Robust statistical approaches for local planar surface fitting in 3d laser scanning data

Abdul Nurunnabi [a,*], David Belton [b], Geoff West [b]

[a, b]Department of Spatial Sciences, Curtin University, Perth, Western Australia-6845, Australia

[a]abdul.nurunnabi@postgrad.curtin.edu.au, [b]{D.Belton, G.West}@curtin.edu.au

[a,b]Cooperative Research Centre for Spatial Information

ABSTRACT

This paper proposes robust methods for local planar surface fitting in 3D laser scan data. Searching through the literature revealed that many authors frequently used Least Squares (LS) and Principal Component Analysis (PCA) for point cloud processing without any treatment of outliers. It is known that LS and PCA are sensitive to outliers and can give inconsistent and misleading estimates. RANdom SAmple Consensus (RANSAC) is one of the most well-known robust methods used for model fitting when noise and outliers are present. We concentrate on the recently introduced Deterministic Minimum Covariance Determinant estimator and robust PCA, and propose two variants of statistically robust algorithms for fitting planar surfaces to 3D laser scanning point cloud data. The performance of the proposed robust methods is demonstrated by qualitative and quantitative analysis through several synthetic and mobile laser scanning 3D data sets for different applications. Using simulated data, and comparisons with LS, PCA, RANSAC, variants of RANSAC and other robust statistical methods, we demonstrate that the new algorithms are significantly more efficient, faster, and produce more accurate fits and robust local statistics (e.g., surface normals), necessary for many point cloud processing tasks. Consider one example dataset used consisting of 100 points with 20% outliers representing a plane. The proposed methods called DetRD-PCA and DetRPCA, produce bias angles (angle between the fitted planes with and without outliers) of 0.20^o and 0.24^o respectively, whereas LS, PCA and RANSAC produce worse bias angles of 52.49^o, 39.55^o and 0.79^o respectively. In terms of speed, DetRD-PCA takes 0.033s on average for fitting a plane, which is approximately 6.5, 25.4 and 25.8 times faster than RANSAC, and two other robust statistical methods, respectively. The estimated robust surface normals and curvatures from the new methods have been used for plane fitting, sharp feature preservation and segmentation in 3D point clouds obtained from laser scanners The results are significantly better and more efficiently computed than for existing methods.

Keywords: *3D Modelling, Feature Extraction, Normal Estimation, Outlier, Plane Fitting, Point Cloud, Robustness, Segmentation, Surface Reconstruction*

1. Introduction

Fitting planes and estimating the plane parameters are essential for the analysis of 3D point cloud based representations. Much work has been carried out for accurate local surface fitting and local point set property estimation (e.g. surface normals). In surface reconstruction, the quality of the approximation of the output surface depends on how well the estimated normals approximate the true normals of the sampled surface (Tamal et al., 2005). Surface segmentation, reconstruction, object modelling and rendering are related to each other, are closely related to local normal and curvature estimation and mostly depend on accurate plane fitting (Hoppe et al.,

* Corresponding author.

1992; Li et al., 2010). Moreover, the accuracy of the plane extraction and fitting is important for later steps such as object modelling. Two methods and their variants are popular for plane fitting. These are the Least Squares (LS) technique for model parameter estimation, and Principal Component Analysis (PCA) usually applied for dimension reduction. It is known that these techniques are influenced by outliers and can lead to inconsistent and misleading results (Mitra and Nguyen, 2003). Point cloud data is acquired mostly by various measurement processes using a number of instruments (sensors) and can easily be distorted by the presence of noise and/or outliers. Generally, the physical limitations of the sensors, boundaries between 3D features, occlusions, multiple reflectance and noise can produce off-surface points that appear to be outliers (Sotoodeh, 2006). Many people use RANdom SAmple Consensus (RANSAC) to reduce outlier/noise effects and for robust model parameter estimation (Schnabel et al., 2007). We will show that the RANSAC algorithm is not completely free from the effect of outliers and requires more processing time for large datasets. LS, PCA and RANSAC are currently the three most popular techniques for fitting and/or extracting planes (Hoppe et al., 1992; Pauly et al., 2002; Schnabel et al., 2007; Klasing et al., 2009; Deschaud and Goulette, 2010).

