PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Rejecting Outliers and Estimating Errors in an Orthogonal-Regression Framework

Larry S. Shapiro and Michael Brady

Phil. Trans. R. Soc. Lond. A 1995 **350**, 407-439
doi: 10.1098/rsta.1995.0022

Rejecting outliers and estimating errors in an orthogonal-regression framework

By Larry S. Shapiro and Michael Brady

Robotics Research Group, Department of Engineering Science, Oxford University, Parks Road, Oxford OX1 3PJ, U.K.

Least squares minimization is by nature global and, hence, vulnerable to distortion by outliers. We present a novel technique to reject outliers from an m-dimensional data set when the underlying model is a hyperplane (a line in two dimensions, a plane in three dimensions). The technique has a sound statistical basis and assumes that Gaussian noise corrupts the otherwise valid data. The majority of alternative techniques available in the literature focus on *ordinary least squares*, where a single variable is designated to be dependent on all others – a model that is often unsuitable in practice. The method presented here operates in the more general framework of *orthogonal regression*, and uses a new regression diagnostic based on eigendecomposition. It subsumes the traditional residuals scheme and, using matrix perturbation theory, provides an error model for the solution once the contaminants have been removed.

Contents

Phil. Trans. R. Soc. Lond. A (1995) **350**, 407–439
Printed in Great Britain

1. Introduction

The problem of *outliers* (unrepresentative or 'rogue' observations) plagues data analysis techniques such as linear least squares regression in a wide variety of scientific investigations. This problem is well known in the statistics literature (Belsley *et al.* 1980; Cook & Weisberg 1982; Barnett & Lewis 1984; Hawkins 1980; Hampel *et al.* 1986; Huber 1981; Rousseeuw & Leroy 1987; Weisberg 1985), and arises when a given set of data actually comprises two subsets: a large dominant subset (the main body of valid data) and a relatively small subset of 'outliers' (the contaminants). The task of removing the contaminants is further complicated when, as is normally the case, the data in the dominant subset have also been perturbed by *noise*. The outlier problem is important since an analysis based both on the real data and the outliers distorts conclusions about the underlying process (figure 1). It is therefore of interest to seek a means of effectively rejecting such 'maverick' points, thereby restoring the propriety of the data and improving parameter estimation.

We examine this problem in the context of hyperplane fitting. Let the points $\{\boldsymbol{r}_i, i = 1, \ldots, n\}$ be given in $\mathbb{R}^m$ and let $\bar{\boldsymbol{r}}$ be their centroid. Then the points lie on an $(m-1)$-dimensional hyperplane π of $\mathbb{R}^m$ if, and only if, there exists a non-zero vector $\boldsymbol{n} \in \mathbb{R}^m$ such that

$$\boldsymbol{n}^{\perp}(\boldsymbol{r}_i - \bar{\boldsymbol{r}}) = 0, \quad i = 1, \ldots, n, \tag{1.1}$$

where $\boldsymbol{n}$ is the normal to the hyperplane. This paper addresses the case where the measurements $\boldsymbol{r}_i$ are contaminated by Gaussian noise and there is, in addition, a relatively small set of outliers. The aim is to identify and eliminate the outliers, in order to estimate the hyperplane by least squares fitting to the remaining data.

One means of achieving this is to use *regression diagnostics*, which involves calculating an initial fit to the data and then assessing the validity of each point based on a computed residual or influence measure. Such schemes are well-suited to our problem domain (see § 6); unfortunately, the vast majority of existing diagnostics apply to *ordinary least squares* (OLS), which is inappropriate for our purposes. The framework we require is *orthogonal regression* (OR), and the primary goal of this paper is to devise a suitable regression diagnostic. Our technique applies to general problems involving hyperplane fitting and relaxes some of the statistical assumptions imposed by many comparable schemes.

The paper is structured as follows. Section 2 introduces OLS and OR, and contrasts their relative strengths and weaknesses. Section 3 reviews existing solutions to outlier rejection and § 4 presents our new approach. Once the outliers have been rejected, it is important to know the remaining uncertainty in the fit (caused by the inevitable noise): § 5 derives the appropriate variance estimates and covariance matrices for this purpose. Section 6 applies our technique to a computer vision problem and § 7 concludes with directions for future research.

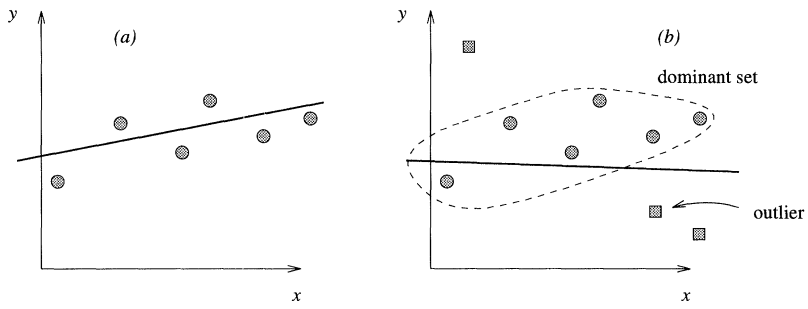


Figure 1. Noise and outliers: (*a*) a least squares fit minimizes the distances of the noise-perturbed data (circles) from the fitted line; (*b*) outliers (squares) distort the correct fit.

2. Linear regression

Regression is used to study relationships between measurable variables; *linear regression* deals with a particular class of relationships, namely those that can be described by straight lines or by their generalizations to many dimensions, typically called 'hyperplanes'. The objective here is to fit a hyperplane to a set of m-dimensional points, for instance a line in two dimensions or a plane in three dimensions.

Consider n data vectors in $\mathbb{R}^m$, denoted $\hat{\boldsymbol{r}}_i = (\hat{x}_{i1}, \hat{x}_{i2}, \ldots, \hat{x}_{im})^\top$, which satisfy the linear relation

$$\boldsymbol{n}^\top \hat{\boldsymbol{r}}_i + d = n_1\hat{x}_{i1} + n_2\hat{x}_{i2} + \cdots + n_m\hat{x}_{im} + d = 0, \quad i = 1 \ldots n. \tag{2.1}$$

The vector $\boldsymbol{n} = (n_1, n_2, \ldots, n_m) \in \mathbb{R}^m$ contains the parameters (or *regression coefficients*) to be estimated, and equation (2.1) has the following geometric interpretation:

(i) $\boldsymbol{n}$ is the direction of the normal to the $(m-1)$-dimensional hyperplane, with magnitude $|\boldsymbol{n}|$;

(ii) $|d|/|\boldsymbol{n}|$ is the perpendicular distance from the hyperplane to the origin.

Equation (2.1) has m unknowns since only the *ratios* of $n_1 : n_2 : \cdots : n_m : d$ can be recovered; thus, m data points are usually sufficient to determine $\boldsymbol{n}$ and d (up to an arbitrary scale factor). When more points are available ($n > m$), the system is overdetermined, and because noise makes it unlikely that all n points will lie precisely on the same hyperplane, an optimization problem arises. We initially ignore outliers and assume that each data point is perturbed by a noise vector $\delta\boldsymbol{r}_i \in \mathbb{R}^m$ to give the measurement $\boldsymbol{r}_i$, where

$$\boldsymbol{r}_i = \hat{\boldsymbol{r}}_i + \delta\boldsymbol{r}_i. \tag{2.2}$$

The perpendicular distance from $\boldsymbol{r}_i$ to the hyperplane π is then $\ell_i = (\boldsymbol{n}^\top \boldsymbol{r}_i + d)/|\boldsymbol{n}|$. Section 2 *a* justifies the use of OR to determine π and § 2 *b* describes the solution method.

(*a*) *Choice of objective function*

Harter (1974*a*) traces the method of OLS to Legendre & Gauss in the early nineteenth century (1805–1809). OLS computes the hyperplane π such that the sum of squared distances from the points to π is minimized in a particular direction. Figure 2 *a* shows the familiar two-dimensional case where a line is fitted to minimize the distances in the vertical direction (i.e. along the y-axis). The method of OR,

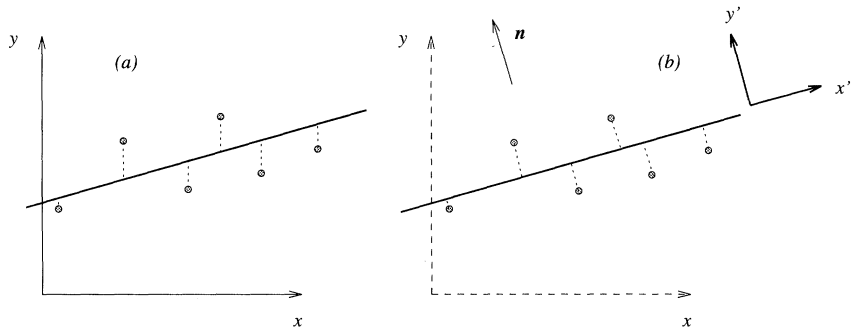


Figure 2. Comparison between OLS and OR: (*a*) in OLS, the minimization direction is specified to be the y-axis; (*b*) in OR, an intrinsic coordinate system (x', y') is computed and the minimization direction is automatically set to be the normal direction $\boldsymbol{n}$.

also termed *total least squares* or *principal component regression*, was proposed later by Adcock (1877, 1878), Kummel (1879) and Pearson (1901). OR minimizes the sum of squared distances *perpendicular* to the fitted hyperplane, i.e. along the direction of its normal $\boldsymbol{n}$ (figure 2*b*).

These approaches differ in two important ways. First, OLS requires the explicit definition of axes and a minimization direction, i.e. an external coordinate system is imposed on the data. In contrast, OR automatically computes an intrinsic coordinate system on the basis of the least squares criterion (figure 2*b*), the new axis variables being linear combinations of the originals. The OLS approach is more appropriate in a range of applications, not least in the social and behavioural sciences, where combinations of dissimilar variables (e.g. height and weight) yield meaningless quantities. However, when (as in our case) the variables are spatial coordinates, one coordinate system is as meaningful as any other, and the facility to select a natural reference frame independent of the external coordinate system is a distinct advantage of the OR method.

Second, the two methods make different assumptions about the error distribution of the variables. OLS only accommodates errors along the minimization axis (the *dependent* variable), and assumes that all remaining variables are independent and known accurately. Thus, in figure 2*a*, all x-components are treated as noise-free, and all errors are ascribed to the y-components. The choice of dependent variable affects the fit, as noted by Pearson (1901, p. 559): 'we get one straight line or plane if we treat some one variable as independent, and a quite different one if we treat another variable as the independent variable'. In contrast, OR caters for errors in *all* coordinate directions and simply seeks a functional relationship between the variables.

Thus, although the OLS formulae are simpler (and easier to solve), OR is generally better suited to the problem of hyperplane fitting, and the objective function we minimize is the sum of the squared perpendicular distances

$$\varepsilon(\boldsymbol{n}, d) = \sum_{i=1}^{n} \ell_i^2 = \sum_{i=1}^{n} (\boldsymbol{n}^\top \boldsymbol{r}_i + d)^2 / |\boldsymbol{n}|^2. \tag{2.3}$$

Least-squares fitting yields maximum-likelihood estimates of the parameters if the measurement errors $\delta \boldsymbol{r}_i$ are independent and follow a normal distribution with zero mean and common variance. Alternative objective functions that one

might minimize include (Berztiss 1964)

$$\varepsilon_{\mathrm{G}} = \sum_{i=1}^{n} |\ell_i| \quad \text{and} \quad \varepsilon_{\mathrm{C}} = \max_{1 \leqslant i \leqslant n} |\ell_i|. \tag{2.4}$$

The Gerschgorin norm, ε_{G}, sums the absolute errors. It is less sensitive to outliers than ε, performing better for long-tailed error distributions (Narula & Wellington 1982). However, ε_{G} is difficult to minimize, since discontinuities in the derivatives thwart general nonlinear equation solvers and function minimizers (Press *et al.* 1988); moreover, the solution is not guaranteed to be unique (Narula & Wellington 1982; Harter 1974*b*). The Chebyshev (or minimax) norm, ε_{C}, measures the maximum error, and whereas ε minimizes the average square error at the cost of potentially large deviations at some points, ε_{C} permits a larger average square error and keeps the maximum deviation to a minimum. This approach is even *less* robust than least squares (Rousseeuw & Leroy 1987) and ε_{C} is also hard to minimize.

Thus, with its straightforward unique closed-form solution, ε remains the preferred cost function. Indeed, Weisberg (1985) remarks that least squares estimation has been used for 180 years precisely because it is computationally simple, geometrically elegant and optimal in several important respects (given some assumptions).

(*b*) *Orthogonal regression*

We solve equation (2.3) for $\boldsymbol{n}$ and d by letting $\boldsymbol{n}$ be a *unit* vector ($|\boldsymbol{n}| = 1$) and minimizing

$$\varepsilon(\boldsymbol{n}, d) = \sum_{i=1}^{n} (\boldsymbol{n}^{\top} \boldsymbol{r}_i + d)^2 \quad \text{subject to} \quad |\boldsymbol{n}|^2 = 1. \tag{2.5}$$

The resulting hyperplane passes through the centroid $\bar{\boldsymbol{r}}$ (see appendix A), so

$$d = -\boldsymbol{n}^{\top} \bar{\boldsymbol{r}}. \tag{2.6}$$

We therefore eliminate d by first centring the data points, writing $\boldsymbol{v}_i = \boldsymbol{r}_i - \bar{\boldsymbol{r}}$:

$$\varepsilon(\boldsymbol{n}) = \sum_{i=1}^{n} (\boldsymbol{n}^{\top} \boldsymbol{v}_i)^2 = \boldsymbol{n}^{\top} \left(\sum_{i=1}^{n} \boldsymbol{v}_i \boldsymbol{v}_i^{\top} \right) \boldsymbol{n} = \boldsymbol{n}^{\top} \boldsymbol{W} \boldsymbol{n}. \tag{2.7}$$

The solution $\boldsymbol{n}$ is well known to be the unit eigenvector of $\boldsymbol{W}$ corresponding to the smallest eigenvalue (appendix A). The eigenvalues of $\boldsymbol{W}$ are the m roots of its characteristic polynomial $p(\lambda) = \det(\boldsymbol{W} - \lambda \boldsymbol{I})$, where $\boldsymbol{I}$ is the $m \times m$ identity matrix. These eigenvalues are denoted $\lambda(\boldsymbol{W}) = \{\lambda_1, \ldots, \lambda_m\}$ and arranged in non-decreasing order. If the corresponding normalized eigenvectors are $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$, then $\boldsymbol{n} = \boldsymbol{u}_1$. Since $\boldsymbol{W}$ is real and symmetric, these eigenvectors can form an orthonormal basis; furthermore, since $\boldsymbol{W}$ is also positive semi-definite ($\boldsymbol{v}_i^{\top} \boldsymbol{W} \boldsymbol{v}_i \geqslant 0$), the eigenvalues are all non-negative, i.e. $0 \leqslant \lambda_1 \leqslant \cdots \leqslant \lambda_m$.

The matrix $\boldsymbol{W}$, termed a scatter matrix, measures the dispersion of the data about the means in each of the m variables $(x_1, x_2, \ldots, x_m)$. In statistics parlance, we are performing *principal component regression* on the *covariance* matrix. (One might also use a *correlation* matrix, where the data points are additionally normalized in terms of variance; this is unnecessary when, as in our case, the variables are comparable in magnitude of variance and units of measurement (Krazanowski 1988).) The algorithm is summarized as follows.

Task 1. *Given* $\boldsymbol{r}_i \in \mathbb{R}^m$ $(i = 1, \ldots, n)$, *where* $n \geqslant m$, *compute* $\boldsymbol{n}$ *and* d *to minimize*

$$\varepsilon(\boldsymbol{n}, d) = \sum_{i=1}^{n} (\boldsymbol{n}^\top \boldsymbol{r}_i + d)^2, \quad \text{where } |\boldsymbol{n}| = 1.$$

Algorithm 1.

(i) Compute the data centroid $\bar{\boldsymbol{r}}$ *and centre the points, writing* $\boldsymbol{v}_i = \boldsymbol{r}_i - \bar{\boldsymbol{r}}$.

(ii) Construct the scatter matrix $\boldsymbol{W} = \sum_{i=1}^{n} \boldsymbol{v}_i \boldsymbol{v}_i^\top$.

(iii) Find the unit eigenvector $\boldsymbol{n}$ *corresponding to the minimum eigenvalue of* $\boldsymbol{W}$ $(\boldsymbol{W}\boldsymbol{n} = \lambda_1 \boldsymbol{n})$.

(iv) Calculate $d = -\boldsymbol{n}^\top \bar{\boldsymbol{r}}$.

Geometry provides useful insight into how OR works. As mentioned earlier, OR finds an optimal set of axes (an 'intrinsic coordinate system') to describe the data, where 'optimal' refers to the best summarization of the data. The 'best' axis is the line which the cloud of points is closest to in Euclidean space, i.e. the line onto which the projections of the points have maximum variance (figure 3). For instance, if all the points lie on a single line, that line is the most descriptive axis and no further axes are needed. The second best-fitting axis (perpendicular to the first) defines the best-fitting plane, the third best-fitting axis (perpendicular to the first two) defines the best-fitting three-dimensional hyperplane, and so on until all m dimensions have been explained. The new axes are the *eigenvectors* and the variances of the projections onto these axes are the *eigenvalues*, specifying the relative order of importance (or 'explaining power') of the various axes.

