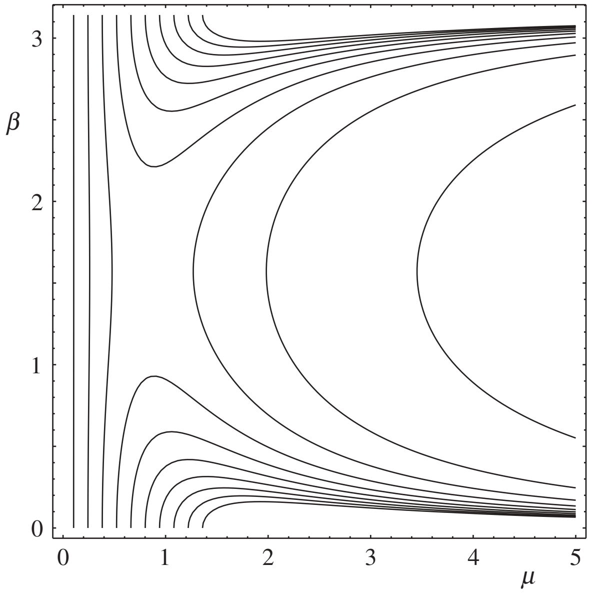

Figure 5. Contour plot of $(\mu,\beta)\mapsto\tau_{\mathrm{A}}(\mu,\beta)$.

numerical minimization of

$$\sum_{i=1}^{7}(\phi_{r,i}-F(\phi_{q,i},\theta))^2$$

with the result $\hat{\theta}=(1.106\,72, 3.169\,56, 1.560\,02)^{\mathrm{T}}$. The corresponding 2×2 matrix $\hat{H}$ is

$$\hat{H}=\begin{pmatrix}-0.903\,267 & 0.019\,028\,5\\ -0.025\,269\,9 & -1.106\,56\end{pmatrix}.$$

The estimated covariance C of the error in $\hat{\theta}$ is $C=7^{-1}J(\hat{\theta})^{-1}$, which yields

$$C=\sigma^2\begin{pmatrix}91.6169 & 0.014\,252\,7 & -4.628\,23\\ 0.014\,252\,7 & 0.143\,131 & -0.000\,497\,125\\ -4.628\,23 & -0.000\,497\,125 & 0.233\,988\end{pmatrix}$$

The standard deviation in the measurements of the coordinates of the points q_i, r_i is estimated to be one pixel. The value of σ is estimated at $\sigma=1/50$ rad, after noting that $\|q_0-q_i\|$, $\|r_0-r_i\|$, $1\leqslant i\leqslant 7$, are in the range of 96–186 pixels and after taking into account the effects of the errors in q_0, r_0 and in the q_i, r_i, $1\leqslant i\leqslant 7$, on the orientations of the lines. The estimated standard deviations of the errors in the components $\hat{\mu}$, $\hat{\alpha}$, $\hat{\beta}$ of $\hat{\theta}$ are, respectively, 0.19, 0.0076, 0.0097, working to two significant figures. The numerical error in $\hat{\mu}$ is likely to be much larger than the numerical errors in $\hat{\alpha}$ and $\hat{\beta}$.

In many practical applications the coordinates of points or the positions of lines are measured automatically (Sonka *et al.* 1999).

5. Affine transformations

The projective transformation (2.2) is an affine transformation if and only if $c=0$,

$$x_2=d^{-1}(ax_1+b).$$

As noted in § 2 *b*, if a projective transformation is required to map a specified point in the domain to a specified point in the range, then coordinates can be chosen such that both points have coordinates $(1,0)^{\mathrm{T}}$. The projective transformation is then affine.

Under certain conditions a general projective transformation can be approximated by an affine transformation. For example, suppose that a line k in space has images m_1, m_2 taken by two cameras placed such that each image plane is near parallel to k. The projective transformation from m_1 to m_2 defined by pairs of corresponding points is approximated by an affine transformation if m_1, m_2 are each parametrized such that the point at infinity in the image is $(1,0)$.

The affine transformations form a subgroup $\mathrm{A}(2,\mathbb{R})$ of $\mathrm{PSL}(2,\mathbb{R})$. Under the parametrization $\theta = (\mu,\alpha,\beta)$ of $\mathrm{PSL}(2,\mathbb{R})$, $\mathrm{A}(2,\mathbb{R})$ is the hypersurface defined by $\alpha = 0$.