Point clouds come from a number of different sources. Currently most point clouds are captured by laser scanning systems. These produce quite accurate point clouds but suffer from noise, outliers and other effects. If the uncertainties of the sampled points are known, then the outliers can be tested against prior knowledge. However, this is not always possible, or is non-trivial. It has been demonstrated that the uncertainty of a point is highly depended on the attributes of the scanner and the scanner geometry, such as distance and surface orientation (Bae et al., 2005; Soudarissanane et al., 2011). Often this information is not available, such as when a scene comprises multiple co-registered scans acquired from different positions. The properties only relate to a single point, not to the local sampled surface model. The surface properties will be based on a pooled or mixture of variance models from overlapping scans. In recognition of these factors, this paper focuses on examining the points robustly, based on the local neighbourhood distribution.

Robust and diagnostic statistics are two branches of statistics that deal with the problem of outliers. In order to be resilient to outliers, robust statistics have procedures, which are stable with respect to small changes (deviations) in the data, and even large changes in the underlying data pattern cannot cause a complete failure of the procedures. In diagnostic statistics, the outliers are identified, removed, and LS or other traditional methods used to fit the model to the remaining data (Rousseeuw and Leroy, 1987; Stahel and Weisberg, 1991).

In addition to accuracy, we want fast fitting of planar surfaces to be able to efficiently process point clouds that can consist of a large numbers of points, typically in the millions. We describe two novel variants of Deterministic MCD (DetMCD) based diagnostic and robust statistical approaches for planar surface fitting in 3D point cloud data, which are able to find outliers and estimate robust parameters. We compare the new methods with LS, PCA, RANSAC and MSAC (M-estimator SAmple Consensus). We also compare the new algorithms with our previously proposed Fast-MCD (FMCD) based diagnostic and robust PCA dependent methods (Nurunnabi et al., 2012a). The new robust plane fitting methods also produce robust normals and curvatures. The accuracy and robustness of the methods are compared and contrasted with respect to the size of the data, presence of outliers, point density variation, computational speed, and through applying the methods to a number of tasks (e.g. segmentation) in point cloud processing.

The remainder of the paper is arranged as follows: Section 2 presents a short literature review. Section 3 contains brief discussions about a number of relevant established methods, which are compared and contrasted with our new algorithms. In Section 4, we describe our diagnostic and robust statistical algorithms for fitting planes, and for robust normal and curvature estimation. In Section 5, we detail experiments and evaluations and compare the results for the proposed techniques to other existing methods using simulated and real Mobile Laser Scanning (MLS) data sets. Section 6 concludes the paper.

2. Literature review

Many methods related to plane fitting have been explored in different disciplines including computer vision, computer graphics, computational geometry, robotics, photogrammetry, remote sensing, machine learning and statistics. The methods were developed for general plane fitting (Wang et al., 2001; Deschaud and Goulette, 2010), surface reconstruction (Yoon et al., 2007; Sheung and Wang, 2009), sharp feature preserving (Fleishman et al., 2005; Weber et al., 2012) and normal and/or curvature estimation (Mitra and Nguyen, 2003; Crosilla et al., 2009; Li et al., 2010; Boulch and Marlet, 2012). The three main approaches (LS, PCA and RANSAC) have been thoroughly studied and have been used as the foundation for many of the more recently developed methods.