In fitting the best hyperplane, our interest lies in the axis with smallest eigenvalue, i.e. the axis onto which the projections of the data points have *minimum* variance. These projections are the *residuals* of the fit, i.e. the residual for the ith point is the projection of $\boldsymbol{v}_i$ onto the unit axis $\boldsymbol{u}_1$,

$$\ell_i = \boldsymbol{u}_1^\top \boldsymbol{v}_i, \tag{2.8}$$

which gives the perpendicular distance from $\boldsymbol{v}_i$ to π. The residuals would be zero in the absence of noise, since the data set only has $m-1$ independent dimensions. The *residual vector* $\boldsymbol{\ell} = (\ell_1, \ell_2, \ldots, \ell_n)^\top$ is defined as

$$\boldsymbol{\ell}^\top = \boldsymbol{u}_1^\top \boldsymbol{V}, \tag{2.9}$$

where $\boldsymbol{V} = [\boldsymbol{v}_1 | \boldsymbol{v}_2 | \cdots | \boldsymbol{v}_n]$. Note that ε_{G} and ε_{C} in equation (2.4) are simply different norms of $\boldsymbol{\ell}$ (Berztiss 1964), namely $\varepsilon_{\mathrm{G}} = \|\boldsymbol{\ell}\|_1$ and $\varepsilon_{\mathrm{C}} = \|\boldsymbol{\ell}\|_\infty$ (see appendix B).

Although algorithm 1 caters for Gaussian noise in an optimal way, the problem of *outliers* remains; least squares estimation is *global* and outliers distort the solution. We address this problem in the sections that follow.

3. Previous work on outlier rejection

Two main approaches to the outlier problem have evolved in the form of *regression diagnostics* (Belsley *et al.* 1980; Cook & Weisberg 1982; Barnett & Lewis 1984; Hawkins 1980) and *robust statistics* (Hampel *et al.* 1986; Huber 1981; Rousseeuw & Leroy 1987). The diagnostic methods compute an initial fit to the data, pinpoint outliers, reject them and then reanalyse the remaining data (possibly in an iterative process). Thus, they simultaneously 'build and criticize' the model (Myers 1990). In contrast, the robust statistics methods first find a fit

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

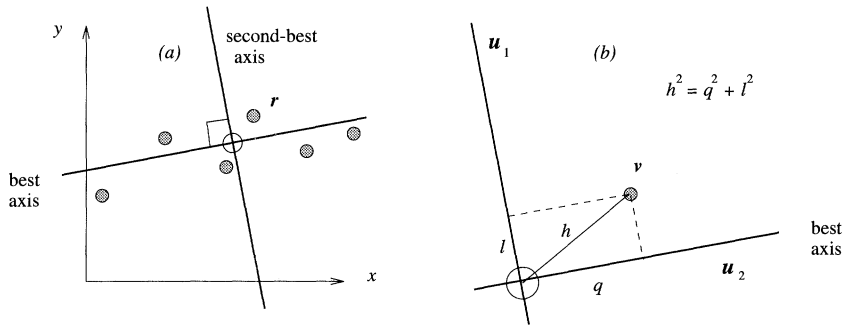


Figure 3. Orthogonal regression: (*a*) the best axis passes closest to the points in terms of perpendicular distance, i.e. the projections of the points onto this axis have maximum variance; (*b*) maximizing the projection of $\boldsymbol{v}$ onto $\boldsymbol{u}_2$ (length q) is equivalent to minimizing the projection onto $\boldsymbol{u}_1$ (length ℓ), since the distance h from $\boldsymbol{v}$ to the centroid O is fixed, and $h^2 = q^2 + \ell^2$.

to explain the majority of the data without removing the contaminants. The outliers are then identified (if needed) as those points which are inconsistent with the dominant fit. The robust approach is said to 'accommodate' the outliers. In some applications, the two schemes yield identical results; in others, they differ significantly, and several papers have debated their relative merits and shortcomings (see Rousseeuw & Leroy 1987).

(*a*) *Regression diagnostics*

The classical least squares approach to outlier rejection computes the initial fit, determines the residual for each data point and rejects all points whose residuals exceed a predetermined threshold (based, say, on a chosen confidence level and a prior statistical-noise model). The procedure is then repeated with the reduced set of points until all outliers have been removed. This approach works well when the percentage of outliers is small and their deviations from the valid data are not too large; unfortunately, a single outlier far removed from the data centroid can strongly distort the fit, yet still have a very small residual.

A refinement of the above scheme uses *influence measures* to pinpoint potential outliers. These measures assess the extent to which a particular point influences the fit by determining the change in the solution when that point is omitted. Examples include Cook's D distance (Cook & Weisberg 1982; Weisberg 1985) and the DFFITS/DFBETAS statistics (Belsley *et al.* 1980), which measure the effect of point deletion on various regression parameters. Several such measures were evaluated by Torr & Murray (1993*a*); unfortunately, all were designed for the ordinary least squares formulation.

We therefore formulate our diagnostic directly in terms of the *eigensolution*. Much of the previous work in this area revolves around principal component regression (PCR), and the proposed solutions generally have an *ad hoc* intuitively justified basis, with little sound statistical foundation (Barnett & Lewis 1984). In general, a p-dimensional data point is transformed into a different p-dimensional point (its principal component vector) by projecting the original data point onto each of the p new principal component axes. Gnanadesikan & Kettenring (1972) surveyed the field and suggested highlighting different types of outlier by using the first few and last few principal component vectors of the data, the former being sensitive to outliers inflating variances/covariances, and the latter to outliers adding spurious dimensions to the data. They proposed no formal tests; instead, they recommended graphical methods (such as bivariate plotting of the different components) to elicit putative contaminants.

Hawkins (1980) used two formal test statistics; however, his tests (like many others) require the strong assumption that the data points are a random sample from a *multivariate normal distribution*. When the underlying data distribution is unknown (as in most cases, including our own), his theory only holds asymptotically as the number of points $n \to \infty$, so that the central-limit theorem can be invoked (yielding an approximately normal distribution of principal components). He pointed out that, in practice, this requirement on n can be prohibitively large. He concluded that despite the appealing properties of tests involving principal component residuals, they are not generally valid for formal testing with controlled probability of type I error since the underlying (null) data distribution is unknown for n of small to moderate size. (Type I errors occur when one incorrectly rejects a valid point as an outlier.) Chatfield & Collins (1980) concurred that the available sampling theory for PCR is of limited use (even under the assumption of multivariate normality), and went so far as to suggest that PCR be viewed solely as 'a mathematical technique with no underlying statistical model'.

We address these concerns in § 4 *e* and illustrate how our particular formulation (which uses only the minimum variance principal component) overcomes them for our class of problem.

(*b*) *Robust statistics*

As mentioned in § 2 *a*, alternatives to the least squares estimator can be found, and concern for sensitivity to outliers spawned a search for 'robust' estimators that would better tolerate the perturbations (Huber 1981; Rousseeuw & Leroy 1987). In this context, 'robust' means 'insensitive to small departures from the ideal assumptions for which the estimator was optimized', often implying large departures for a small number of points. One class of robust estimators is the 'maximum-likelihood type' or M-estimators (Huber 1981), which include the least-absolute-deviation estimator.

Torr & Murray (1993*b*) suggested that robust estimators are appropriate when the number of outliers is large or when the outliers possess structure. Their least median-squares solution repeatedly sampled the universal set of data, computing a statistic from each subset and averaging the results of many such 'trials' (via robust statistical methods) to compute the dominant fit. The robustness of this method stems from the fact that it only considers a subset of the universal data set at any one time. Although this approach is reliable and applies equally to the OLS and OR frameworks, it has a significant computational overhead. Moreover, since in many cases (including our own) the outliers constitute a small percentage of the data and their maximum deviation is bounded, the least squares methods suffice.

4. Outlier rejection techniques

We examine two methods of identifying outliers. The first, a novel approach, computes the improvement in the minimum eigenvalue of the scatter matrix $\boldsymbol{W}$ when a data point is deleted. The second, the traditional residuals method, serves as a benchmark to evaluate the performance of the first method. We show that the residuals method is in fact *subsumed* by the minimum-eigenvalue method, the results coinciding when a first-order perturbation model is used in the minimum-eigenvalue scheme.

(*a*) *The minimum-eigenvalue method*

The intuition behind this approach is as follows. The minimum eigenvalue λ_1 of the scatter matrix $\boldsymbol{W}$ measures the total error in the fit; if this error is statistically

significant (i.e. partly due to outliers rather than pure random noise), then the point whose removal most decreases the error is identified as an outlier, and deleted. The parameters are then recomputed for the reduced set of points, and the process continues until the termination criteria are satisfied. Section 4 *a* (i) outlines the basic algorithm, and § 4 *a* (ii) discusses ways to improve its efficiency.

(i) *Basic algorithm*

Consider the effect of deleting the ith data point. The data centroid changes from $\bar{\boldsymbol{r}}$ to $\bar{\boldsymbol{r}}^\star$, and the points $\boldsymbol{v}_j$ $(j = 1 \ldots n, j \neq i)$ acquire new coordinates $\boldsymbol{v}_j^\star$. The scatter matrix is modified from

$$\boldsymbol{W} = \sum_{j=1}^{n} (\boldsymbol{r}_j - \bar{\boldsymbol{r}})(\boldsymbol{r}_j - \bar{\boldsymbol{r}})^\top$$

(summed over n points) to

$$\boldsymbol{W}^\star = \sum_{j=1, j\neq i}^{n} (\boldsymbol{r}_j - \bar{\boldsymbol{r}}^\star)(\boldsymbol{r}_j - \bar{\boldsymbol{r}}^\star)^\top$$

(summed over $n-1$ points). It is straightforward to show that the new quantities are:

$$\boldsymbol{W}^\star = \boldsymbol{W} - \frac{n}{n-1}(\boldsymbol{r}_i - \bar{\boldsymbol{r}})(\boldsymbol{r}_i - \bar{\boldsymbol{r}})^\top = \boldsymbol{W} - \frac{n}{n-1}\boldsymbol{v}_i\, \boldsymbol{v}_i^\top, \tag{4.1}$$

$$\bar{\boldsymbol{r}}^\star = \bar{\boldsymbol{r}} - \frac{1}{n-1}(\boldsymbol{r}_i - \bar{\boldsymbol{r}}) = \bar{\boldsymbol{r}} - \frac{1}{n-1}\boldsymbol{v}_i, \tag{4.2}$$

$$\boldsymbol{v}_j^\star = \boldsymbol{v}_j + \frac{1}{n-1}\boldsymbol{v}_i\ . \tag{4.3}$$

The eigensolution also changes when $\boldsymbol{W}$ is perturbed to $\boldsymbol{W}^\star$. We now identify λ_1 as an appropriate statistical variable for assessing the improvement in the fit when a point is deleted. By definition, $\boldsymbol{W}\boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1$, so

$$\lambda_1 = \boldsymbol{u}_1^\top \boldsymbol{W}\boldsymbol{u}_1 = \sum_{i=1}^{n} (\boldsymbol{u}_1^\top \boldsymbol{v}_i)^2 = \sum_{i=1}^{n} \ell_i^2.$$

Thus, λ_1 equals the sum of the squared distances ℓ_i^2 between the data points and the fitted hyperplane π. If the residuals ℓ_i are independent random variables drawn from a zero-mean Gaussian distribution with variance σ_ℓ^2, then λ_1/σ_ℓ^2 is distributed as χ^2 with $(n-m)$ degrees of freedom (see, for instance, Porrill *et al.* 1986). Expressions for the variances and covariances of ℓ_i are derived in § 5 *b*, where it is shown that the assumption of a univariate distribution on ℓ is not strictly true; the residual variances differ slightly from point to point, and are also correlated. However, we show these effects to be minor, and express σ_ℓ^2 as $(n-1)\sigma^2/n$, where σ^2 is the variance of the zero-mean independent isotropic Gaussian noise in the original points $\boldsymbol{r}_i$.

A one-tailed significance test can therefore be performed on λ_1/σ_ℓ^2, corresponding to the (null) hypothesis that π explains the data to a predetermined confidence level. If, for instance, $\lambda_1/\sigma_\ell^2 < \chi^2_{0.95}$ (95% confidence level), there are deemed to be no outliers to the fit; otherwise, we delete that point whose removal maximally reduces λ_1, and decrement the degrees of freedom on the χ^2 variable by one. A new fit is then computed using the remaining data, and the data centroid and scatter matrix are updated for the next iteration. This process continues until λ_1

falls within the specified confidence interval (indicating that no outliers remain), or until a specified number of iterations/outliers has been reached (m is a lower bound on the number of retained points).

Task 2. *Given $\boldsymbol{r}_i \in \mathbb{R}^m$ $(i = 1 \ldots n,\ n \geqslant m)$ perturbed by zero-mean independent isotropic Gaussian noise with variance σ^2, reject outliers from the set $\{\boldsymbol{r}_i\}$ to a $\zeta\%$ confidence level.*

Algorithm 2.

(i) Compute the data centroid $\bar{\boldsymbol{r}}$ and centre the points, writing $\boldsymbol{v}_i = \boldsymbol{r}_i - \bar{\boldsymbol{r}}$. Construct the scatter matrix $\boldsymbol{W} = \sum_{i=1}^{n} \boldsymbol{v}_i \boldsymbol{v}_i^{\top}$.

(ii) Find the eigenvector $\boldsymbol{u}_1$ corresponding to the minimum eigenvalue λ_1 of $\boldsymbol{W}$.

(iii) Perform a one-tailed χ^2 significance test on λ_1/σ_ℓ^2, with $\sigma_\ell^2 = (n-1)\sigma^2/n$; if the test falls within the $\zeta\%$ confidence bound, set $\boldsymbol{n} = \boldsymbol{u}_1$ and goto step (vi).

(iv) For each point i, delete it from the data set and compute the new minimum eigenvalue, $\lambda_1^\star(i)$. Hence determine the change in λ_1 when point i is deleted, namely $\Delta\lambda_1(i) = \lambda_1 - \lambda_1^\star(i)$.

(v) Delete the point i for which $\Delta\lambda_1(i)$ is greatest. Update the scatter matrix and centroid using equations (4.1) and (4.2), and return to step (ii).

(vi) Calculate $d = -\boldsymbol{n}^{\top}\bar{\boldsymbol{r}}$. The final hyperplane parameters are $\{\boldsymbol{n}, d\}$.

The full eigensolution is recomputed to determine the change in λ_1 when point i is removed, so to remove k outliers, $k(2n - k + 1)/2$ eigendecompositions (of an $m \times m$ matrix) are required. Ways of reducing this computational cost are discussed in § 4 *a* (ii). We do not delete more than one outlier at a time since the new fit may yield substantially different influence values for the remaining data. While this might appear cautious, this caution is justified by the example in § 4 *d*.

(ii) *Efficiency considerations*

To determine the point with maximum influence, algorithm 2 recomputes the full eigensolution for every point at each iteration. This is inefficient, and we present two techniques to redress this problem. In both cases, complete eigendecomposition is required only at the beginning of the calculation and after the deletion of each outlier. The removal of k outliers therefore involves only $k + 1$ eigendecompositions, substantially reducing the previous computational cost.

The first technique uses matrix-perturbation theory and the second uses an exact eigenvalue identity; the latter is shown to be more suitable, and is thus adopted for the remainder of this paper. Both methods express equation (4.1) as $\boldsymbol{W}^\star = \boldsymbol{W} + \Delta\boldsymbol{W}_i$, where $\Delta\boldsymbol{W}_i = -n\boldsymbol{v}_i\boldsymbol{v}_i^{\top}/(n-1)$ corresponds to removal of point i. Note that $\Delta\boldsymbol{W}$ is a dyadic product and hence a real-symmetric matrix of rank one.

Perturbation model. This approach uses *matrix-perturbation theory* (Golub & van Loan 1989; Wilkinson 1965) to evaluate the change in λ_1 induced by deleting point i. Various worst-case bounds for eigenvalue perturbation exist in the literature (e.g. Weyl theorem, Wielandt–Hoffman theorem), derived mainly from Gerschgorin disk theory (Stewart & Sun 1990; Wilkinson 1965). These bounds impose restrictions on the maximum variation of the eigenvalues; however, while useful for devising numerical algorithms, their worth is limited by their conservatism (Weng *et al.* 1989).