Let J_{ij}, $1 \leqslant i,j \leqslant 3$, be the components (3.6) of the Fisher information for $\mathrm{PSL}(2,\mathbb{R})$. The Fisher information J_{A}, canonical measure $\tau_{\mathrm{A}}(\mu,\beta)\,\mathrm{d}\mu\,\mathrm{d}\beta$, and scalar curvature κ_{A} of $\mathrm{A}(2,\mathbb{R})$ under the parametrization $\theta_{\mathrm{A}} = (\mu,\beta)$, $0 \leqslant \mu < \infty$, $0 < \beta < \pi$, are given by

$$J_{\mathrm{A}}(\theta_{\mathrm{A}}) = \begin{pmatrix} J_{11} & J_{13} \\ J_{31} & J_{33} \end{pmatrix},$$

$$\tau_{\mathrm{A}}(\theta_{\mathrm{A}}) = \frac{\mu(1+2\mu^2\sin^2\beta)^{1/2}}{\sigma^2(\cos^2\beta + \sin^2\beta(\mu^2+1)^2)},$$

$$\kappa_{\mathrm{A}}(\theta_{\mathrm{A}}) = \frac{-2\sigma^2(1+3\mu^2\sin^2\beta)(\cos^2\beta+\sin^2\beta(\mu^2+1)^2)}{\mu^2\sin^2\beta(1+2\mu^2\sin^2\beta)^2}.$$

The function $\tau_{\mathrm{A}}(\mu,\beta)$ has a saddle point at

$$\mu = (1+\sqrt{17})^{1/2}/\sqrt{8}, \qquad \beta = \pi/2.$$

The scalar curvature κ_{A} is always negative, and it has no global maximum. A contour plot of $\tau_{\mathrm{A}}(\mu,\beta)$ is shown in figure 5.

Calculations with Mathematica show that $\mathrm{A}(2,\mathbb{R})$ is an Einstein manifold with Einstein curvature equal to zero (Gallot *et al.* 1990; Jost 1995; Misner *et al.* 1973).

(*a*) *Volume of* A(2, ℝ)

The volume of $\mathrm{A}(2,\mathbb{R})$ under the Rao metric is infinite. To prove this, let $\omega_{\mathrm{A}}(\mu)$ be the function defined by

$$\omega_{\mathrm{A}}(\mu) = \int_0^{\pi} \tau_{\mathrm{A}}(\mu,\beta)\,\mathrm{d}\beta. \tag{5.1}$$

The function $\tau_{\mathrm{A}}(\mu,\beta)$ is bounded from below:

$$\tau_{\mathrm{A}}(\mu,\beta) \geqslant \frac{\mu}{\sigma^2(\cos^2\beta+\sin^2\beta(\mu^2+1)^2)}.$$

Thus

$$\omega_{\mathrm{A}}(\mu) \geqslant \int_0^{\pi} \frac{\mu}{\sigma^2(\cos^2\beta + \sin^2\beta(\mu^2+1)^2)}\,\mathrm{d}\beta,$$
$$= \frac{\pi\mu}{\sigma^2(\mu^2+1)}, \quad 0 \leqslant \mu < \infty. \tag{5.2}$$

It follows from (5.2) that the integral of $\omega_{\mathrm{A}}(\mu)$ over $[0,\infty)$ is infinite; thus $\mathrm{A}(2,\mathbb{R})$ has infinite volume.

In a similar manner to § 3 *e*, let $B_{\mathrm{A}}(\mu_m)$, $V_{\mathrm{A}}(\mu_m)$ be defined by

$$B_{\mathrm{A}}(\mu_m) = \{(\mu,\beta) \in \mathrm{A}(2,\mathbb{R}), 0 \leqslant \mu \leqslant \mu_m\},$$
$$V_{\mathrm{A}}(\mu_m) = \int_{B_{\mathrm{A}}(\mu_m)} \tau_{\mathrm{A}}(\theta_{\mathrm{A}})\,\mathrm{d}\theta_{\mathrm{A}}.$$

It follows from (5.1) that

$$V_{\mathrm{A}}(\mu_m) = \int_0^{\mu_m} \omega_{\mathrm{A}}(\mu)\,\mathrm{d}\mu, \quad 0 \leqslant \mu_m < \infty.$$

The function $\tau_{\mathrm{A}}(\theta_{\mathrm{A}})/V_{\mathrm{A}}(\mu_m)$ is a probability density on $B_{\mathrm{A}}(\mu_m)$.