One of the earliest methods of plane fitting and normal estimation proposed by Hoppe et al. (1992) used PCA to estimate tangent planes from the local neighbours of each sampled point. Many authors used the PCA based approach (Pauly et al., 2002; Rabbani, 2006; Belton, 2008) for point cloud processing. The PCA based approach can be formulated as geometric optimization that minimizes a LS cost function, and can be shown to be equivalent to maximum likelihood estimation (Kanatani, 1996). In a study, Klasing et al. (2009) compared a number of optimization and averaging methods and showed that when using a *k*-nearest neighbourhood the *PlaneSVD* (LS) and the *PlanePCA* (PCA) are the two most efficient methods for plane fitting and normal estimation in terms of both quality of results and speed. It is evident that the results from PCA are affected by outlying observations, because the mean and covariance matrix have an unbounded influence function and zero breakdown point (Hampel et al., 1986). Although LS and PCA are not able to identify outliers, many outlier detection approaches such as the 'w-test' (Teunissen, 2000) can be used before fitting by LS and PCA after the identified outliers have been removed. To avoid the influence of outliers/noise on the estimates from PCA, robust versions of PCA have been introduced (Hubert et al., 2005; Feng et al., 2012). Nurunnabi et al. (2012a and 2012b) used the Fast-MCD based outlier detection approach (Rousseeuw and Driessen, 1999) and robust PCA (Hubert et al., 2005) for plane fitting and extraction, respectively. Fleishman et al. (2005) proposed a forward-search based robust moving least squares technique (Alexa et al., 2001; Levin, 2003) for reconstructing piecewise smooth surfaces and for reliable normal estimation. The method can deal with multiple outliers, but requires very dense sampling and a robust initial estimator to start the forward-search. Sheung and Wang (2009) showed that forward-search misclassifies the noisy regions at corners since it fails to obtain a good initial fit in these regions.

The RANSAC algorithm is a model-based approach used frequently for planar surface fitting, extraction and normal estimation (Schnabel et al., 2007; Masuda et al., 2013). Boulaassal et al. (2007) used RANSAC for automatic extraction of planar parts from building façades. Schnabel et al. (2007) developed two optimizations to RANSAC that Deschaud and Goulette (2010) claimed are slow for large datasets. They showed that RANSAC is very efficient for detecting large planes in noisy point clouds but very slow for detecting small planes in large point clouds. The Hough transform (Duda and Hart, 1972) is another model-based method used for detecting parameterized objects in which each data point casts its vote in a parameter space. Vosselman et al. (2004) used the Hough transform to detect geometric shapes in point clouds. However, Deschaud and Goulette (2010) argued that it is too time consuming for fitting a model to a large dataset. Tarsha-Kurdi et al. (2007) showed that the Hough-transform is sensitive to the segmentation parameters values, and RANSAC is more efficient in terms of processing time.

3. Methods used for comparison

3.1. Least squares

Least Squares (LS) minimizes the sum of the squared residuals and has been used in different ways for plane fitting in different applications (Wang et al., 2001; Klasing et al., 2009). For a set of 3D data points $\{p_i(x_i, y_i, z_i); i = 1, \dots, n\}$, the plane equation can be defined as:

$$ax + by + cz + d = 0, \quad (1)$$

where a, b and c are the slope parameters, and d measures the distance of the plane to the origin. In classical LS, the data points are expressed by a functional relation, $z = f(x, y)$, and the sum of the squared residuals in the z direction is minimized:

$$\min \sum_{i=1}^{n} r_i^2 = \min \sum_{i=1}^{n} d_{vi}^2 = \min \sum_{i=1}^{n} (z_i - \hat{z}_i)^2, \quad (2)$$

where the ith residual r_i (d_{vi}) is the vertical distance between the ith point and the fit ($\hat{z}_i$), see Fig. 1a. Minimization of vertical squared errors is not ideal, because it considers errors only the one vertical or z direction (Kwon et al., 2004). To overcome the bias in one direction, the total least squares (Huffel and Vandewalle, 1991) approach is used that minimizes the squared sum of the orthogonal distances (d_{oi}) between the points and the fitted plane, see Fig. 1b:

$$\min \sum_{i=1}^{n} r_i^2 = \min \sum_{i=1}^{n} d_{oi}^2. \quad (3)$$

One of the most common approaches for plane parameter estimation uses the eigenvalue method, which minimizes $\sum_{ij}(ax_{ij} + by_{ij} + cz_{ij} + d)^2$ under the constraint: $a^2 + b^2 + c^2 + d^2 = 1$. This minimization is equivalent to finding the eigenvector that corresponds to the least eigenvalue of the matrix:

$$M = \frac{1}{n}\sum_{ij}\left(x_{ij}, y_{ij}, z_{ij}, 1\right)^T (x_{ij}, y_{ij}, z_{ij}, 1). \quad (4)$$

This method is also known as the *PlaneSVD* method (Klasing et al., 2009).