We use instead a Taylor series expansion to compute the first- and second-order eigenvalue variations. Consider a perturbation that is proportional to ϵ in

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

the scatter matrix, and let $\lambda_1(\epsilon)$ denote the dependence of the least eigenvalue on ϵ, where $\lambda_1 = \lambda_1(0)$ and

$$\lambda_1(\epsilon) = \lambda_1(0) + \dot{\lambda}_1(0)\epsilon + \tfrac{1}{2}\ddot{\lambda}_1(0)\epsilon^2 + O(\epsilon^3).$$

The quantity ϵ is the two-norm of $\Delta \boldsymbol{W}$, defined formally in appendix B, where it is also shown that the first- and second-order perturbation terms in λ_1 due to deleting point i are

$$\dot{\lambda}_1(0)\epsilon = -\frac{n}{n-1}(\boldsymbol{u}_1^\top \boldsymbol{v}_i)^2, \tag{4.4}$$

$$\tfrac{1}{2}\ddot{\lambda}_1(0)\epsilon^2 = -\frac{n^2}{(n-1)^2}(\boldsymbol{u}_1^\top \boldsymbol{v}_i)^2 \sum_{k=2}^{m} \frac{(\boldsymbol{u}_k^\top \boldsymbol{v}_i)^2}{\lambda_k - \lambda_1}. \tag{4.5}$$

This scheme is incorporated into algorithm 2 by replacing step (iv) as follows:

Step (iv). *For every point i, compute the perturbation in λ_1 (up to second order) when that point is deleted, namely*

$$\Delta\lambda_1(i) = \frac{n^2}{(n-1)^2}(\boldsymbol{u}_1^\top \boldsymbol{v}_i)^2\left[\frac{n-1}{n} + \sum_{k=2}^{m}\frac{(\boldsymbol{u}_k^\top \boldsymbol{v}_i)^2}{\lambda_k - \lambda_1}\right].$$

Once it has been decided which point to remove, the accurate eigensolution is computed to prevent errors accumulating over time. Thus, while perturbation theory is used to speed up the search for the outlier, the solution is calculated accurately once the point is actually removed.

Unfortunately, the use of this approximation may sometimes identify an outlier incorrectly, though improvements can be made to the above algorithm to reduce the chance of this occurring. For instance, one could use higher-order perturbations, or compute the accurate solutions for several of the points with large perturbations (ensuring that the worst offender is pinpointed correctly). A better approach is to use the eigenvalue identity described in the following subsection, which computes the new eigenvalues precisely. Indeed, the main value of the above perturbation analysis is the insight it provides into the operation of the outlier rejection method (elegantly demonstrating in § 4 *c* how the minimum eigenvalue and residual methods are related), and the tractability of the noise analysis it facilitates in § 5.

Eigenvalue identity. Let $\boldsymbol{D}$ be an $m \times m$ non-singular matrix and let $\boldsymbol{D} - c\boldsymbol{v}\boldsymbol{v}^\top$ be singular, where c is a non-zero scalar and $\boldsymbol{v}$ is a non-zero m-vector. Then the vector $\boldsymbol{D}^{-1}\boldsymbol{v}$ lies in the null space of the singular matrix. (The proof is simple: a non-zero vector $\boldsymbol{e}$ lying in this null space satisfies $(\boldsymbol{D} - c\boldsymbol{v}\boldsymbol{v}^\top)\boldsymbol{e} = \boldsymbol{0}$, giving $\boldsymbol{e} = c(\boldsymbol{e}^\top \boldsymbol{v})\boldsymbol{D}^{-1}\boldsymbol{v}$, which is parallel to $\boldsymbol{D}^{-1}\boldsymbol{v}$.) Thus, if $\lambda^\star$ is an eigenvalue of $\boldsymbol{D} - c\boldsymbol{v}\boldsymbol{v}^\top$ with associated eigenvector $\boldsymbol{u}^\star$, then $(\boldsymbol{D} - \lambda^\star \boldsymbol{I} - c\boldsymbol{v}\boldsymbol{v}^\top)\boldsymbol{u}^\star = \boldsymbol{0}$ and the eigenvector has direction $(\boldsymbol{D} - \lambda^\star \boldsymbol{I})^{-1}\boldsymbol{v}$. The eigenvalue therefore satisfies the relation

$$(\boldsymbol{D} - \lambda^\star \boldsymbol{I} - c\boldsymbol{v}\boldsymbol{v}^\top)(\boldsymbol{D} - \lambda^\star \boldsymbol{I})^{-1}\boldsymbol{v} = \boldsymbol{v} - c\boldsymbol{v}\boldsymbol{v}^\top(\boldsymbol{D} - \lambda^\star \boldsymbol{I})^{-1}\boldsymbol{v} = \boldsymbol{0},$$

yielding the identity (Golub 1973)

$$c\boldsymbol{v}^\top(\boldsymbol{D} - \lambda^\star \boldsymbol{I})^{-1}\boldsymbol{v} = 1.$$

In the notation of equation (4.1), we have $\boldsymbol{D} = \boldsymbol{W}$, $\boldsymbol{v} = \boldsymbol{v}_i$ and $c = n/(n-1)$, so

$\lambda^\star$ is an eigenvalue of $\boldsymbol{W}^\star$ if

$$\boldsymbol{v}_i^\top (\boldsymbol{W} - \lambda^\star \boldsymbol{I})^{-1} \boldsymbol{v}_i = \frac{n-1}{n}. \tag{4.6}$$

Now the eigensolution for $\boldsymbol{W}$ is known to be $\boldsymbol{W}\boldsymbol{u}_k = \lambda_k \boldsymbol{u}_k$, where

$$\boldsymbol{W} = \sum_{k=1}^{m} \lambda_k \boldsymbol{u}_k \boldsymbol{u}_k^\top \quad \text{and} \quad \boldsymbol{I} = \sum_{k=1}^{m} \boldsymbol{u}_k \boldsymbol{u}_k^\top.$$

Thus,

$$\boldsymbol{W} - \lambda^\star \boldsymbol{I} = \sum_{k=1}^{m} (\lambda_k - \lambda^\star) \boldsymbol{u}_k \boldsymbol{u}_k^\top \quad \text{and} \quad (\boldsymbol{W} - \lambda^\star \boldsymbol{I})^{-1} = \sum_{k=1}^{m} \frac{\boldsymbol{u}_k \boldsymbol{u}_k^\top}{\lambda_k - \lambda^\star}.$$

Equation (4.6) therefore reduces to

$$\sum_{k=1}^{m} \frac{(\boldsymbol{u}_k^\top \boldsymbol{v}_i)^2}{\lambda_k - \lambda^\star} = \frac{n-1}{n}. \tag{4.7}$$

This equation has one root $\lambda^\star$ such that $\lambda^\star < \lambda_1$ if $\boldsymbol{u}_1^\top \boldsymbol{v}_i$ is non-zero. (If $\boldsymbol{u}_1^\top \boldsymbol{v}_i = 0$, then $\{\lambda_1, \boldsymbol{u}_1\}$ is an eigensolution of both $\boldsymbol{W}$ and $\boldsymbol{W}^\star$.) This suggests an efficient way to check whether deleting point i causes a bigger change in λ_1 than deleting one of the other points that have already been examined. Let $\underline{\lambda_1^\star}$ be the smallest of the minimum eigenvalues calculated so far, that is $\underline{\lambda_1^\star} = \min\{\lambda_1^\star(1), \lambda_1^\star(2), \ldots, \lambda_1^\star(i-1)\}$. Then the minimum eigenvalue for point i, $\lambda_1^\star(i)$, will be smaller than $\underline{\lambda_1^\star}$ if

$$\sum_{k=1}^{m} \frac{(\boldsymbol{u}_k^\top \boldsymbol{v}_i)^2}{\lambda_k - \underline{\lambda_1^\star}} > \frac{n-1}{n},$$

in which case point i becomes the one favoured for deletion. The true value of $\lambda_1^\star(i)$ can then be computed to high accuracy, because on average this will only be done on relatively few occasions. The modification to step (iv) of the algorithm is given below.

Step (iv). *Delete point 1 from the data set, compute the new minimum eigenvalue* $\lambda_1^\star(1)$, *and set* $\underline{\lambda_1^\star} = \lambda_1^\star(1)$. *For each remaining point* i, *where* $i = 2, \ldots, n$,

$$\textit{if} \sum_{k=1}^{m} \frac{(\boldsymbol{u}_k^\top \boldsymbol{v}_i)^2}{\lambda_k - \underline{\lambda_1^\star}} > \frac{n-1}{n},$$

(*a*) *delete point* i *from the data set and compute the new minimum eigenvalue* $\lambda_1^\star(1)$, *and*
(*b*) *set* $\underline{\lambda_1^\star} = \lambda_1^\star(i)$ *and* $\Delta\lambda_1(i) = \lambda_1 - \lambda_1^\star(i)$.

Importantly, computing this true value $\lambda_1^\star(i)$ does *not* require eigendecomposition; it can be done using a modified Newton–Raphson procedure on equation (4.7). Indeed, for large n, the work of such a method would be less than the work of generating all the scalar products $\boldsymbol{u}_k^\top \boldsymbol{v}_i$, which will have been performed already. Thus, as in the perturbation method above, only one complete eigendecomposition is needed for the deletion of each outlier; however, the eigenvalue-identity method has the advantage of computing the *exact* value of λ_1, not just an approximation to it.

(*b*) *Method of residuals*

We now turn to the second method of identifying outliers, namely the method of residuals. Recall equation (2.9), $\boldsymbol{\ell}^\top = \boldsymbol{u}_1^\top \boldsymbol{V}$, where $\boldsymbol{\ell} = (\ell_1, \ell_2, \ldots, \ell_n)^\top$. Clearly, ℓ_i is a linear combination of random Gaussian variables and is thus itself a Gaussian variable, with $E\{\ell_i\} = 0$ and $\mathrm{Var}\{\ell_i\} = \sigma_\ell^2$ (σ_ℓ is derived in § 5 *b*). The elements of $\boldsymbol{\ell}$ thus form a Gaussian distribution, and standard statistical tests can be performed on them to identify outliers. Since the residuals can be positive or negative ('behind' or 'in front of' the hyperplane), a two-tailed test is needed, such as $-1.96 \leqslant \ell_i/\sigma_\ell \leqslant 1.96$ for a 95% confidence level. The algorithm is summarized as follows.

Task 3. *Given $\boldsymbol{r}_i \in \mathbb{R}^m$ ($i = 1, \ldots, n$, $n \geqslant m$), with $\boldsymbol{r}_i$ perturbed by zero-mean independent isotropic Gaussian noise having variance σ^2, reject outliers from the set $\{\boldsymbol{r}_i\}$ to a ζ% confidence level.*

Algorithm 3.

(i) Compute the data centroid $\bar{\boldsymbol{r}}$ and centre the points, writing $\boldsymbol{v}_i = \boldsymbol{r}_i - \bar{\boldsymbol{r}}$. Construct the scatter matrix $\boldsymbol{W} = \sum_{i=1}^n \boldsymbol{v}_i \boldsymbol{v}_i^\top$.

(ii) Find the eigenvector $\boldsymbol{u}_1$ corresponding to the minimum eigenvalue λ_1 of $\boldsymbol{W}$, and compute the residual vector $\boldsymbol{\ell} = \boldsymbol{u}_1^\top \boldsymbol{V}$.

(iii) Find the maximum residual $\ell_{\max} = \max_{1 \leqslant i \leqslant n} |\ell_i|$, and perform a two-tailed Gaussian significance test on $\ell_{\max}/\sigma_\ell$. If this is within acceptable limits, set $\boldsymbol{n} = \boldsymbol{u}_1$ and goto step (v).

(iv) Delete the point with maximum residual. Update the scatter matrix and centroid according to equations (4.1) and (4.2), and return to step (ii).

(v) Calculate $d = -\boldsymbol{n}^\top \bar{\boldsymbol{r}}$. The final hyperplane parameters are $\{\boldsymbol{n}, d\}$.

(*c*) *Comparison between methods*

The two outlier rejection schemes in § 4 *a* and § 4 *b* share many similarities, and the following analogy can be made. Consider n samples $\{t_1, t_2, \ldots, t_n\}$ drawn from a univariate population with mean $\bar{t}$ and variance σ_t^2. The statistic

$$T^2 = \frac{(t_1 - \bar{t})^2 + (t_2 - \bar{t})^2 + \cdots + (t_n - \bar{t})^2}{\sigma_t^2}$$

has a χ^2 distribution if the underlying population is Gaussian. Then $(t_i - \bar{t})/\sigma_t$ is the 'residual' (individual distance) and T^2 is the 'eigenvalue' (sum of the squared distances). A large residual causes a corresponding increase in the eigenvalue, and the reliability of both schemes hinges on the validity of the Gaussian distribution assumption and the assumed value of σ_t. The difference is that whereas the residual scheme simply rejects the point that deviates most from the current fit, the influence function rejects the point whose exclusion will result in the best fit on the *next* iteration. Put differently, the residual scheme looks only at the *existing* fit to identify the 'villain', while the influence fit 'looks ahead' to the *next* fit to see what improvements will actually materialize.

The two schemes often agree on which point to discard, though this is not true in general. To see why, recall from equations (4.4) and (4.5) that the first- and second-order approximations to $\Delta\lambda_1$ following the removal of point i are

$$\dot{\lambda}_1(0)\epsilon = -\frac{n}{n-1}(\boldsymbol{u}_1^\top \boldsymbol{v}_i)^2 \quad \text{and} \quad \tfrac{1}{2}\ddot{\lambda}_1(0)\epsilon^2 = -\frac{n^2}{(n-1)^2}(\boldsymbol{u}_1^\top \boldsymbol{v}_i)^2 \sum_{k=2}^{m} \frac{(\boldsymbol{u}_k^\top \boldsymbol{v}_i)^2}{\lambda_k - \lambda_1},$$

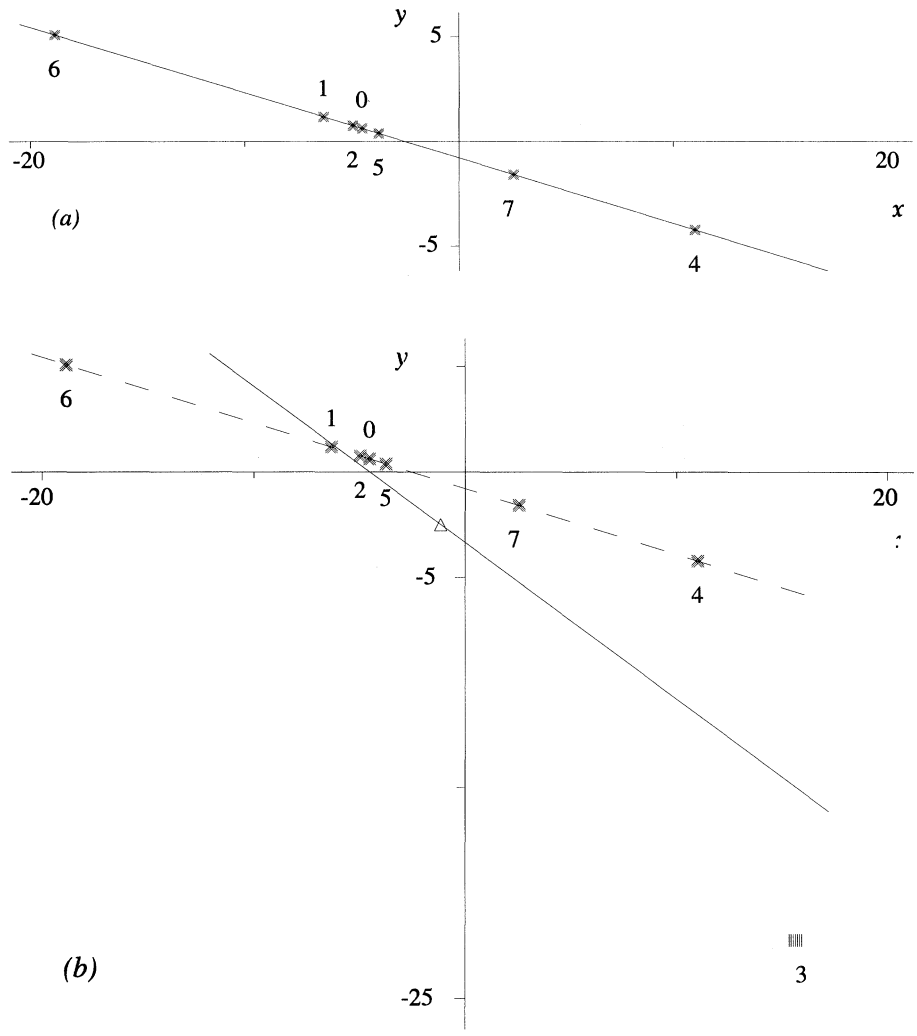


Figure 4. Data lying on the line $3.8x + 12.2y + 9.6 = 0$ (crosses) with a single outlier (square): (*a*) the correct fit; (*b*) the OR fit affected by the outlier (point 3). The triangle shows the centroid of the data set.

while the residual for point i is (2.8)

$$\ell_i = \boldsymbol{u}_1^\top \boldsymbol{v}_i.$$

Evidently, *there is agreement at the first-order approximation*, since $\Delta\lambda_1$ is then simply proportional to ℓ_i^2. In other words, the point with largest residual will also be that point inducing maximum change in λ_1 at a first-order expansion! The residual method is therefore subsumed by the influence function method; the results are identical if a first-order eigenvalue perturbation model is used. The second- (and higher-) order terms start to account for the change in eigenvector

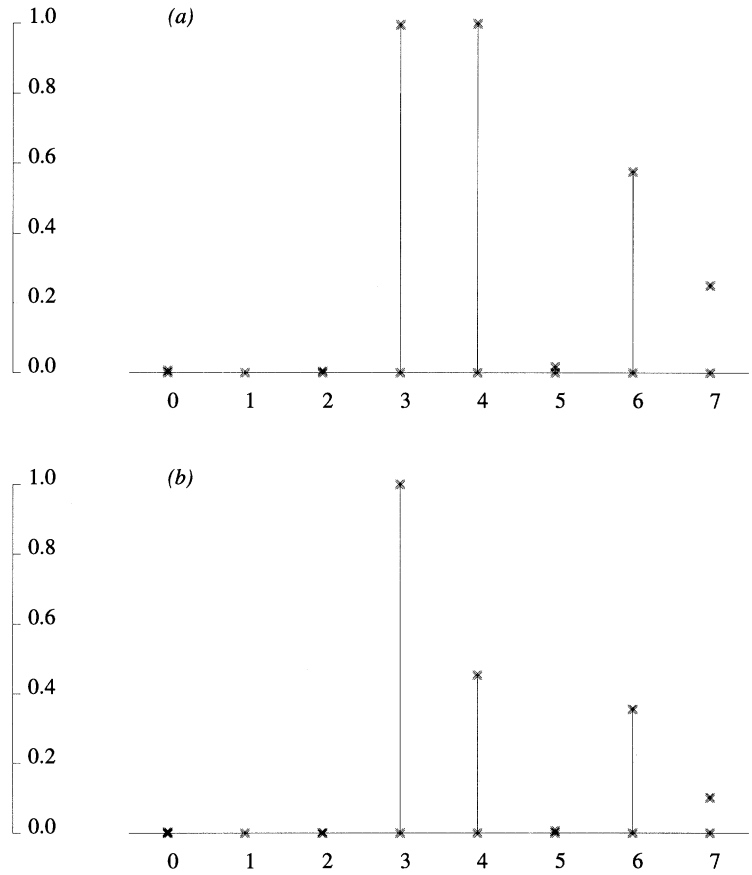


Figure 5. Superior discrimination power of minimum eigenvalue method versus residual method: point 3 is the outlier; (*a*) residual ℓ_i^2 for each data point relative to the largest; (*b*) actual change in minimum eigenvalue, $\Delta\lambda_1$, for each data point relative to the largest.

structure (the 'look-ahead effect'), so a different point might then have a larger overall influence.