The ratio $V_{\mathrm{A}}(\mu_m)/V(\mu_m)$ is bounded above and below as follows. The function $\tau_{\mathrm{A}}(\mu,\beta)$ is bounded above by

$$\tau_{\mathrm{A}}(\mu,\beta) \leqslant \frac{\mu(1+\mu^2\sin^2\beta)}{\sigma^2(\cos^2\beta + (1+\mu^2)^2\sin^2\beta)};$$

thus

$$\omega_{\mathrm{A}}(\mu) \leqslant \frac{2\pi\mu}{\sigma^2(2+\mu^2)} \leqslant \frac{2\pi\mu}{\sigma^2(1+\mu^2)}, \quad 0 \leqslant \mu < \infty. \tag{5.3}$$

It follows from (5.2) and (5.3) that

$$\frac{\sigma}{\pi} \leqslant \frac{V_{\mathrm{A}}(\mu_m)}{V(\mu_m)} \leqslant \frac{2\sigma}{\pi}, \quad 0 \leqslant \mu_m < \infty. \tag{5.4}$$

An expression for $\omega_{\mathrm{A}}(\mu)$ can be obtained in terms of elliptic integrals. Let $K(m)$ be the complete elliptic integral of the first kind and let $\Pi(n \mid m)$ be the complete elliptic integral of the third kind (Abramowitz & Stegun 1965; Wolfram 1999). The integral (5.1) is evaluated using Mathematica, to yield

$$\omega_{\mathrm{A}}(\mu) = \frac{2\mu}{\sigma^2(2+\mu^2)}\,(2K(-2\mu^2) + \mu^2\Pi(-2\mu^2-\mu^4 \mid -2\mu^2)).$$

Numerical calculation shows that $\omega_{\mathrm{A}}(\mu)$ has a global maximum at $\mu = 1.224\,35\ldots$.

(*b*) *Geodesic distance to* $\mathrm{A}(2,\mathbb{R})$

Let $d(\theta,\mathrm{A})$ be the signed geodesic distance from θ to $\mathrm{A}(2,\mathbb{R})$. To be specific about the sign, let $d(\theta,\mathrm{A})$ be positive if α is small and positive. Let θ_{A} be the closest point of $\mathrm{A}(2,\mathbb{R})$ to θ and let $h = (0,1,0)^{\mathrm{T}}$. The manifold $\mathrm{A}(2,\mathbb{R})$ is the set of $\theta \in \mathrm{PSL}(2,\mathbb{R})$

such that $\theta \cdot h = 0$. If θ is close to $\mathrm{A}(2,\mathbb{R})$, then θ_{A} is estimated by minimizing $(\theta - u)^{\mathrm{T}} J(\theta)(\theta - u)$ over u subject to the constraint $u \cdot h = 0$. Let λ_1 be a Lagrange multiplier and define V_1 by

$$V_1(u) = (u-\theta)^{\mathrm{T}} J(\theta)(u-\theta) + 2\lambda_1 u \cdot h.$$

On solving $\partial V_1(u)/\partial u = 0$ and $\partial V_1(u)/\partial \lambda_1 = 0$, it follows that, to leading order

$$\left.\begin{aligned} \theta_{\mathrm{A}} &= \theta - (\theta \cdot h)(h^{\mathrm{T}} J(\theta)^{-1} h)^{-1} J(\theta)^{-1} h, \\ d(\theta, \mathrm{A}) &= (h^{\mathrm{T}} J(\theta)^{-1} h)^{-1/2} (\theta \cdot h). \end{aligned}\right\} \tag{5.5}$$

6. Model selection

The model-selection problem, as described in § 1, is to select the best parameter manifold for the data from a list of candidates $M_1, M_2, \ldots$. Consider the two parameter manifolds $\mathrm{PSL}(2,\mathbb{R})$ and $\mathrm{A}(2,\mathbb{R})$. After scaling, the Rao measures $\tau(\theta)\,\mathrm{d}\theta$, $\tau_{\mathrm{A}}(\mu,\beta)\,\mathrm{d}\mu\,\mathrm{d}\beta$ define prior densities on suitable subsets of $\mathrm{PSL}(2,\mathbb{R})$, $\mathrm{A}(2,\mathbb{R})$. With these prior densities, model selection can be based on a Bayesian decision rule: select the model with the greatest probability, given the data. The aim in this section is to approximate the probability of misclassification, i.e. the probability that the decision rule selects $\mathrm{A}(2,\mathbb{R})$ when the true model is $\mathrm{PSL}(2,\mathbb{R})$. Numerical calculations suggest that the probability of misclassification tends to a limit as the chosen subsets of $\mathrm{PSL}(2,\mathbb{R})$, $\mathrm{A}(2,\mathbb{R})$ increase in size.

The probability of misclassification is a fundamental property of the family of densities (2.9). Its value depends on the balance between the 'number' of distinguishable probability density functions corresponding to points in $\mathrm{PSL}(2,\mathbb{R})$ but near to $\mathrm{A}(2,\mathbb{R})$ and the bias of the Bayes decision rule in favour of the model $\mathrm{A}(2,\mathbb{R})$ with fewer parameters.

(a) *Expression for the probability of misclassification*

Let $E = \{e_1, \ldots, e_N\}$ be a set of N independent samples from the probability density function $p(\phi \mid t)$ defined by (2.9), where t is the true value of θ. The two manifolds $\mathrm{PSL}(2,\mathbb{R})$, $\mathrm{A}(2,\mathbb{R})$ provide different models for E. Let $\mathcal{P}$, for projective transformation, be the model in which t is drawn from $\mathrm{PSL}(2,\mathbb{R})$ and let $\mathcal{A}$ be the model in which t is drawn from $\mathrm{A}(2,\mathbb{R})$.

Let $C(\mathcal{P})$ be the event that $\mathcal{P}$ is selected as the best model and let $C(\mathcal{A})$ be the event that $\mathcal{A}$ is selected as the best model. A misclassification occurs if $\mathcal{A}$ is selected when $\mathcal{P}$ is the true model for E. Let $B(\mu_m)$ be the set defined by (3.9). Suppose initially that t is drawn from $B(\mu_m)$ according to the density $\tau(\theta)/V(\mu_m)$. The probability $p_{\mathrm{M}}(N, \mu_m)$ of misclassification is

$$p_{\mathrm{M}}(N, \mu_m) = V(\mu_m)^{-1} \int_{B(\mu_m)} \tau(\theta) P(C(\mathcal{A}) \mid \theta)\,\mathrm{d}\theta. \tag{6.1}$$

The integral on the right-hand side of (6.1) includes cases in which θ is a point of $\mathrm{A}(2,\mathbb{R})$. It is still appropriate to call $p_{\mathrm{M}}(N, \mu_m)$ the probability of misclassification because the contribution of such points to $p_{\mathrm{M}}(N, \mu_m)$ is zero.

In the following subsections the probability $p_{\mathrm{M}}(N, \mu_m)$ is estimated as a function of the number of samples N and the standard deviation σ of the measurement errors.

The probability of misclassification for t drawn from $\mathrm{PSL}(2,\mathbb{R})$ is obtained as the limit of $p_{\mathrm{M}}(N,\mu_m)$ for $\mu_m \to \infty$.

(*b*) *Criterion for model selection*

The Bayesian approach to model selection is taken. Let $P(E)$ be the prior probability of E, and let $P(\mathcal{P})$, $P(\mathcal{A})$ be the prior probabilities of $\mathcal{P}$, $\mathcal{A}$. Let $p(E \mid \theta)$ be the probability of E given θ:

$$p(E \mid \theta) = \prod_{i=1}^{N} p(e_i \mid \theta).$$

An application of Bayes's rule (Balasubramanian 1997) yields

$$\begin{aligned} P(\mathcal{P} \mid E) &= \frac{P(\mathcal{P})}{P(E)V(\mu_m)} \int_{B(\mu_m)} \tau(\theta) p(E \mid \theta)\, \mathrm{d}\theta, \\ P(\mathcal{A} \mid E) &= \frac{P(\mathcal{A})}{P(E)V_{\mathrm{A}}(\mu_m)} \int_{B_{\mathrm{A}}(\mu_m)} \tau_{\mathrm{A}}(\theta_{\mathrm{A}}) p(E \mid \theta_{\mathrm{A}})\, \mathrm{d}\theta_{\mathrm{A}}. \end{aligned}$$

The model $\mathcal{A}$ is selected if $P(\mathcal{A} \mid E) \geqslant P(\mathcal{P} \mid E)$, otherwise $\mathcal{P}$ is selected.

The prior probability $P(E)$ does not affect the selection of a model because it is the same for $\mathcal{A}$ and $\mathcal{P}$. In the absence of any further information about the correct model it is assumed that $P(\mathcal{A}) = P(\mathcal{P}) = 1/2$. Define $\chi(E)$, $\chi_{\mathrm{A}}(E)$ by

$$\begin{aligned} \chi(E) &= -\ln\biggl(V(\mu_m)^{-1} \int_{B(\mu_m)} \tau(\theta) p(E \mid \theta)\, \mathrm{d}\theta\biggr) \\ \chi_{\mathrm{A}}(E) &= -\ln\biggl(V_{\mathrm{A}}(\mu_m)^{-1} \int_{B_{\mathrm{A}}(\mu_m)} \tau_{\mathrm{A}}(\theta_{\mathrm{A}}) p(E \mid \theta_{\mathrm{A}})\, \mathrm{d}\theta_{\mathrm{A}}\biggr) \end{aligned}$$

The model $\mathcal{A}$ is selected if $\chi_{\mathrm{A}}(E) \leqslant \chi(E)$, otherwise $\mathcal{P}$ is selected.