We demonstrate this by means of a simple two-dimensional example. Consider the set of points in figure 4*a*. The dominant data (7 points) are noise-free and lie on the straight line $3.8x + 12.2y + 9.6 = 0$. There is a single contaminant (point 3), and the OR fit to the full data set is shown in figure 4*b*. The outlier has a significant detrimental effect, 'pulling' the line over towards point 3. The residuals ℓ_i are listed in table 1, and it is clearly impossible to identify the outlier correctly on the basis of these perpendicular distances. In fact, point 4 has a larger residual than point 3! The minimum eigenvalue scheme offers a much clearer distinction, discriminating between point 3 and point 4 by a factor greater than two. This discrimination power is illustrated graphically in figure 5, where the relative values of ℓ_i^2 are shown in (*a*) and those of $\Delta\lambda_1$ in (*b*). We therefore only consider the minimum eigenvalue method in the sections that follow.

Table 1 also lists the second-order eigenvalue perturbations to show that they provide a fairly good approximation to the true changes in λ_1. The approximations are worse for points further from the centroid (e.g. points 3 and 6), but still identify the most influential point correctly.

Table 1. *Data for the example in figure 4*

(The eigenvalue scheme provides better discrimination between the outlier (point 3) and the valid data than the residual scheme, which identifies the wrong point (point 4). The important comparison is between ℓ_i^2 and the true change in λ_1 (columns 3 and 4). The last column (column 5) gives the second-order approximation to $\Delta\lambda_1$, which still identifies the outlier correctly.)

point	residual (ℓ_i, ℓ_i^2)		$\Delta\lambda_1$ (true, approx)	
0	0.5337	0.2849	0.2911	0.2910
1	−0.1285	0.0165	0.0172	0.0172
2	0.3753	0.1409	0.1446	0.1445
3	−6.2510	39.0751	97.8422	64.6086
4	6.2607	39.1961	44.2979	43.8900
5	0.8231	0.6775	0.6877	0.6875
6	−4.7517	22.5791	34.7413	30.7303
7	3.1384	9.8497	9.9089	9.9090

(*d*) *Experiments*

Although our theory generalizes to any number of dimensions, the data used in this section are two-dimensional, since higher-dimensional spaces are harder to visualize (four-dimensional data are used in § 6). Our first example illustrates the facility of our algorithm to cope with multiple outliers, while the second illustrates 'masking'.

Figure 6 shows the OR fit for 16 data points perturbed from the line $-6.5x + 2.1y + 3.2 = 0$ by Gaussian noise ($\sigma = 1$). There are seven contaminants (points 1, 3, 4, 5, 9, 17 and 21), and a 95% significance level is required for the χ^2 test. Table 2 gives the changes in minimum eigenvalue at each iteration, with the largest change identifying the point to be deleted. The six iterations eliminate contaminants 5, 21, 1, 3, 17 and 4 in order.

We make several observations. First, the procedure terminates automatically, with the algorithm itself deciding (based on a statistical decision) when to stop removing outliers. Errors still remain after the final iteration, but they fall within acceptable levels, i.e. they are consistent with the assumed levels of noise. Second, as predicted by theory, λ_1 decreases with the deletion of each point. Third, a contaminant is sometimes sufficiently consistent with the underlying fit to be indistinguishable from a valid noisy point; point 9 is not recognized as an outlier since it lies in the centre of the data. This is not serious since the error introduced by the outlier falls within the tolerable bound, so is not severely disruptive. Indeed, a principal dilemma in outlier detection lies in deciding whether a potential outlier is an extreme (but valid) perturbation of the dominant set, or a contaminant from another population. Such a decision is necessarily statistical and guidance must be obtained from the application (e.g. in the form of prior knowledge of σ). Finally, we note that the outliers are confidently detected, despite comprising a third of the data set.

Our second example illustrates why only one point is deleted at a time. Figure 7 shows points displaced from their line $-2.5x + 8y + 7.1 = 0$ by Gaussian noise ($\sigma = 1$). Four outliers are added (points 5, 8, 17 and 19) and table 3 summarizes the iterations for selected points. We observe that a poor initial fit can incorrectly attribute influence to points which are actually correct; for instance, point 18 has a large influence in iteration 1 even though it isn't an outlier. This influence

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Table 2. *Table of data for the example in figure 6*

(Each column refers to an iteration and gives the change in the smallest eigenvalue λ_1 caused by deleting each point. The largest change (underlined) determines which point to delete.)

point	1	2	3	4	5	6
0	5.8659	0.9802	0.0193	2.6291	0.3270	0.3791
1	228.4422	260.0809	291.4066	—	—	—
2	2.9063	1.4266	0.1850	0.0002	0.4895	0.7661
3	324.4088	293.1343	264.0601	236.4234	—	—
4	21.4060	29.5761	39.6450	50.1246	39.7099	33.5104
5	375.3948	—	—	—	—	—
6	1.3932	0.3566	0.0342	0.4938	0.0003	0.0779
7	0.0059	0.2449	1.5407	2.4455	0.7764	0.6662
8	1.1019	0.2493	0.0753	0.5680	0.0019	0.0327
9	11.5642	6.8780	3.1768	1.0760	3.3019	5.3344
10	2.6872	0.5479	0.0214	1.0614	0.0443	0.1708
11	1.4537	0.1342	2.4536	10.2752	4.3754	0.6000
12	2.3191	0.0079	1.0828	6.3807	2.1837	0.0872
13	2.5193	0.0900	0.5740	4.3251	1.2076	0.0042
14	0.9171	0.1090	0.2154	1.0694	0.0924	0.0001
15	3.1934	0.7044	0.0044	1.0119	0.0280	0.2600
16	3.7753	1.5449	0.1829	0.0515	0.2761	0.8574
17	159.4227	134.3600	113.5430	92.5173	110.5653	—
18	1.6782	0.2114	0.1622	1.4740	0.1745	0.0182
19	7.4104	2.2989	0.2475	0.5377	0.0458	1.3712
20	1.9686	0.3277	0.0830	1.1962	0.0894	0.0617
21	306.1983	339.7714	—	—	—	—
22	0.9484	0.1331	0.1800	0.9458	0.0612	0.0022
delete	5	21	1	3	17	4
λ_1	1019.4191	679.6477	388.2411	151.8177	41.2524	7.7420

decreases as the fit improves. It is also important to recompute fits after removing points since outliers are sometimes 'masked' by other outliers. For instance, the fact that point 19 is an outlier only becomes apparent in iteration 3, once outliers 17 and 8 have been removed.

(*e*) *Discussion*

Section 3*a* mentioned some drawbacks of existing outlier tests for OR, and chief among them was the often unrealistic requirement for a Gaussian null distribution. This requires that the data points $\boldsymbol{r}_i$ occupy a Gaussian-like hyperellipsoid about the centroid $\bar{\boldsymbol{r}}$. In many applications (including our own), there is no justification for such an assumption, since the data $\boldsymbol{r}_i$ are often *not* random variables in the sense that they estimate a stationary mean. Instead, they may simply represent an arbitrary structure perturbed by noise, where the *noise errors* are the fundamental random variables. In our case, for instance, fixed points are displaced

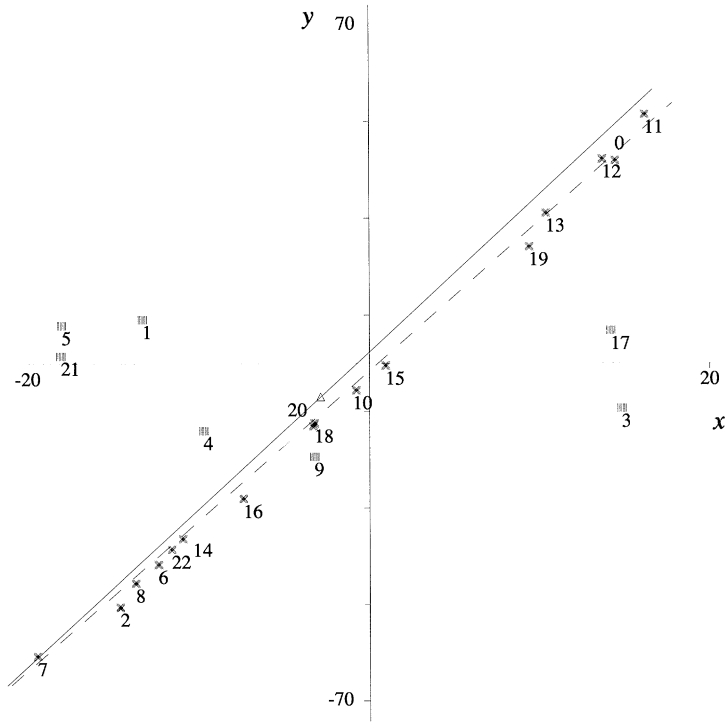


Figure 6. Data points (crosses) originally lying on the line $-6.5x + 2.1y + 3.2 = 0$ (dashed) are perturbed by Gaussian noise ($\sigma = 1$) and contaminants are added (squares). The OR fit (solid) is affected by both contaminants and the noise. The centroid of the entire data set is also shown (triangle).

from their hyperplane π by small noise vectors. The 'small signal' therefore introduces randomness into the 'large signal', and the statistical assumptions should therefore only be imposed on the noise, not on the distribution of features within the hyperplane.

Since we have m-dimensional data with a single constraint, the first $m - 1$ eigenvectors explain the structure of the hyperplane (irrelevant from a noise viewpoint), and the last eigenvector explains the noise; if there was no noise in the system, λ_1 would be zero. By only focussing on this last dimension, we avoid imposing conditions on the distribution of the data itself. This contrasts with other schemes (see, for example, Hawkins 1980; Gnanadesikan & Kettenring 1972) that analyse the full set of principal components, and thus require knowledge of how the data are distributed along all other axes.

5. Error analysis

Once the outliers have been removed, it is useful to relate the noise in the input to the OR solution. Several researchers have recently stressed the advantages of systematically studying the statistical-error behaviour of an algorithm (Weng *et al.* 1989; Kanatani 1993). We additionally emphasize the importance of having first removed the outliers, since their presence can severely distort such calculations, leading to serious bias or reduction in precision (Barnett & Lewis 1984). Qualitatively, outliers tend to deflate correlations and inflate variances. Indeed,

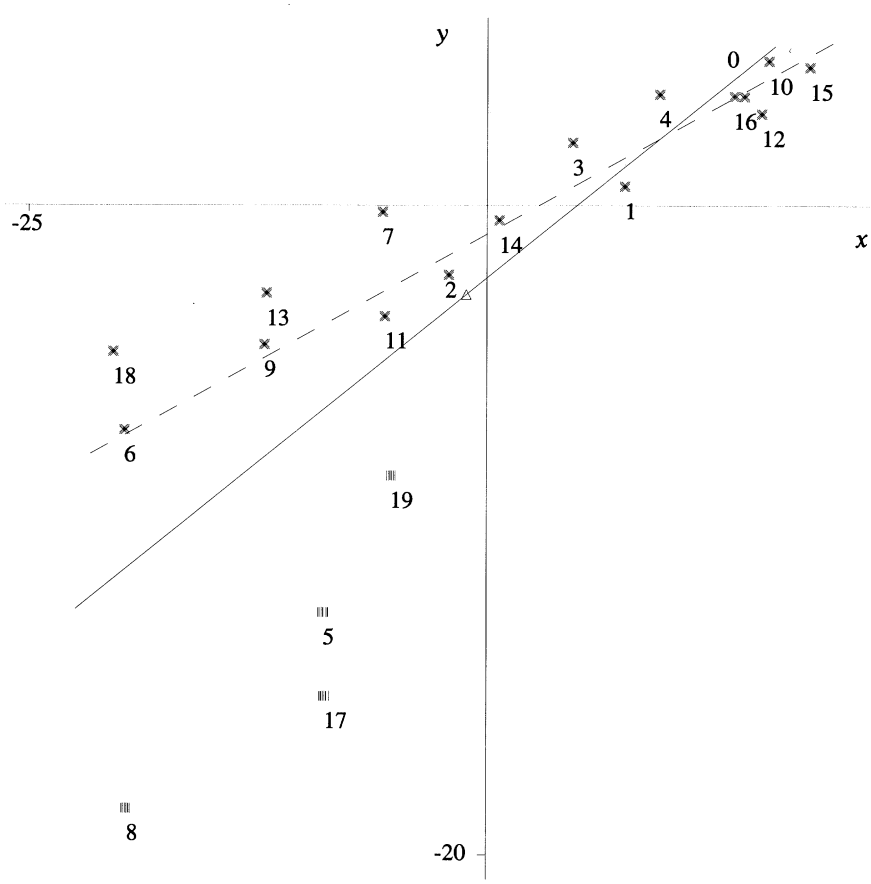


Figure 7. Data points (crosses) originally lying on the line $-2.5x + 8y + 7.1 = 0$ (dashed) are perturbed by Gaussian noise ($\sigma = 1$) and contaminants are added (squares). The OR fit (solid) and data centroid (triangle) are shown.

Barnett and Lewis (1984, p. 249) mention in this regard that 'even one or two outliers in a large set can wreak havoc!'

In § 5 *a*, we perform a similar error analysis to that of Weng *et al.* (1989), and in § 5 *b* we derive the variance and covariance expressions for the residuals. We discover that this variance differs from point to point, and modify our previous algorithm accordingly.

(*a*) *Hyperplane covariance matrix*

Let the data point $\hat{\boldsymbol{r}}$ be perturbed by *independent isotropic additive Gaussian noise* $\delta\boldsymbol{r}$, giving the measurement $\boldsymbol{r} = \hat{\boldsymbol{r}} + \delta\boldsymbol{r}$. It is assumed that each noise perturbation has zero mean ($E\{\delta\boldsymbol{r}_i\} = \boldsymbol{0}$) with variance σ^2, i.e. $E\{\boldsymbol{r}_i\} = E\{\hat{\boldsymbol{r}}_i + \delta\boldsymbol{r}_i\} = \hat{\boldsymbol{r}}_i$, $\mathrm{Var}\{\boldsymbol{r}_i\} = \mathrm{Var}\{\delta\boldsymbol{r}_i\}$ and

$$\boldsymbol{\Gamma}_{\boldsymbol{r}} = E\{\delta\boldsymbol{r}_i\ \delta\boldsymbol{r}_i^{\top}\} = \sigma^2\boldsymbol{I}. \tag{5.1}$$

We further assume (as in Weng *et al.* 1989) that the data points have independent errors, i.e. $E\{\delta\boldsymbol{r}_i\delta\boldsymbol{r}_j^{\top}\} = \delta_{ij}\boldsymbol{\Gamma}_{\boldsymbol{r}}$, where δ_{ij} is the Kronecker delta product. Our objective is to assess the effects of these errors on $\boldsymbol{u}_1$ and, hence, obtain a confidence estimate in the computed normal $\boldsymbol{n}$.