(*c*) *Estimation of* $\chi_{\mathrm{A}}(E) - \chi(E)$

Let $p(\phi \mid t)$ be the true density from which the elements of E are sampled, and let f be the Taylor expansion of $\theta \mapsto -N^{-1}\ln(p(E \mid \theta))$ centred at t and truncated after the second order terms:

$$f(\theta) = a + b \cdot (\theta - t) + \tfrac{1}{2}(\theta - t)^{\mathrm{T}} B (\theta - t). \tag{6.2}$$

The minimum value of f is $a - \frac{1}{2}b^{\mathrm{T}}B^{-1}b$ and it is achieved at $\hat{\theta} = t - B^{-1}b$. It is assumed that f is a close approximation to $\theta \mapsto -N^{-1}\ln(p(E \mid \theta))$ and that $\hat{\theta}$ is a close approximation to the maximum-likelihood estimate $\operatorname{argmax}_\theta\{p(E \mid \theta)\}$ of t.

A rearrangement of the right-hand side of (6.2) yields

$$f(\theta) = a - \tfrac{1}{2}b^{\mathrm{T}}B^{-1}b + \tfrac{1}{2}(\theta - \hat{\theta})^{\mathrm{T}} B (\theta - \hat{\theta}).$$

It follows that as $N \to \infty$, $\chi(E)$ has the asymptotic expansion (Wong 1989)

$$\begin{aligned}\chi(E) &\sim N(a - \tfrac{1}{2}b^{\mathrm{T}}B^{-1}b) - \ln\biggl(V(\mu_m)^{-1}\int_{B(\mu_m)} \tau(\theta)\exp(-\tfrac{1}{2}N(\theta-\hat{\theta})^{\mathrm{T}}B(\theta-\hat{\theta}))\biggr) \\ &\sim N(a - \tfrac{1}{2}b^{\mathrm{T}}B^{-1}b) + \tfrac{1}{2}\ln(\det(B)) \\ &\qquad + \tfrac{3}{2}\ln\biggl(\frac{N}{2\pi}\biggr) - \ln(\tau(\hat{\theta})V(\mu_m)^{-1}) + O(N^{-1}). \qquad (6.3)\end{aligned}$$

Let $\hat{\theta}_{\mathrm{A}}$ be the point in $\mathrm{A}(2,\mathbb{R})$ at which the restriction of f to $\mathrm{A}(2,\mathbb{R})$ attains its minimum value. As in § 5 *b*, a point θ is in $\mathrm{A}(2,\mathbb{R})$ if and only if $\theta \cdot h = 0$, where $h = (0,1,0)^{\mathrm{T}}$. The point $\hat{\theta}_{\mathrm{A}} = (\hat{\mu}_{\mathrm{A}}, \hat{\beta}_{\mathrm{A}})$ is estimated by minimizing $f(\theta) + \lambda\theta\cdot h$, where λ is a Lagrange multiplier. The minimization yields

$$\left.\begin{aligned}(\hat{\mu}_{\mathrm{A}}, 0, \hat{\beta}_{\mathrm{A}}) &= \hat{\theta} - \lambda B^{-1}h, \\ f(\hat{\theta}_{\mathrm{A}}) &= a - 2^{-1}b^{\mathrm{T}}B^{-1}b + 2^{-1}\lambda^2 h^{\mathrm{T}}B^{-1}h.\end{aligned}\right\} \qquad (6.4)$$

Let B_{A} be the Hessian matrix of the restriction of f to $\mathrm{A}(2,\mathbb{R})$,

$$B_{\mathrm{A}} = \begin{pmatrix} B_{11} & B_{13} \\ B_{31} & B_{33}\end{pmatrix}.$$

It follows that as $N \to \infty$, $\chi_{\mathrm{A}}(E)$ has the asymptotic expansion

$$\begin{aligned}\chi_{\mathrm{A}}(E) &\sim N(a - \tfrac{1}{2}b^{\mathrm{T}}B^{-1}b + \tfrac{1}{2}\lambda^2h^{\mathrm{T}}B^{-1}h) + \tfrac{1}{2}\ln(\det(B_{\mathrm{A}})) \\ &\qquad + \ln\biggl(\frac{N}{2\pi}\biggr) - \ln(\tau(\hat{\theta}_{\mathrm{A}})V_{\mathrm{A}}(\mu_m)^{-1}) + O(N^{-1}). \qquad (6.5)\end{aligned}$$