Table 3. *Subset of the data table for figure 7*

(Each column refers to an iteration and gives the decrease in minimum eigenvalue λ_1 caused by deleting each point.)

point	1	2	3	4
5	35.3568	46.2664	69.0109	—
8	57.6128	80.7649	—	—
11	1.6243	0.4591	0.0631	1.0289
17	71.6464	—	—	—
18	47.8599	34.2896	14.7933	5.8793
19	12.0306	17.0934	27.0519	36.3879
delete	17	8	5	19
λ_1	201.4694	120.7046	51.6937	15.3057

Consider the centred data points $\boldsymbol{v}_i$. The noise in $\boldsymbol{r}_i$ induces an error $\delta\boldsymbol{v}_i$ in $\boldsymbol{v}_i$:

$$\boldsymbol{v}_i = \boldsymbol{r}_i - \bar{\boldsymbol{r}} = (\hat{\boldsymbol{r}}_i - \hat{\bar{\boldsymbol{r}}}) + (\delta\boldsymbol{r}_i - \delta\bar{\boldsymbol{r}}) = \hat{\boldsymbol{v}}_i + \delta\boldsymbol{v}_i.$$

The covariance matrix for $\delta\boldsymbol{v}$ differs from that of $\delta\boldsymbol{r}$, because by centring the data we confound the noise of the centroid with the noise of each individual point. Noting that $\delta\bar{\boldsymbol{r}} = \sum_{i=1}^{n} \delta\boldsymbol{r}_i/n$, we have

$$\begin{aligned} E\{\delta\boldsymbol{v}_i\ \delta\boldsymbol{v}_i^\top\} &= E\{(\delta\boldsymbol{r}_i - \delta\bar{\boldsymbol{r}})\ (\delta\boldsymbol{r}_i - \delta\bar{\boldsymbol{r}})^\top\} \\ &= E\{\delta\boldsymbol{r}_i\ \delta\boldsymbol{r}_i^\top\} - E\{\delta\bar{\boldsymbol{r}}\delta\boldsymbol{r}_i^\top\} - E\{\delta\boldsymbol{r}_i\delta\bar{\boldsymbol{r}}^\top\} + E\{\delta\bar{\boldsymbol{r}}\delta\bar{\boldsymbol{r}}^\top\} \\ &= \sigma^2\boldsymbol{I} - \frac{1}{n}\sigma^2\boldsymbol{I} - \frac{1}{n}\sigma^2\boldsymbol{I} + \frac{1}{n^2}(n\sigma^2\boldsymbol{I}) = \frac{n-1}{n}\sigma^2\boldsymbol{I}, \end{aligned}$$

and for $i \neq j$,

$$\begin{aligned} E\{\delta\boldsymbol{v}_i\ \delta\boldsymbol{v}_j^\top\} &= E\{(\delta\boldsymbol{r}_i - \delta\bar{\boldsymbol{r}})\ (\delta\boldsymbol{r}_j - \delta\bar{\boldsymbol{r}})^\top\} \\ &= E\{\delta\boldsymbol{r}_i\delta\boldsymbol{r}_j^\top\} - E\{\delta\bar{\boldsymbol{r}}\delta\boldsymbol{r}_j^\top\} - E\{\delta\boldsymbol{r}_i\delta\bar{\boldsymbol{r}}^\top\} + E\{\delta\bar{\boldsymbol{r}}\delta\bar{\boldsymbol{r}}^\top\} \\ &= \boldsymbol{0} - \frac{1}{n}\sigma^2\boldsymbol{I} - \frac{1}{n}\sigma^2\boldsymbol{I}_m + \frac{1}{n^2}(n\sigma^2\boldsymbol{I}) = -\frac{1}{n}\sigma^2\boldsymbol{I}. \end{aligned}$$

The covariance matrix for $\boldsymbol{v}_i$ is thus

$$\boldsymbol{\Gamma}_v = E\{\delta\boldsymbol{v}_i\ \delta\boldsymbol{v}_i^\top\} = \frac{\sigma^2}{n}(n-1)\boldsymbol{I} = \frac{n-1}{n}\boldsymbol{\Gamma}_r, \tag{5.2}$$

and for $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ $(i \neq j)$ is

$$\boldsymbol{\Upsilon}_v = E\{\delta\boldsymbol{v}_i\delta\boldsymbol{v}_j^\top\} = -\frac{\sigma^2}{n}\boldsymbol{I} = -\frac{1}{n}\boldsymbol{\Gamma}_r. \tag{5.3}$$

When n is large, the covariance matrices for the centred data $\boldsymbol{v}_i$ tend towards those for the uncentred data $\boldsymbol{r}_i$, since $(n-1)/n \to 1$ and $-1/n \to 0$. In matrix form, $\boldsymbol{V} = [\boldsymbol{v}_1|\boldsymbol{v}_2|\cdots|\boldsymbol{v}_n]$ and $\delta\boldsymbol{V} = [\delta\boldsymbol{v}_1|\delta\boldsymbol{v}_2|\cdots|\delta\boldsymbol{v}_n]$. The perturbation caused in the scatter matrix $\boldsymbol{W} = \boldsymbol{V}\boldsymbol{V}^\top$ due to the noise in $\boldsymbol{V}$ is shown in the following

equation:

$$\boldsymbol{W} = (\hat{\boldsymbol{V}} + \delta\boldsymbol{V})(\hat{\boldsymbol{V}} + \delta\boldsymbol{V})^{\top} = \hat{\boldsymbol{V}}\,\hat{\boldsymbol{V}}^{\top} + \hat{\boldsymbol{V}}\delta\boldsymbol{V}^{\top} + \delta\boldsymbol{V}\,\hat{\boldsymbol{V}}^{\top} + \delta\boldsymbol{V}\delta\boldsymbol{V}^{\top}.$$

We write $\boldsymbol{W} = \hat{\boldsymbol{W}} + \delta\boldsymbol{W}$ and note that $\hat{\boldsymbol{W}} = \hat{\boldsymbol{V}}\,\hat{\boldsymbol{V}}^{\top}$. Then, using a first-order approximation (Weng *et al.* 1989),

$$\delta\boldsymbol{W} \approx \hat{\boldsymbol{V}}\delta\boldsymbol{V}^{\top} + \delta\boldsymbol{V}\,\hat{\boldsymbol{V}}^{\top}. \tag{5.4}$$

Finally we can consider the eigenvector $\hat{\boldsymbol{u}}_1$. Since $\hat{\boldsymbol{W}}\hat{\boldsymbol{u}}_j = \hat{\lambda}_j\hat{\boldsymbol{u}}_j$ gives the noise-free solution, $\hat{\lambda}_1 = 0$. Moreover, the noise-free residuals $\hat{\ell}_i = \hat{\boldsymbol{u}}_1^{\top}\hat{\boldsymbol{v}}_i$ are zero, because all points then lie *on* the hyperplane. Since $\delta\boldsymbol{W}$ is a real symmetric matrix, the first-order change in $\hat{\boldsymbol{u}}_1$ is (appendix B):

$$\delta\boldsymbol{u}_1 = -\sum_{k=2}^{m}\frac{(\hat{\boldsymbol{u}}_k^{\top}\,\delta\boldsymbol{W}\,\hat{\boldsymbol{u}}_1)\hat{\boldsymbol{u}}_k}{\hat{\lambda}_k} = -\left(\sum_{k=2}^{m}\frac{\hat{\boldsymbol{u}}_k\hat{\boldsymbol{u}}_k^{\top}}{\hat{\lambda}_k}\right)\delta\boldsymbol{W}\,\hat{\boldsymbol{u}}_1.$$

Now $\delta\boldsymbol{W}\,\hat{\boldsymbol{u}}_1 = (\hat{\boldsymbol{V}}\delta\boldsymbol{V}^{\top} + \delta\boldsymbol{V}\,\hat{\boldsymbol{V}}^{\top})\hat{\boldsymbol{u}}_1 = \hat{\boldsymbol{V}}\delta\boldsymbol{V}^{\top}\hat{\boldsymbol{u}}_1$ since $\hat{\boldsymbol{V}}^{\top}\hat{\boldsymbol{u}}_1 = \boldsymbol{0}$ (the noise-free residuals $\hat{\boldsymbol{u}}_1^{\top}\hat{\boldsymbol{v}}_i$ equal 0), so $\delta\boldsymbol{u}_1$ can be written

$$\delta\boldsymbol{u}_1 = \hat{\boldsymbol{J}}\,\hat{\boldsymbol{V}}\delta\boldsymbol{V}^{\top}\hat{\boldsymbol{u}}_1 = \hat{\boldsymbol{J}}\sum_{i=1}^{n}\hat{\boldsymbol{v}}_i(\delta\boldsymbol{v}_i^{\top}\hat{\boldsymbol{u}}_1), \quad \text{where} \quad \hat{\boldsymbol{J}} = -\sum_{k=2}^{m}\frac{\hat{\boldsymbol{u}}_k\hat{\boldsymbol{u}}_k^{\top}}{\hat{\lambda}_k} \tag{5.5}$$

and the 'hat' indicates noise-free quantities. Finally, the covariance matrix for $\boldsymbol{u}_1$ can be computed, giving a measure of confidence in the eigenvector solution (proof in appendix C *a*):

$$\boldsymbol{\Gamma}_{\boldsymbol{u}_1} = E\{\delta\boldsymbol{u}_1\delta\boldsymbol{u}_1^{\top}\} = -\sigma^2\hat{\boldsymbol{J}}. \tag{5.6}$$

Many of the above equations require the true noise-free quantities (e.g. $\hat{\boldsymbol{V}}$, $\hat{\boldsymbol{u}}_1$, $\hat{\boldsymbol{J}}$), which are not available in general. Weng *et al.* (1989) pointed out that if one writes, for instance, $\hat{\boldsymbol{V}} = \boldsymbol{V} - \delta\boldsymbol{V}$ and substitutes this in the relevant equations, the terms in $\delta\boldsymbol{V}$ disappear in the first-order expressions, allowing $\boldsymbol{V}$ to be simply interchanged with $\hat{\boldsymbol{V}}$, and so on. The covariance matrix for $\boldsymbol{u}_1$ can thus be expressed in terms of directly measurable quantities:

$$\boldsymbol{\Gamma}_{\boldsymbol{u}_1} = -\sigma^2\sum_{k=2}^{m}\frac{\boldsymbol{u}_k\boldsymbol{u}_k^{\top}}{\lambda_k}. \tag{5.7}$$

(*b*) *Residual variance and covariance*

The residual is

$$\ell_i = \hat{\ell}_i + \delta\ell_i = (\hat{\boldsymbol{u}}_1 + \delta\boldsymbol{u}_1)^{\top}(\hat{\boldsymbol{v}}_i + \delta\boldsymbol{v}_i) = \hat{\boldsymbol{u}}_1^{\top}\hat{\boldsymbol{v}}_i + \hat{\boldsymbol{u}}_1^{\top}\delta\boldsymbol{v}_i + \hat{\boldsymbol{v}}_i^{\top}\delta\boldsymbol{u}_1 + \delta\boldsymbol{u}_1^{\top}\delta\boldsymbol{v}_i,$$

where $\hat{\ell}_i = 0$. We neglect second-order terms and obtain an expression for the perturbation:

$$\delta\ell_i \approx \hat{\boldsymbol{u}}_1^{\top}\delta\boldsymbol{v}_i + \hat{\boldsymbol{v}}_i^{\top}\delta\boldsymbol{u}_1.$$

Thus, $E\{\delta\ell_i\} = \boldsymbol{0}$ and its variance $\sigma_{\ell i}^2$ is

$$\begin{aligned}\mathrm{Var}\{\delta\ell_i\} &= E\,\{(\hat{\boldsymbol{u}}_1^{\top}\delta\boldsymbol{v}_i + \hat{\boldsymbol{v}}_i^{\top}\delta\boldsymbol{u}_1)^2\}\\ &= E\,\{(\hat{\boldsymbol{u}}_1^{\top}\delta\boldsymbol{v}_i)^2 + (\hat{\boldsymbol{v}}_i^{\top}\delta\boldsymbol{u}_1)^2 + 2(\hat{\boldsymbol{v}}_i^{\top}\delta\boldsymbol{u}_1)(\hat{\boldsymbol{u}}_1^{\top}\delta\boldsymbol{v}_i)\}\\ &= \hat{\boldsymbol{u}}_1^{\top}\boldsymbol{\Gamma}_v\,\hat{\boldsymbol{u}}_1 + \hat{\boldsymbol{v}}_i^{\top}\boldsymbol{\Gamma}_{\boldsymbol{u}_1}\hat{\boldsymbol{v}}_i + 2\hat{\boldsymbol{v}}_i^{\top}E\{\delta\boldsymbol{u}_1\delta\boldsymbol{v}_i^{\top}\}\hat{\boldsymbol{u}}_1.\end{aligned} \tag{5.8}$$

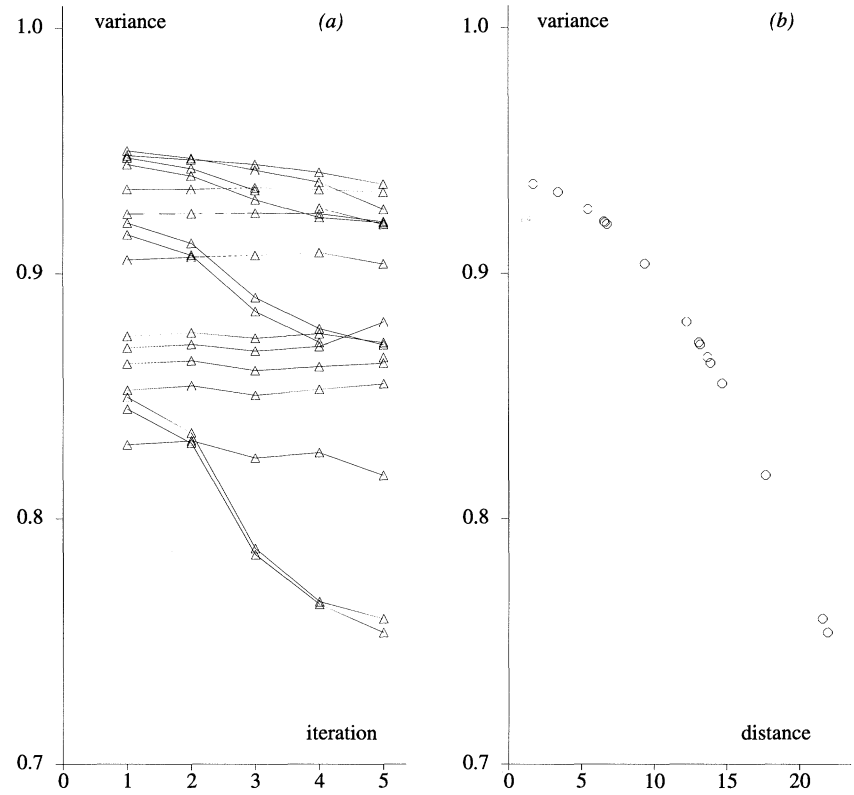


Figure 8. Computed residual variances $\sigma^2_{\ell i}$ for the valid data points in figure 7: (*a*) the presence of outliers tends to *inflate* the variance estimates, which reduce to the correct theoretical value as more outliers are removed; (*b*) the variance falls off with increasing distance from the centroid $d_i = \boldsymbol{u}_2^\top \boldsymbol{v}_i$ (measured along the fitted axis).

Appendix C *b* shows that this reduces to

$$\sigma^2_{\ell i} = \sigma^2 \left[\frac{n-1}{n} - \sum_{k=2}^{m} \frac{(\boldsymbol{v}_i^\top \boldsymbol{u}_k)^2}{\lambda_k} \right]. \tag{5.9}$$

The variance in the residual for point i therefore consists of a *constant* term $\sigma^2_\ell = (n-1)\sigma^2/n$, dependent on the variance of the raw data, and a *variable* term, dependent on the specific location of $\boldsymbol{v}_i$. Evidently, no single variance applies to all the residuals; *the residual error distribution is different for every point.* As a rough guide, points further away from the data centroid have smaller residuals. This is because the further a point is from the centre along a given axis, the greater its potential influence in altering that axis. This is analogous to a lever, where the moment caused by a constant force varies with the distance between the pivot and the point of application.

To illustrate this, we return to the example in figure 7, where points 5, 8, 17 and 19 were identified as outliers using algorithm 2. Figure 8 *a* graphs the residual variances $\sigma^2_{\ell i}$ for the valid points at each iteration, illustrating that the presence of outliers tends to *inflate* the computed variances. Figure 8 *b* plots the final variances against the projected distances $d_i = (\boldsymbol{u}_2^\top \boldsymbol{v}_i)$, showing that points

further from the centroid (measured along the axis $\boldsymbol{u}_2$) have smaller variances. The relation is a simple quadratic in the two-dimensional case:

$$\sigma_{\ell i}^2 = \sigma^2 \left[\frac{n-1}{n} - \frac{d_i^2}{\lambda_2} \right].$$

This phenomenon of inhomogeneous residual variances (heteroscedasticity) is termed 'ballooning' (Cook & Weisberg 1982; Barnett & Lewis 1984). It is a common effect, arising even with a simple linear regression model (i.e. two-dimensional case with a single regressor variable). It is inconvenient because if the residuals all have different Gaussian distributions, there is no way to compare them sensibly; a single χ^2 test is invalid. This problem was ignored in the algorithms of § 4 *a* and § 4 *b*, which assumed the residuals ℓ_i to be from a univariate distribution. One solution is to *compute* the variance for each point and to scale the residual appropriately (Barnett & Lewis 1984), thereby ensuring that all residuals have unit standard deviation. The difficulty here is that the formula for computing $\sigma_{\ell i}$ assumes that there are no outliers! When there *are* outliers, the computed variances are distorted, rendering them useless.