It follows from (6.3) and (6.5) that

$$\begin{aligned}\chi_{\mathrm{A}}(E) - \chi(E) &\sim \tfrac{1}{2}N\lambda^2h^{\mathrm{T}}B^{-1}h \\ &\qquad + \ln\biggl(\frac{\sqrt{\det(B_{\mathrm{A}})}}{\tau_{\mathrm{A}}(\hat{\theta}_{\mathrm{A}})}\biggr) - \ln\biggl(\frac{\sqrt{\det(B)}}{\tau(\hat{\theta})}\biggr) \\ &\qquad - \tfrac{1}{2}\ln\biggl(\frac{N}{2\pi}\biggr) + \ln\biggl(\frac{V_{\mathrm{A}}(\mu_m)}{V(\mu_m)}\biggr) + O(N^{-1}).\end{aligned}$$

The matrix B is given by

$$B_{ij} = -\frac{1}{N}\sum_{k=1}^{N}\frac{\partial^2\ln(p(e_k \mid \theta))}{\partial\theta_i\partial\theta_j}\bigg|_{\theta=t}, \quad 1 \leqslant i,j \leqslant 3.$$

Let $t = (\mu,\alpha,\beta)$. It follows from the central limit theorem that with a high probability

$$\begin{aligned}B &= J(t) + O(N^{-1/2}), \\ B_{\mathrm{A}} &= J_{\mathrm{A}}(\mu,\beta) + O(N^{-1/2}).\end{aligned}$$

Let $\eta \sim \mathcal{N}(0,I)$ be a Gaussian random variable. The maximum-likelihood estimate $\hat{\theta}$ of t is to a first approximation a Gaussian random variable (Lindsey 1996):

$$\hat{\theta} = t + N^{-1/2}J(t)^{-1/2}\eta. \qquad (6.6)$$

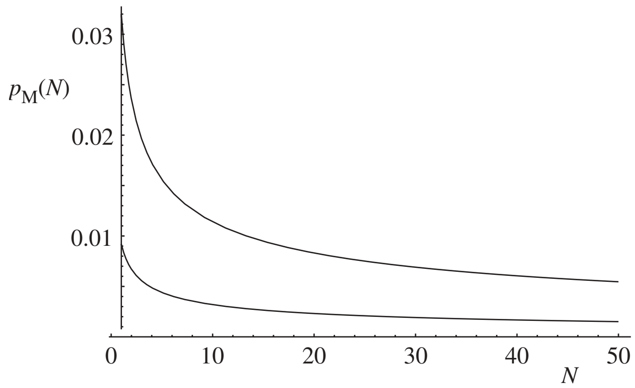

Figure 6. Graphs of $p_M(N)$ for $\sigma = 1/64$ (upper) and $\sigma = 1/256$ (lower).

It follows that with a high probability

$$\ln(\sqrt{\det(B)}/\tau(\hat{\theta})) = O(N^{-1/2}),$$
$$\ln(\sqrt{\det(B_A)}/\tau_A(\hat{\theta}_A)) = O(N^{-1/2});$$

thus

$$\chi_A(E) - \chi(E) \sim \tfrac{1}{2}N\lambda^2 h^T J(t)^{-1}h + \ln\left(\frac{V_A(\mu_m)}{V(\mu_m)}\right) - \tfrac{1}{2}\ln\left(\frac{N}{2\pi}\right) + O(N^{-1/2}). \quad (6.7)$$

It follows from (6.4) and (6.6) that

$$\lambda = \frac{\hat{\theta}\cdot h}{h^T B^{-1}h} = \left(\frac{t\cdot h + N^{-1/2}h^T J(t)^{-1}\eta}{h^T J(t)^{-1}h}\right)(1 + O(N^{-1/2})). \quad (6.8)$$

Let u be the unit vector defined by

$$u = (h^T J(t)^{-1}h)^{-1/2} J(t)^{-1/2}h.$$

It follows from (5.5), (6.7) and (6.8) that

$$\chi_A(E) - \chi(E) \sim \tfrac{1}{2}(N^{1/2}d(t,A) + u\cdot\eta)^2 - \tfrac{1}{2}\ln\left(\frac{N}{2\pi}\right) + \ln\left(\frac{V_A(\mu_m)}{V(\mu_m)}\right) + O(N^{-1/2}) \quad (6.9)$$

Author: is this variable or differential ‘d’?

(*d*) *Estimation of* $p_M(N, \mu_m)$

It is assumed that N is large enough to ensure that

$$NV(\mu_m)^2 \geqslant 2\pi V_A(\mu_m)^2. \quad (6.10)$$

It follows from (5.4) that (6.10) holds provided that $N \geqslant 8\sigma^2/\pi$, which is the case in most applications because $\sigma \ll 1$. Define ζ by

$$\zeta(N, \mu_m) = \sqrt{\ln\left(\frac{NV(\mu_m)^2}{2\pi V_A(\mu_m)^2}\right)}. \quad (6.11)$$

It follows from (6.9) that if N is sufficiently large, then to a good approximation $\mathcal{A}$ is selected if and only if

$$|N^{1/2}d(t,A) + u\cdot\eta| \leqslant \zeta(N, \mu_m).$$

The unit vector u is independent of η thus $u\cdot\eta$ has a normal distribution, $u\cdot\eta\sim\mathcal{N}(0,1)$. Let Φ be the cumulative distribution function for $\mathcal{N}(0,1)$. It follows from (6.9) and (6.11) that

$$P(C(\mathcal{A})\mid t)=\Phi(N^{1/2}d(t,\mathrm{A})+\zeta(N,\mu_m))-\Phi(N^{1/2}d(t,\mathrm{A})-\zeta(N,\mu_m)). \tag{6.12}$$

If $t\notin\mathrm{A}(2,\mathbb{R})$, then $P(C(\mathcal{A})\mid t)\to 0$ as $N\to\infty$. If $t\in\mathrm{A}(2,\mathbb{R})$, then $P(C(\mathcal{A})\mid t)\to 1$ as $N\to\infty$.

Let coordinates μ, z, β be chosen in a neighbourhood of $\mathrm{A}(2,\mathbb{R})$, where μ, β are as in § 2 *a*, and z is the signed distance from (μ,z,β) to $\mathrm{A}(2,\mathbb{R})$. If N is sufficiently large, then $P(C(\mathcal{A})\mid t)$ drops rapidly to zero as the z-component of t increases in magnitude. Near to $\mathrm{A}(2,\mathbb{R})$, $\tau(\theta)\,\mathrm{d}\theta$ is approximated by $\tau_{\mathrm{A}}(\mu,\beta)\,\mathrm{d}\mu\,\mathrm{d}\beta\,\mathrm{d}z$. The probability $p_{\mathrm{M}}(N)$ is thus approximated by

$$p_{\mathrm{M}}(N,\mu_m)=V(\mu_m)^{-1}\int_{\mu=0}^{\mu_m}\int_{\beta=0}^{\pi}\int_{z=-\infty}^{\infty}\tau_{\mathrm{A}}(\mu,\beta)P(C(\mathcal{A})\mid\theta)\,\mathrm{d}z\,\mathrm{d}\beta\,\mathrm{d}\mu. \tag{6.13}$$

The integral over z can be carried out exactly, using the expression (6.12) for $P(C(\mathcal{A})\mid t)$ and the fact that $d(t,A)=z$:

$$\int_{-\infty}^{\infty}P(C(\mathcal{A})\mid t)\,\mathrm{d}z=\frac{2\zeta(N,\mu_m)}{N^{1/2}}. \tag{6.14}$$

It follows from (6.13) and (6.14) that

$$\begin{aligned}p_{\mathrm{M}}(N,\mu_m)&=\frac{2\zeta(N,\mu_m)}{N^{1/2}V(\mu_m)}\int_0^{\mu_m}\int_0^{\pi}\tau_{\mathrm{A}}(\mu,\beta)\,\mathrm{d}\beta\,\mathrm{d}\mu,\\&=\frac{2V_{\mathrm{A}}(\mu_m)\zeta(N,\mu_m)}{N^{1/2}V(\mu_m)}.\end{aligned} \tag{6.15}$$

Numerical calculations, reinforced by the bounds (5.4), suggest that

$$\lim_{\mu_m\to\infty}\frac{V_{\mathrm{A}}(\mu_m)}{V(\mu_m)}=0.345\,081\cdots\sigma. \tag{6.16}$$

Let $p_{\mathrm{M}}(N)=\lim_{\mu_m\to\infty}p_{\mathrm{M}}(N,\mu_m)$. It follows from (6.15) and (6.16) that

$$p_{\mathrm{M}}(N)\approx\frac{0.690\,162\sigma}{N^{1/2}}\sqrt{\ln\biggl(\frac{1.336\,528N}{\sigma^2}\biggr)}. \tag{6.17}$$

Graphs of $p_{\mathrm{M}}(N)$ as a function of N are shown in figure 6 for noise levels $\sigma=1/256$ rad and $\sigma=1/64$ rad. These noise levels are of an order appropriate for the task of locating lines in images 512 × 512 pixels in size.

On examining the two graphs in figure 6, it appears that there is at these noise levels little benefit in using more than about $N=10$ points for classifying affine transformations and projective transformations. Beyond $N=10$, a large increase in the number of pairs of corresponding points produces only small decrease in $p_{\mathrm{M}}(N)$.