We solve this problem by proceeding in two stages. First we note that $\sigma_\ell^2 = (n-1)\sigma^2/n$ is the *upper bound* for $\sigma_{\ell i}^2$; as points move further away from the centroid, the variance *decreases*, and for these points the upper bound exceeds the correct value. We therefore initially use σ_ℓ^2 as the common variance for *all* points. The fact that we have overestimated the variance for some points errs on the side of conservatism, since all *valid* points will then fall *well* inside the allowed probability region – indeed, further inside this region than the specified confidence level merits. Consequently, any points falling *outside* the limit are certain to be contaminants.

Once the worst outliers have been removed, we refine the dividing line by introducing the computed variances, assured that $\sigma_{\ell i}$ will not be too inaccurate. Each residual is then scaled by its individual $\sigma_{\ell i}$ value, giving the new χ^2 test statistic

$$s = \sum_{i \in \mathcal{V}} \frac{\ell_i^2}{\sigma_{\ell i}^2}, \tag{5.10}$$

where $\mathcal{V}$ is the current set of valid points, a subset of the initial data set with some outliers removed. As before, the modified algorithm proceeds until the scaled variances fall within specified probability limits, indicating that no further outliers can be detected. Details of these algorithms, along with experimental results, are given in Shapiro & Brady (1993).

Finally, we note that the residuals are not mutually independent; the covariance between any two residuals ℓ_i and ℓ_j is given by (appendix C *c*):

$$\mathrm{Cov}\{\delta\ell_i, \delta\ell_j\} = E\{\delta\ell_i \delta\ell_j\} = -\sigma^2 \left[\frac{1}{n} + \sum_{k=2}^{m} \frac{(\boldsymbol{v}_i^\top \boldsymbol{u}_k)(\boldsymbol{v}_j^\top \boldsymbol{u}_k)}{\lambda_k} \right]. \tag{5.11}$$

We have found empirically that this is a minor effect.

6. Computer vision application

We have encountered the outlier problem in the course of our computer vision research into the structure from motion (SFM) problem, where our algorithms require hyperplane fitting (Shapiro 1993; Shapiro *et al.* 1994). This aspect of data fitting is frequently either ignored in the vision literature or treated heuristically, adding further to the already considerable difficulty of devising algorithms

that work reliably on real imagery. Indeed, much of the early SFM work (see, for example, Ullman 1979; Longuet-Higgins 1981; Tsai & Huang 1984) ignored both noise perturbations and outliers. There has subsequently been a trend towards consideration of noise (Faugeras *et al.* 1987; Weng *et al.* 1989; Kanatani 1993), but the problem of outliers remains largely unexplored. Thus, although many SFM algorithms now have partial immunity to noise (indeed some estimate this noise (Weng *et al.* 1989; Kanatani 1993)), they generally lack immunity to contaminants (a notable exception being Torr & Murray (1993*a*, *b*)).

The application we describe concerns the computation of the epipolar geometry parameters for an affine camera. Section 6 *a* provides a brief background to the problem and § 6 *b* gives results on real data, demonstrating the successful operation of the outlier rejection scheme.

(*a*) *Affine epipolar geometry*

Suppose a three-dimensional scene point $(X, Y, Z)^\top$ is observed in two images as $(x, y)^\top$ and $(x', y')^\top$. We model the projection operation by means of an *affine camera* (Mundy & Zisserman 1992), a generalization of the scaled orthographic camera model. The affine camera preserves parallelism (parallel scene lines appear as parallel lines in the image) and approximates the more accurate perspective projection model well when the field of view is small and the variation of depth of the scene along the camera's line of sight is small compared with the average scene depth. It is shown in Zisserman (1992) and Shapiro (1992) that $(x, y)^\top$ is then related to its counterpart $(x', y')^\top$ by the *affine epipolar equation*,

$$n_1 x' + n_2 y' + n_3 x + n_4 y + d = 0. \tag{6.1}$$

The notion of an epipolar constraint is well known in the stereo and motion literature, and specifies a line in each image (the 'epipolar line') on which the point must lie, thereby reducing the search for 'matching points' from two dimensions to one. When there are $n > 4$ correspondences (in the presence of noise), the points do not lie exactly on the computed epipolar lines, and $\{\boldsymbol{n}, d\}$ is determined by minimizing (Shapiro *et al.* 1994):

$$\sum_{i=1}^{n} (n_1 x'_i + n_2 y'_i + n_3 x_i + n_4 y_i + d)^2 \quad \text{subject to} \quad n_1^2 + n_2^2 + n_3^2 + n_4^2 = 1.$$

This is precisely equation (2.5) with $m = 4$.

(*b*) *Experiments*

We illustrate the outlier rejection scheme on some real image sequences. To obtain the initial matches (x', y', x, y), corner features are extracted from each image using the Wang–Brady corner detector (Wang & Brady 1992) and tracked over time (Shapiro *et al.* 1992). System noise arises from quantization error, corner localization error and simplified camera model assumptions, and is estimated at $\sigma = 0.7$ pixels. Outliers arise mainly from mismatches in the correspondence algorithm, failures in the segmentation algorithm (e.g. inclusion of a feature which doesn't belong to the object of interest), corners on extremal boundaries (e.g. object silhouette) and false corners (e.g. the conjunction of edges at different depths). A 95% confidence level is used throughout.

Figure 9 *a* shows the first frame in a sequence taken from a camera moving to the right in a static scene. The computed flow vectors (for 219 corners) are shown in figure 9 *b*, and while they certainly convey the motion with good qualitative accuracy, there are clearly outliers arising from the error sources discussed earlier. The black left-hand edge of the image also provides an interesting test of

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

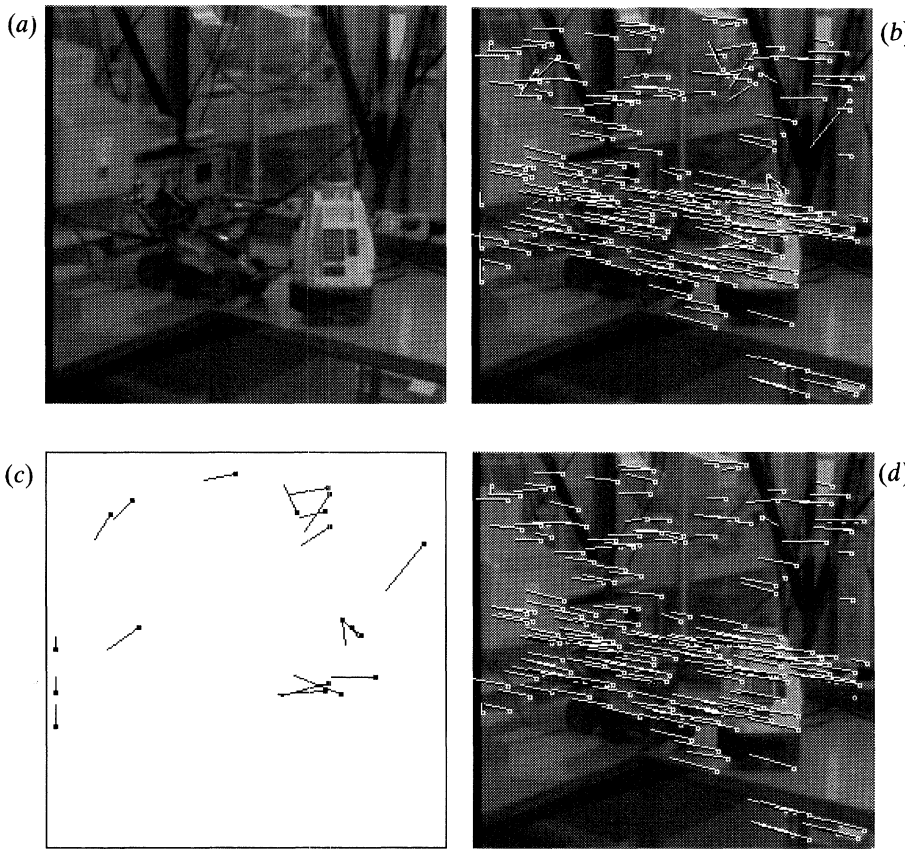


Figure 9. Outlier rejection with a moving camera and static scene (camera moves right). Motion vectors are shown double length for clarity: (*a*) first frame; (*b*) initial motion vectors; (*c*) rejected motion vectors; (*d*) final motion vectors (outliers removed).

the algorithm's performance. Figure 9*c* shows the 22 outliers identified by our method, all of which were confirmed as incorrect matches (by manual inspection). Figure 9*d* shows the final set of data. Not all the outliers have been removed, but those that remain have a negligible effect on the fit. The minimum eigenvalue was reduced from $\lambda_1 = 370.13$ to $\lambda_1 = 42.94$. A more intuitive error measure is the sum of the squared perpendicular image distances from the points to their respective epipolar lines:

$$\kappa = \sum_{i=1}^{n} \frac{(n_1 x_i' + n_2 y_i' + n_3 x_i + n_4 y_i + d)^2}{n_1^2 + n_2^2} + \sum_{i=1}^{n} \frac{(n_1 x_i' + n_2 y_i' + n_3 x_i + n_4 y_i + d)^2}{n_3^2 + n_4^2}.$$

In our example, κ is reduced from 1933.36 to 171.83, and the final fit therefore has an average perpendicular distance of 0.66 pixels between each corner and its epipolar line, a reasonable result given $\sigma = 0.7$. Figure 10 shows two additional examples, where the camera is stationary and an object moves in the scene. This introduces an additional source of error, namely the segmentation of the stationary background from the moving points (e.g. some motion of the shirt is included with the head motion in figure 10*b*). The majority of outliers are successfully rejected.

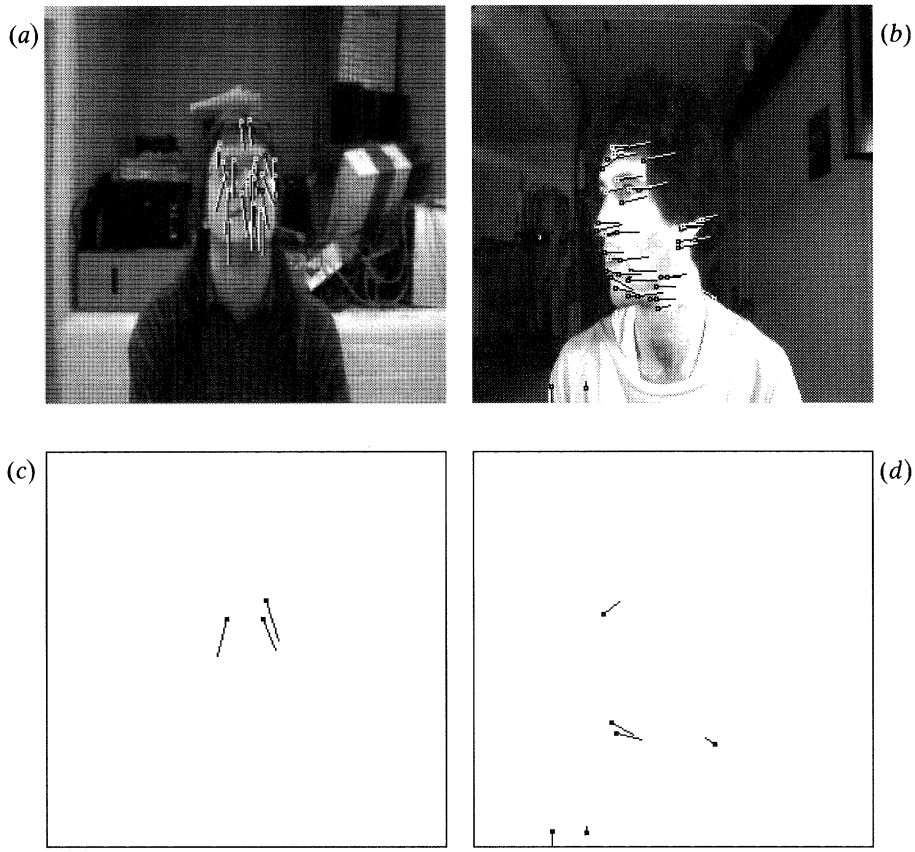


Figure 10. Outlier rejection with a static camera and moving object. Motion vectors are shown double length for clarity: (*a*), (*b*) Initial motion vectors; (*c*), (*d*) Rejected motion vectors.

Although the technique evidently performs well when the outliers are 'randomly' distributed, it is not designed to cope with 'structured noise'. It is therefore unsuitable for segmenting independent motions in a scene, such as multiple moving objects. This is because least squares estimation is severely distorted by multiple populations. Larger-scale segmentation techniques must therefore be used first (see, for example, Torr & Murray 1994), and each object can then be handled in turn by our algorithm.

7. Conclusion

We have proposed a novel scheme for rejecting contaminants from a set of data lying on an $(m-1)$-dimensional hyperplane. The method operates in the OR framework and is based on the simple (yet powerful) principle of an influence function. By assessing the change in the minimum eigenvalue of the scatter matrix when a point is deleted, we can represent the total error in the fit without needing to model how the data points themselves are distributed. The algorithm termination is based on a statistical decision rather than prior knowledge of the number of outliers present. We have also shown this minimum eigenvalue scheme to subsume the more familiar method of residuals, and have investigated its error

characteristics. The successful operation of the scheme has been demonstrated on data from a real application.

One interesting direction for future study is the possibility of weighting the contributions of the various data points based on their influence, downgrading those which appear to be contaminants rather than simply rejecting them outright.

L.S.S. thanks Andrew Zisserman of the Robotics Research Group (RRG) for introducing him to matrix perturbation theory and for insightful comments on early drafts of this paper. We had valuable discussions with Brian Ripley and Matthew Eagle of the Oxford University Statistics Department, as well as with Andrew Blake, Phil McLauchlan and Phil Torr of the RRG. We also thank one of our anonymous referees for his many constructive comments and for pointing out the Golub reference which led to equation (4.6). L.S.S. is supported by an Overseas Research Students Award and by the Foundation for Research Development (RSA). M.B. thanks Nikki Clack and Sara Morris for heroic defence against telephone marauders.

Appendix A. Orthogonal regression

The derivation given below is standard (see, for example, Porrill *et al.* 1986; Murtagh & Heck 1987) and is repeated for completeness. The error function to be minimized is given by equation (2.5),

$$\varepsilon(\boldsymbol{n}, d) = \sum_{i=1}^{n} (\boldsymbol{n}^\top \boldsymbol{r}_i + d)^2 \quad \text{subject to } |\boldsymbol{n}|^2 = 1.$$

We express the constraint by means of a Lagrange multiplier μ,

$$\varepsilon'(\boldsymbol{n}, d) = \varepsilon - \mu(\boldsymbol{n}^\top \boldsymbol{n} - 1),$$

and solve by setting the partial derivatives of the Lagrangian to zero. First,

$$\frac{\partial \varepsilon'}{\partial d} = 0 = \sum_{i=1}^{n} (2\boldsymbol{n}^\top \boldsymbol{r}_i + 2d) \implies d = -\frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{n}^\top \boldsymbol{r}_i) = -\boldsymbol{n}^\top \bar{\boldsymbol{r}},$$

so the hyperplane passes through the data centroid $\bar{\boldsymbol{r}}$. Defining $\boldsymbol{v}_i = \boldsymbol{r}_i - \bar{\boldsymbol{r}}$, $\boldsymbol{W} = \sum_i \boldsymbol{v}_i \boldsymbol{v}_i^\top$ and substituting for d in ε gives

$$\varepsilon'(\boldsymbol{n}) = \sum_{i=1}^{n} (\boldsymbol{n}^\top \boldsymbol{v}_i)^2 - \mu(\boldsymbol{n}^\top \boldsymbol{n} - 1) = \boldsymbol{n}^\top \boldsymbol{W} \boldsymbol{n} - \mu(\boldsymbol{n}^\top \boldsymbol{n} - 1)$$

$$\frac{\partial \varepsilon'}{\partial \boldsymbol{n}} = 0 = 2\boldsymbol{W}\boldsymbol{n} - 2\mu\boldsymbol{n} \implies \boldsymbol{W}\boldsymbol{n} = \mu\boldsymbol{n}.$$

Evidently, $\boldsymbol{n}$ is a unit eigenvector of $\boldsymbol{W}$ corresponding to the eigenvalue μ. To decide *which* eigenvalue, we substitute back into ε:

$$\varepsilon = \boldsymbol{n}^\top \boldsymbol{W} \boldsymbol{n} = \boldsymbol{n}^\top \mu \boldsymbol{n} = \mu.$$

Thus, for minimum ε, μ must be the *minimum* eigenvalue of $\boldsymbol{W}$, and $\boldsymbol{n}$ its associated eigenvector.

Appendix B. Matrix perturbation theory

Let $\boldsymbol{A}$ be an $m \times m$ matrix with eigenvalues $\lambda(\boldsymbol{A}) = \{\lambda_1, \ldots, \lambda_m\}$. The non-zero m-vectors satisfying $\boldsymbol{A}\boldsymbol{u}_j = \lambda_j \boldsymbol{u}_j$ $(j = 1, \ldots, m)$ are the *right* eigenvectors (or simply the eigenvectors) of $\boldsymbol{A}$, while those satisfying $\boldsymbol{q}_j^\top \boldsymbol{A} = \lambda_j \boldsymbol{q}_j^\top$ are the *left*

eigenvectors. The p-norm of the vector $\boldsymbol{x}$ is

$$\|\boldsymbol{x}\|_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{1/p}, \quad 1 \leqslant p \leqslant \infty,$$

so the 2-norm is simply the Euclidean length $|\boldsymbol{x}|$. The p-norm of $\boldsymbol{A}$ is

$$\|\boldsymbol{A}\|_p = \sup_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_p}{\|\boldsymbol{x}\|} = \max_{\|\boldsymbol{x}\|_p = 1} \|\boldsymbol{A}\boldsymbol{x}\|,$$

the p-norm of the longest vector obtained by applying A to a unit p-norm vector. Importantly, $\|\boldsymbol{A}\|_2$ is the square-root of the largest eigenvalue of $\boldsymbol{A}^\top\boldsymbol{A}$ (Wilkinson 1965).