The data from the numerical example in § 4 are used to estimate $\chi_{\mathrm{A}}(E)-\chi(E)$, by setting $u\cdot\eta=0$, $t=\hat{\theta}=(1.106\,72,3.169\,56,1.560\,02)^{\mathrm{T}}$, $N=7$, $\sigma=1/50$ rad on the right-hand side of (6.9), and using the approximation (5.5) to $d(t,A)$. The value obtained is $\chi_{\mathrm{A}}(E)-\chi(E)\approx 614\,143$. This large positive value suggests that the Bayes rule strongly rejects the affine model. The result is consistent with the fact that the two images in figure 3 show a strong perspective distortion of the front face of the nearest building.

7. Conclusion

Closed form expressions have been obtained for the Fisher information and the Rao measure on both the group PSL$(2, \mathbb{R})$ of projective transformations and on the subgroup A$(2, \mathbb{R})$ of affine transformations. The Rao measures have been used to obtain an approximation to the probability, $p_{\mathrm{M}}(N)$, of misclassifying a projective transformation as an affine transformation. The approximation is a relatively simple function (6.17) of (i) the number N of pairs of corresponding points and (ii) the limiting ratio of the 2-volume of the affine transformations to the 3-volume of the projective transformations. It is only through the latter ratio that the standard deviation of the measurement noise contributes to the probability of misclassification.

OK?

The probabilistic model (2.9) used in the above calculations is a tractable simplification of the models used in practical estimation problems. It remains to be seen how the calculations are affected when more complicated probabilistic models are used.

I thank James Ferryman for permission to use images from the PETS2001 database, and the referees for their comments on an earlier draft of this paper.

References

Abramowitz, M. & Stegun, I. A. (eds) 1965 *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* New York: Dover.

Amari, S.-I. 1985 *Differential geometrical methods in statistics.* Lecture Notes in Statistics, vol. 28. Springer.

Balasubramanian, V. 1996 A geometric formulation of Occam's razor for inference of parametric distributions. Report No. PUPT-1588, Department of Physics, Princeton University, Princeton, NJ, USA. (Available from http://www.arxiv.org/list/nlin.AO/9601.)

Balasubramanian, V. 1997 Statistical inference, Occam's razor and statistical mechanics on the space of probability distributions. *Neural Comput.* **9**, 349–368.

Cover, T. M. & Thomas, J. A. 1991 *Elements of information theory.* Wiley.

Faugeras, O. D. 1993 *Three-dimensional computer vision.* Cambridge, MA: MIT Press.

Ferryman, J. M. 2001 PETS'2001 database. (Available at http://www.visualsurveillance.org/PETS2001.)

Fisher, R. A. 1922 On the mathematical foundations of theoretical statistics. *Phil. Trans. R. Soc. Lond.* A **222**, 309–368.

Gallot, S., Hulin, D. & Lafontaine, J. 1990 *Riemannian geometry.* Universitext, Springer.

Hartley, R. & Zisserman, A. 2000 *Multiple view geometry in computer vision.* Cambridge University Press.

Jeffreys, H. 1961 *Theory of probability.* Oxford: Clarendon Press.

Jost, J. 1995 *Riemannian geometry and geometric analysis.* Springer.

Kotz, S. & Johnson, N. L. (eds) 1992 *Breakthroughs in statistics. 1. Foundations and basic theory.* Springer Series in Statistics. Springer.

Lindsey, J. K. 1996 *Parametric statistical inference.* Oxford University Press.

Misner, C. W., Thorne, K. S. & Wheeler, J. A. 1973 *Gravitation.* San Francisco, CA: W. H. Freeman.

Myung, J., Balasubramanian, V. & Pitt, M. A. 2000 Counting probability distributions: differential geometry and model selection. *Proc. Natl Acad. Sci. USA* **97**, 11 170–11 175.

Rao, C. R. 1945 Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **37**, 81–91.

Semple, J. G. & Kneebone, G. T. 1952 *Algebraic projective geometry.* Oxford University Press.

Sonka, M., Hlavac, V. & Boyle, R. 1999 *Image processing, analysis and machine vision.* Boston, MA: PWS Publishing.

Thurston, W. P. 1997 *Three-dimensional geometry and topology.* Princeton, NJ: Princeton University Press.

Torr, P. H. S. & Zisserman, A. 1998 Concerning Bayesian motion segmentation, model averaging, matching and the trifocal tensor. In *Computer vision: ECCV'98* (ed. H. Burkhardt & B. Neumann). Lecture Notes in Computer Science, vol. 1406, pp. 511–527. Springer.

Torr, P. H. S., Dick, A. R. & Cipolla, R. 2000 Layer extraction with a Bayesian model of shapes. In *Computer vision: ECCV'2000* (ed. D. Vernon). Lecture Notes in Computer Science, vol. 1843, pp. 273–289. Springer.

Wolfram, S. 1999 *The* MATHEMATICA® *book*, 4th edn. Cambridge University Press.

Wong, R. 1989 *Asymptotic approximations of integrals.* Academic Press.