Consider the effect on the eigensolution when $\boldsymbol{A}$ is perturbed to $\boldsymbol{A} + \Delta\boldsymbol{A}$. A well-established body of theory describes the sensitivity of eigenvalues and eigenvectors to perturbations in matrix elements (see, for example, Golub & van Loan 1989; Wilkinson 1965). The relations for the general asymmetric case simplify considerably when $\boldsymbol{A}$ is symmetric, since the left and right eigenvectors are then equal ($\boldsymbol{u}_j = \boldsymbol{q}_j$). A further simplification applies in our case, where $\boldsymbol{A} = \boldsymbol{W}$ and $\Delta\boldsymbol{A} = \Delta\boldsymbol{W}_i = -n\boldsymbol{v}_i\boldsymbol{v}_i^\top/(n-1)$ (see § 4), since $\Delta\boldsymbol{A}$ then has unit rank and its 2-norm equals $n|\boldsymbol{v}_i|^2/(n-1)$. In addition, $\Delta\boldsymbol{W}_i$ is semi-negative definite, so λ_j will never increase when $\boldsymbol{W}$ is perturbed by $\Delta\boldsymbol{W}_i$.

(a) *Eigenvalue perturbation*

Suppose λ_j is a simple (i.e. non-repeated) eigenvalue of $\boldsymbol{A}$, and that the left and right eigenvectors of $\boldsymbol{A}$ have unit 2-norm. If the perturbation $\Delta\boldsymbol{A}$ equals $\epsilon\boldsymbol{B}$, where ϵ is small and $\|\boldsymbol{B}\|_2 = 1$, then it can be shown (Golub & van Loan 1989) that in the neighbourhood of the origin there exist differentiable $\boldsymbol{u}_j(\epsilon)$ and $\lambda_j(\epsilon)$ such that

$$(\boldsymbol{A} + \epsilon\boldsymbol{B})\boldsymbol{u}_j(\epsilon) = \lambda_j(\epsilon)\boldsymbol{u}_j(\epsilon), \quad j = 1, \ldots, m.$$

Differentiating with respect to ϵ (and setting $\epsilon = 0$ in the result) yields

$$\boldsymbol{A}\dot{\boldsymbol{u}}_j(0) + \boldsymbol{B}\boldsymbol{u}_j = \dot{\lambda}_j(0)\boldsymbol{u}_j + \lambda_j(0)\dot{\boldsymbol{u}}_j(0),$$

where $\lambda_j(0) = \lambda_j$ and $\lambda_j(\epsilon) = \lambda_j(0) + \dot{\lambda}_j(0)\epsilon + \frac{1}{2}\ddot{\lambda}_j(0)\epsilon^2 + O(\epsilon^3)$. Premultiplying by $\boldsymbol{q}_j^\top$ and simplifying gives (Golub & van Loan 1989; Wilkinson 1965)

$$\dot{\lambda}_j(0) = \frac{\boldsymbol{q}_j^\top\boldsymbol{B}\boldsymbol{u}_j}{\boldsymbol{q}_j^\top\boldsymbol{u}_j}.$$

If $\boldsymbol{A}$ is symmetric ($\boldsymbol{q}_j = \boldsymbol{u}_j$),

$$\dot{\lambda}_j(0) = \boldsymbol{u}_j^\top\boldsymbol{B}\boldsymbol{u}_j,$$

and since in our case $\epsilon\boldsymbol{B} = \Delta\boldsymbol{W}_i = -n\boldsymbol{v}_i\boldsymbol{v}_i^\top/(n-1)$, the first-order change in λ_j is

$$\dot{\lambda}_j(0)\epsilon = -\frac{n}{n-1}(\boldsymbol{u}_j^\top\boldsymbol{v}_i)^2, \quad j \in \{1 \ldots m\}, \quad i \in \{1 \ldots n\}, \tag{B 1}$$

for the ith data vector and the jth eigenvector (corresponding to eigenvalue λ_j). This varies between 0 (no perturbation) and $-n|\boldsymbol{v}_i|^2/(n-1)$ (maximum perturbation). Evidently, the first-order change is always non-positive. The second-order

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

perturbation can be obtained in a similar fashion (Wilkinson 1965; Hinch 1991):

$$\tfrac{1}{2}\ddot{\lambda}_j(0) = \frac{1}{\boldsymbol{q}_j^{\top}\boldsymbol{u}_j}\sum_{\substack{k=1\\k\neq j}}^{m}\frac{(\boldsymbol{q}_k^{\top}\boldsymbol{B}\boldsymbol{u}_j)(\boldsymbol{q}_j^{\top}\boldsymbol{B}\boldsymbol{u}_k)}{(\lambda_j-\lambda_k)\boldsymbol{q}_k^{\top}\boldsymbol{u}_k}.$$

With $\boldsymbol{A}$ symmetric,

$$\tfrac{1}{2}\ddot{\lambda}_j(0) = \sum_{\substack{k=1\\k\neq j}}^{m}\frac{(\boldsymbol{u}_k^{\top}\boldsymbol{B}\boldsymbol{u}_j)(\boldsymbol{u}_j^{\top}\boldsymbol{B}\boldsymbol{u}_k)}{\lambda_j-\lambda_k},$$

and with $\epsilon\boldsymbol{B} = -n\boldsymbol{v}_i\boldsymbol{v}_i^{\top}/(n-1)$, the second-order change in λ_j is

$$\tfrac{1}{2}\ddot{\lambda}_j(0)\epsilon^2 = \frac{n^2}{(n-1)^2}(\boldsymbol{u}_j^{\top}\boldsymbol{v}_i)^2\sum_{\substack{k=1\\k\neq j}}^{m}\frac{(\boldsymbol{u}_k^{\top}\boldsymbol{v}_i)^2}{\lambda_j-\lambda_k}. \tag{B 2}$$

This second-order variation is small when λ_j is well-separated from the other eigenvalues.

(*b*) *Eigenvector perturbation*

The Taylor series expansion for the jth eigenvector is

$$\boldsymbol{u}_j(\epsilon) = \boldsymbol{u}_j + \epsilon\dot{\boldsymbol{u}}_j(0) + O(\epsilon^2),$$

and Golub & van Loan (1989) give

$$\dot{\boldsymbol{u}}_j(0) = \sum_{\substack{k=1\\k\neq j}}^{m}\frac{\boldsymbol{q}_k^{\top}\boldsymbol{B}\boldsymbol{u}_j}{(\lambda_j-\lambda_k)\ \boldsymbol{q}_k^{\top}\boldsymbol{u}_k}\boldsymbol{u}_k,$$

showing that the sensitivity of the jth eigenvector depends both on eigenvalue sensitivity and on the separation of λ_j from the other eigenvalues. Substituting the expressions for $\boldsymbol{A} = \boldsymbol{W}$ and $\epsilon\boldsymbol{B} = \Delta\boldsymbol{W}_i$ gives

$$\dot{\boldsymbol{u}}_j(0)\epsilon = -\frac{n}{n-1}(\boldsymbol{u}_j^{\top}\boldsymbol{v}_i)\sum_{\substack{k=1\\k\neq j}}^{m}\frac{\boldsymbol{u}_k^{\top}\boldsymbol{v}_i}{\lambda_j-\lambda_k}\boldsymbol{u}_k. \tag{B 3}$$

Appendix C. Variance proofs

(*a*) *Eigenvector covariance matrix*

We have from equation (5.6):

$$\begin{aligned}\boldsymbol{\Gamma}_{\boldsymbol{u}_1} &= E\{\delta\boldsymbol{u}_1\delta\boldsymbol{u}_1^{\top}\} = E\{\hat{\boldsymbol{J}}\,\hat{\boldsymbol{V}}\delta\boldsymbol{V}^{\top}\hat{\boldsymbol{u}}_1\hat{\boldsymbol{u}}_1^{\top}\delta\boldsymbol{V}\,\hat{\boldsymbol{V}}^{\top}\hat{\boldsymbol{J}}^{\top}\}\\ &= \hat{\boldsymbol{J}}E\left\{\left(\sum_{i=1}^{n}\hat{\boldsymbol{v}}_i(\delta\boldsymbol{v}_i^{\top}\hat{\boldsymbol{u}}_1)\right)\left(\sum_{j=1}^{n}\hat{\boldsymbol{v}}_j^{\top}(\delta\boldsymbol{v}_j^{\top}\hat{\boldsymbol{u}}_1)\right)\right\}\hat{\boldsymbol{J}}^{\top}\\ &= \hat{\boldsymbol{J}}\left[\sum_{i=1}^{n}\hat{\boldsymbol{v}}_i\left(\sum_{j=1}^{n}\hat{\boldsymbol{v}}_j^{\top}\hat{\boldsymbol{u}}_1^{\top}E\{\delta\boldsymbol{v}_i\delta\boldsymbol{v}_j^{\top}\}\hat{\boldsymbol{u}}_1\right)\right]\hat{\boldsymbol{J}}^{\top}.\end{aligned}$$

Now

$$E\{\delta \boldsymbol{v}_i \delta \boldsymbol{v}_j^\top\} = \begin{cases} \sigma^2 \left(1 - \frac{1}{n}\right) \boldsymbol{I}, & i = j \\ -\sigma^2 \frac{1}{n} \boldsymbol{I}, & i \neq j. \end{cases}$$

and since $\hat{\boldsymbol{u}}_1^\top \hat{\boldsymbol{u}}_1 = 1$,

$$\hat{\boldsymbol{u}}_1^\top E\{\delta \boldsymbol{v}_i \delta \boldsymbol{v}_j^\top\} \hat{\boldsymbol{u}}_1 = \begin{cases} \sigma^2 \left(1 - \frac{1}{n}\right), & i = j \\ -\sigma^2 \frac{1}{n}, & i \neq j. \end{cases}$$

Noting that $\sum_{j=1}^n \hat{\boldsymbol{v}}_j = \boldsymbol{0}$, we obtain

$$\sum_{j=1}^n \hat{\boldsymbol{v}}_j^\top \hat{\boldsymbol{u}}_1^\top E\{\delta \boldsymbol{v}_i \delta \boldsymbol{v}_j^\top\} \hat{\boldsymbol{u}}_1 = \sigma^2 \hat{\boldsymbol{v}}_i^\top - \frac{\sigma^2}{n} \sum_{j=1}^n \hat{\boldsymbol{v}}_j^\top = \sigma^2 \hat{\boldsymbol{v}}_i^\top,$$

giving

$$\sum_{i=1}^n \hat{\boldsymbol{v}}_i \left(\sum_{j=1}^n \hat{\boldsymbol{v}}_j^\top \hat{\boldsymbol{u}}_1^\top E\{\delta \boldsymbol{v}_i \delta \boldsymbol{v}_j^\top\} \hat{\boldsymbol{u}}_1 \right) = \sigma^2 \sum_{i=1}^n \hat{\boldsymbol{v}}_i \hat{\boldsymbol{v}}_i^\top = \sigma^2 \hat{\boldsymbol{W}}.$$

Since $\hat{\boldsymbol{W}} \hat{\boldsymbol{u}}_j = \hat{\lambda}_j \hat{\boldsymbol{u}}_j$ and $\hat{\boldsymbol{u}}_i^\top \hat{\boldsymbol{u}}_j = \delta_{ij}$,

$$\begin{aligned} \boldsymbol{\Gamma}_{\boldsymbol{u}_1} = \sigma^2 \hat{\boldsymbol{J}} \hat{\boldsymbol{W}} \hat{\boldsymbol{J}}^\top &= \sigma^2 \left(\sum_{k=2}^m \frac{\hat{\boldsymbol{u}}_k \hat{\boldsymbol{u}}_k^\top}{\hat{\lambda}_k} \right) \left(\sum_{l=2}^m \frac{\hat{\boldsymbol{W}} \hat{\boldsymbol{u}}_l \hat{\boldsymbol{u}}_l^\top}{\hat{\lambda}_l} \right) \\ &= \sigma^2 \left(\sum_{k=2}^m \frac{\hat{\boldsymbol{u}}_k \hat{\boldsymbol{u}}_k^\top}{\hat{\lambda}_k} \right) \left(\sum_{l=2}^m \frac{\hat{\lambda}_l \hat{\boldsymbol{u}}_l \hat{\boldsymbol{u}}_l^\top}{\hat{\lambda}_l} \right) = \sigma^2 \sum_{k=2}^m \frac{\hat{\boldsymbol{u}}_k \hat{\boldsymbol{u}}_k^\top}{\hat{\lambda}_k}. \end{aligned}$$

Thus,

$$\boldsymbol{\Gamma}_{\boldsymbol{u}_1} = -\sigma^2 \hat{\boldsymbol{J}}.$$

Note that in § 4, we could not assume $\lambda_1 \approx 0$ in the denominator terms $(\lambda_k - \lambda_1)$ because outliers were present. The approximation does hold once the contaminants are eliminated, since then $\lambda_1 \ll \lambda_k$.

(*b*) *Residual variance*

From equation (5.8), the variance of the residual for point i is

$$\begin{aligned} \mathrm{Var}\{\delta \ell_i\} &= \hat{\boldsymbol{u}}_1^\top \boldsymbol{\Gamma}_v \hat{\boldsymbol{u}}_1 + \hat{\boldsymbol{v}}_i^\top \boldsymbol{\Gamma}_{\boldsymbol{u}_1} \hat{\boldsymbol{v}}_i + 2 \hat{\boldsymbol{v}}_i^\top E\{\delta \boldsymbol{u}_1 \delta \boldsymbol{v}_i^\top\} \hat{\boldsymbol{u}}_1 \\ &= \frac{n-1}{n} \sigma^2 \hat{\boldsymbol{u}}_1^\top \boldsymbol{I} \hat{\boldsymbol{u}}_1 - \sigma^2 \hat{\boldsymbol{v}}_i^\top \hat{\boldsymbol{J}} \hat{\boldsymbol{v}}_i + 2 \hat{\boldsymbol{v}}_i^\top E\left\{ \hat{\boldsymbol{J}} \left(\sum_{j=1}^n \hat{\boldsymbol{v}}_j (\delta \boldsymbol{v}_j^\top \hat{\boldsymbol{u}}_1) \right) \delta \boldsymbol{v}_i^\top \right\} \hat{\boldsymbol{u}}_1 \\ &= \frac{n-1}{n} \sigma^2 \hat{\boldsymbol{u}}_1^\top \hat{\boldsymbol{u}}_1 + \sigma^2 \hat{\boldsymbol{v}}_i^\top \left(\sum_{k=2}^m \frac{\hat{\boldsymbol{u}}_k \hat{\boldsymbol{u}}_k^\top}{\hat{\lambda}_k} \right) \hat{\boldsymbol{v}}_i \\ &\quad + 2 \hat{\boldsymbol{v}}_i^\top \hat{\boldsymbol{J}} \left(\sum_{j=1}^n \hat{\boldsymbol{v}}_j \hat{\boldsymbol{u}}_1^\top E\{\delta \boldsymbol{v}_j \delta \boldsymbol{v}_i^\top\} \hat{\boldsymbol{u}}_1 \right). \end{aligned}$$

Now $\hat{\boldsymbol{u}}_1^\top \hat{\boldsymbol{u}}_1 = 1$ and from apppendix C *a*

$$\sum_{j=1}^{n} \hat{\boldsymbol{v}}_j \hat{\boldsymbol{u}}_1^\top E\{\delta \boldsymbol{v}_i \delta \boldsymbol{v}_j^\top\} \hat{\boldsymbol{u}}_1 = \sigma^2 \hat{\boldsymbol{v}}_i,$$

and thus

$$\mathrm{Var}\{\delta \ell_i\} = \frac{n-1}{n}\sigma^2 + \sigma^2 \sum_{k=2}^{m} \frac{(\hat{\boldsymbol{v}}_i^\top \hat{\boldsymbol{u}}_k)^2}{\hat{\lambda}_k} + 2\sigma^2 \hat{\boldsymbol{v}}_i^\top \hat{\boldsymbol{J}} \hat{\boldsymbol{v}}_i \tag{C 1}$$

$$= \sigma^2 \left[\frac{n-1}{n} + \sum_{k=2}^{m} \frac{(\hat{\boldsymbol{v}}_i^\top \hat{\boldsymbol{u}}_k)^2}{\hat{\lambda}_k} - 2 \sum_{k=2}^{m} \frac{(\hat{\boldsymbol{v}}_i^\top \hat{\boldsymbol{u}}_k)^2}{\hat{\lambda}_k} \right]. \tag{C 2}$$

The final expression therefore simplifies to

$$\mathrm{Var}\{\delta \ell_i\} = \sigma^2 \left[\frac{n-1}{n} - \sum_{k=2}^{m} \frac{(\hat{\boldsymbol{v}}_i^\top \hat{\boldsymbol{u}}_k)^2}{\hat{\lambda}_k - \hat{\lambda}_1} \right].$$

For any point i, $\sum_{k=2}^{m} (\hat{\boldsymbol{v}}_i^\top \hat{\boldsymbol{u}}_k)^2 / \hat{\lambda}_k$ has a maximum value of $(n-1)/n$ and a minimum value of 0. The variance thus has lower and upper bounds of 0 and $(n-1)\sigma^2/n$, respectively.

(*c*) *Residual covariance*

We derive equation (5.11). The covariance between two residuals ℓ_i and ℓ_j $(i \neq j)$ is

$$\begin{aligned} \mathrm{Cov}\{\delta \ell_i, \delta \ell_j\} &= E\{(\hat{\boldsymbol{u}}_1^\top \delta \boldsymbol{v}_i + \hat{\boldsymbol{v}}_i^\top \delta \boldsymbol{u}_1)(\hat{\boldsymbol{u}}_1^\top \delta \boldsymbol{v}_j + \hat{\boldsymbol{v}}_j^\top \delta \boldsymbol{u}_1)\} \\ &= \hat{\boldsymbol{u}}_1^\top E\{\delta \boldsymbol{v}_i \delta \boldsymbol{v}_j^\top\} \hat{\boldsymbol{u}}_1 + \hat{\boldsymbol{v}}_j^\top E\{\delta \boldsymbol{u}_1 \delta \boldsymbol{v}_i^\top\} \hat{\boldsymbol{u}}_1 \\ &\quad + \hat{\boldsymbol{v}}_i^\top E\{\delta \boldsymbol{u}_1 \delta \boldsymbol{v}_j^\top\} \hat{\boldsymbol{u}}_1 + \hat{\boldsymbol{v}}_i^\top \boldsymbol{\Gamma}_{\boldsymbol{u}_1} \hat{\boldsymbol{v}}_j, \end{aligned}$$

and, because the argument in appendix C *a* can also provide the relation $E\{\delta \boldsymbol{u}_1 \delta \boldsymbol{v}_i^\top\} \hat{\boldsymbol{u}}_1 = \hat{\boldsymbol{J}} \sigma^2 \hat{\boldsymbol{v}}_i$,

$$\begin{aligned} \mathrm{Cov}\{\delta \ell_i, \delta \ell_j\} &= -\frac{1}{n}\sigma^2 + \sigma^2 \hat{\boldsymbol{v}}_j^\top \hat{\boldsymbol{J}} \hat{\boldsymbol{v}}_i + \sigma^2 \hat{\boldsymbol{v}}_i^\top \hat{\boldsymbol{J}} \hat{\boldsymbol{v}}_j - \sigma^2 \hat{\boldsymbol{v}}_i^\top \hat{\boldsymbol{J}} \hat{\boldsymbol{v}}_j \\ &= -\frac{1}{n}\sigma^2 - \sigma^2 \hat{\boldsymbol{v}}_j^\top \sum_{k=2}^{m} \frac{\hat{\boldsymbol{u}}_k \hat{\boldsymbol{u}}_k^\top}{\hat{\lambda}_k} \hat{\boldsymbol{v}}_i. \end{aligned}$$

The final form is therefore

$$\mathrm{Cov}\{\delta \ell_i, \delta \ell_j\} = -\sigma^2 \left[\frac{1}{n} + \sum_{k=2}^{m} \frac{(\hat{\boldsymbol{v}}_i^\top \hat{\boldsymbol{u}}_k)(\hat{\boldsymbol{v}}_j^\top \hat{\boldsymbol{u}}_k)}{\hat{\lambda}_k} \right]. \tag{C 3}$$

References

Adcock, R. J. 1877 Note on the method of least squares. *Analyst, Lond.* **4**, 183–184.

Adcock, R. J. 1878 A problem in least squares. *Analyst, Lond.* **5**, 53–54.

Barnett, V. & Lewis, T. 1984 *Outliers in statistical data*, 2nd edn. Wiley.

Belsley, D. A., Kuh, E. & Welsch, R. E. 1980 *Regression diagnostics*. Wiley.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Berztiss, A. T. 1964 Least squares fitting of polynomials to irregularly spaced data. *SIAM Rev.* **6**, 203–227.

Chatfield, C. & Collins, A. J. 1980 *Introduction to multivariate analysis.* Chapman & Hall.

Cook, R. D. & Weisberg, S. 1982 *Residuals and influence in regression.* Chapman & Hall.

Faugeras, O. D., Lustman, F. & Toscani, G. 1987 Motion and structure from motion from point and line matches. In *Proc. 1st Int. Conf. on Computer Vision, London*, pp. 25–34.

Gnanadesikan, R. & Kettenring, J. R. 1972 Robust estimates, residuals and outlier detection with multi-response data. *Biometrics* **28**, 81–124.

Golub, G. H. 1973 Some modified matrix eigenvalue problems. *SIAM Rev.* **15**, 318–334.

Golub, G. H. & van Loan, C. F. 1989 *Matrix computations*, 2nd edn. Johns Hopkins University Press.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. & Stahel, W. A. 1986 *Robust statistics: the approach based on influence functions.* Wiley.

Harter, H. L. 1974*a* The method of least squares and some alternatives. I. *Int. statist. Rev.* **42**, 147–174.

Harter, H. L. 1974*b* The method of least squares and some alternatives. II. *Int. statist. Rev.* **42**, 235–264.

Hawkins, D. M. 1980 *Identification of outliers.* Chapman & Hall.

Hinch, E. J. 1991 *Perturbation methods.* Mexico: Cambridge University Press.

Huber, P. J. 1981 *Robust statistics.* Wiley.

Kanatani, K. 1993 Unbiased estimation and statistical analysis of three-dimensional rigid motion from two views. *IEEE Trans. Pattern Anal. Machine Intell.* **PAMI-15**, 37–50.

Kummel, C. H. 1879 Reduction of observation equations which contain more than one observed quantity. *Analyst, Lond.* **6**, 97–105.

Krazanowski, W. J. 1988 *Principles of multivariate analysis.* Clarendon Press.

Longuet-Higgins, H. C. 1981 A computer algorithm for reconstructing a scene from two projections. *Nature* **293**, 133–135.

Mundy, J. L. & Zisserman, A. 1992 *Geometric invariance in computer vision.* MIT Press.

Murtagh, F. & Heck, A. 1987 *Multivariate data analysis.* Dordrecht: Reidel.

Myers, R. H. 1990 *Classical and modern regression with applications*, 2nd edn. PWS-Kent, USA.

Narula, S. C. & Wellington, J. F. 1982 The minimum sum of absolute errors regression: a state of the art survey. *Int. statist. Rev.* **50**, 317–326.

Pearson, K. 1901 On lines and planes of closest fit to systems of points in space. *Phil. Mag.*, ser. 6, **2**, 559–572.

Porrill, J., Pridmore, T. P., Mayhew, J. E. W. & Frisby, J. P. 1986 Fitting planes, lines and circles to stereo disparity data. *AIVRU Tech. Rep.* 17, University of Sheffield.

Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. 1988 *Numerical recipes in C.* Cambridge University Press.

Rousseeuw, P. J. & Leroy, A. M. 1987 *Robust regression and outlier detection.* Wiley.

Shapiro, L. S. 1993 Affine analysis of image sequences. D.Phil. thesis, Department of Engineering Science, University of Oxford.

Shapiro, L. S. & Brady, J. M. 1993 Rejecting outliers and estimating errors in an orthogonal regression framework. *Tech. Rep.* OUEL 1974/93, Department of Engineering Science, University of Oxford.

Shapiro, L. S., Wang, H. & Brady, J. M. 1992 A matching and tracking strategy for independently moving objects. In *Proc. British Machine Vision Conf.* (ed. D. Hogg & R. Boyle), pp. 139–148. Springer.

Shapiro, L. S., Zisserman, A. & Brady, J. M. 1994 Motion from point matches using affine epipolar geometry. In *Proc. European Conf. on Computer Vision (ECCV'94), Stockholm* (ed. J. O. Eklundh), vol. II, pp. 73–84. Springer.

Stewart, G. W. & Sun, J. 1990 *Matrix perturbation theory.* Academic Press.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Torr, P. H. S. & Murray, D. W. 1993*a* Statistical detection of independent movement from a moving camera. *Image and Vision Computing* **11**, 180–187.

Torr, P. H. S. & Murray, D. W. 1993*b* *Outlier detection and motion segmentation* (ed. P. S. Schenker), vol. 2059, pp. 432–443. *Sensor fusion VI*, Boston, MA: SPIE.

Torr, P. H. S. & Murray, D. W. 1994 Stochastic motion clustering. In *Proc. European Conf. on Computer Vision (ECCV'94), Stockholm* (ed. J. O. Eklundh), vol. II, pp. 328–337. Springer.

Tsai, R. Y. & Huang, T. S. 1984 Uniqueness and estimation of 3D motion parameters of rigid objects with curved surfaces. *IEEE Trans. Pattern Anal. Machine Intell.* **PAMI-6**, 13–27.

Ullman, S. 1979 *The interpretation of visual motion.* MIT Press.

Wang, H. & Brady, J. M. 1992 Corner detection; some new results. *IEE colloquium digest of systems aspects of machine perception and vision.* pp. 1.1–1.4. London: IEE.

Weisberg, S. 1985 *Applied linear regression*, 2nd edn. Wiley.

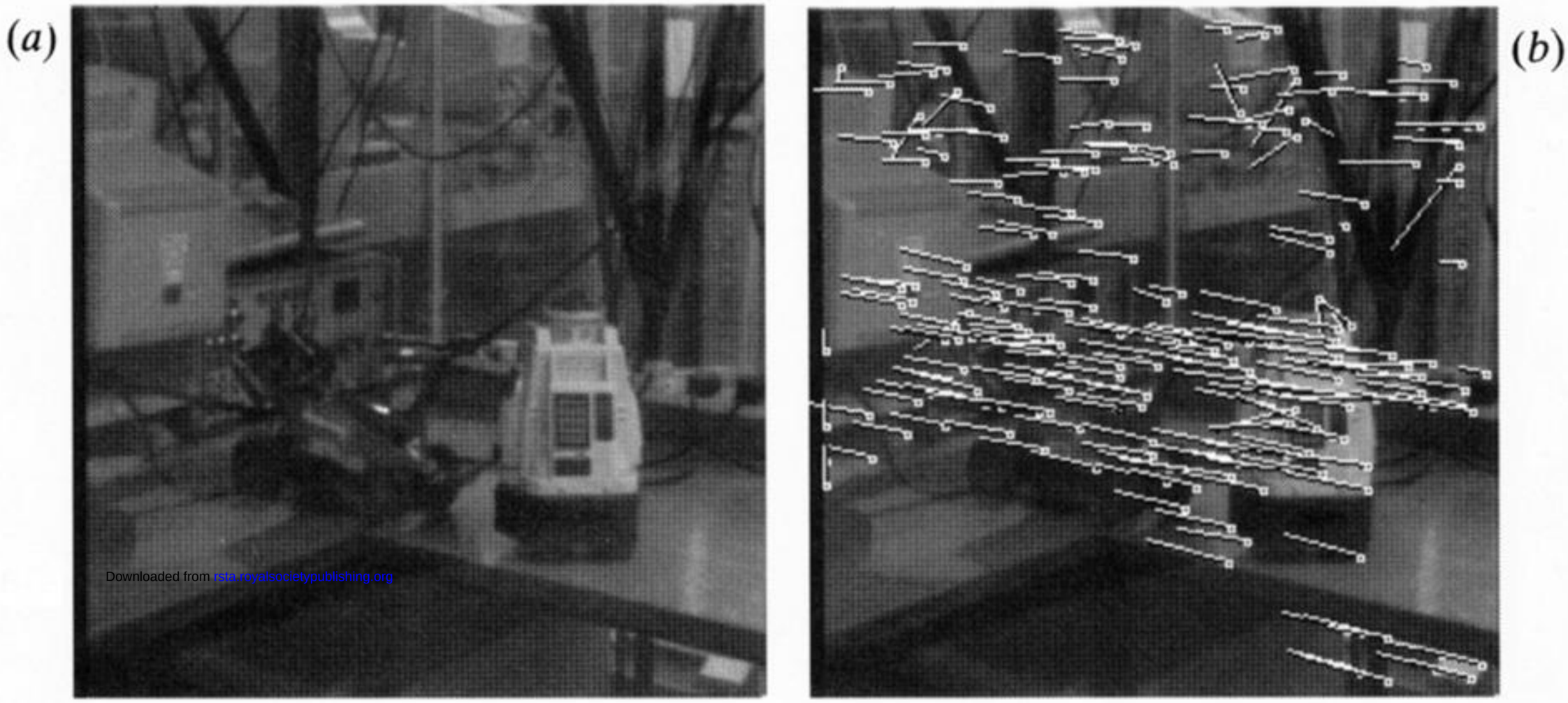
Weng, J., Huang, T. S. & Ahuja, N. 1989 Motion and structure from two perspective views: algorithms, error analysis and error estimation. *IEEE Trans. Pattern Anal. Machine Intell.* **PAMI-11**, 451–476.

Wilkinson, J. H. 1965 *The algebraic eigenvalue problem.* Clarendon Press.

Zisserman, A. 1992 Notes on geometric invariance in vision *BMVC'92 Tutorial*, Leeds.

Received 4 March 1993; revised 2 September 1993; accepted 25 March 1994

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES



PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

igure 9. Outlier rejection with a moving camera and static scene (camera moves right). Motion ectors are shown double length for clarity: (*a*) first frame; (*b*) initial motion vectors; (*c*) rejected otion vectors; (*d*) final motion vectors (outliers removed).

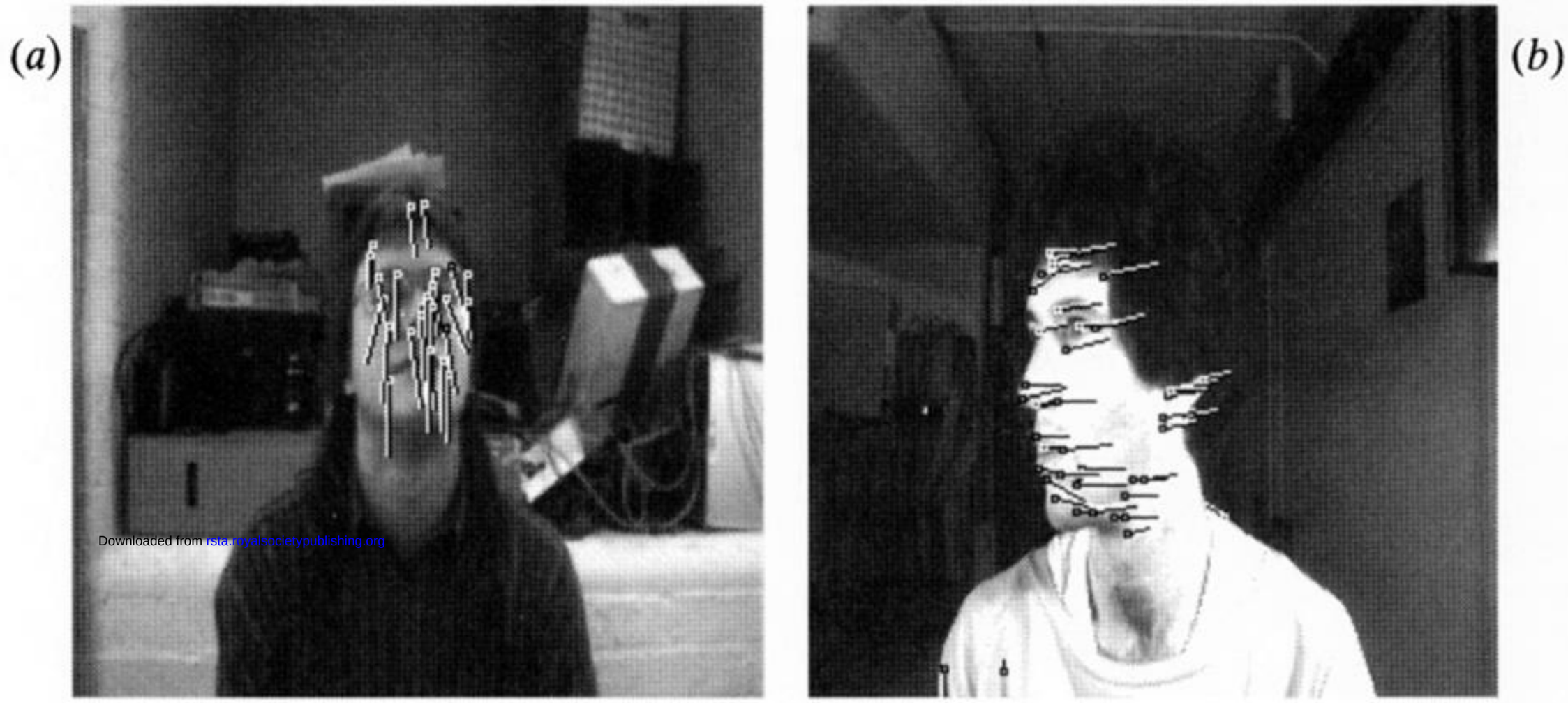


Figure 10. Outlier rejection with a static camera and moving object. Motion vectors are shown double length for clarity: (*a*), (*b*) Initial motion vectors; (*c*), (*d*) Rejected motion vectors.