Figure 1

3.2. Principal component analysis

Principal component analysis is a statistical technique, which is a basis transformation to diagonalize an estimate of the covariance matrix of the data (Schölkopf et al., 1997). PCA works by transforming the variables to a new set of uncorrelated and orthogonal variables that explain the underlying covariance structure of the data. The new set of variables termed Principal Components (PCs) , rank the variability in the data through the variances, and produce the corresponding directions using the eigenvectors of the covariance matrix (Joliffe, 1986; Lay, 2012). For 3D point cloud data, the covariance matrix of n points is defined as:

$$C_{3\times3} = \frac{1}{n}\sum_{i=1}^{n}(p_i - \bar{p})\,(p_i - \bar{p})^T, \quad (5)$$

where $\bar{p}$ is the mean of the data. By performing Singular Value Decomposition (SVD) on the covariance matrix, we get the required PCs or eigenvectors, and the corresponding eigenvalues. The PCs are ranked in order of explanation of the variance, so the first PC is the eigenvector corresponding to the largest eigenvalue (Lay, 2012). For plane fitting, the first two PCs form an orthogonal basis for the plane, and the third PC is orthogonal to the first two and approximates the normal of the fitted plane. PCA gives an equivalent solution to the total least squares formalization of the plane fitting problem. Since the first two PCs explain the variability as much as possible with two dimensions, the fitted plane is the best 2D linear approximation to the data. The third PC corresponding with the least eigenvalue expresses the least amount of variation and is used to get the estimate of the plane parameters. The third eigenvalue can also be used as a measure of the noise level in the data. Classical PCA is also known as *PlanePCA* (Klasing et al., 2009).

3.3. RANSAC and MSAC

Fischler and Bolles (1981) introduced RANdom SAmple Consensus (RANSAC), which is a robust approach used in many applications for extracting shapes and estimating the model parameters from data that may contain outliers. RANSAC classifies data into inliers and outliers. It looks for a minimal subset with maximal support (the number of data points that match with the model). It consists of two steps: hypothesize and test. First, a minimal subset (e.g. three points for a plane) is randomly sampled from the data and the required model parameters are estimated based on the subset. In the second step, the model is compared with the data and its

support is determined. This two-step iterative process continues until the likelihood of getting a model with better support than the current best model is below a given threshold. RANSAC is popular for planar surface fitting because it is conceptually simple and very general (Schnabel et al., 2007). Since its inception, many versions of RANSAC have been proposed (Torr and Zisserman, 2000; Raguram et al., 2008; Choi et al., 2009), but there is no consensus as to which one is the best for each model fitting scenario. For example, RANSAC can be sensitive to the choice of the correct error threshold (T). It finds the minimum of the cost function:

$$C_f = \sum_i \rho(e_i^2), \tag{6}$$

where e_i is the error for the ith observation, and

$$\rho(e^2) = \begin{cases} 0 & e^2 < T^2 \\ \text{constant} & e^2 \geq T^2. \end{cases} \tag{7}$$

Torr and Zisserman (2000) showed that if the threshold (T) is set too high then the robust estimate can be very poor. To address this, they proposed MSAC (M-estimator SAmple Consensus), which minimizes the cost function in Eq. (6) with a robust error function ρ_2 defined as:

$$\rho_2(e^2) = \begin{cases} e^2 & e^2 < T^2 \\ T^2 & e^2 \geq T^2, \end{cases} \tag{8}$$

which is the redescending M-estimator (Huber, 1981; Hampel et al., 1986). The advantage of using MSAC in point cloud data analysis has been demonstrated by Vosselman and Klein (2010). Choi et al. (2009) evaluated the RANSAC family and showed that MSAC is one of the most accurate methods. We consider MSAC for our comparison because it uses a robust M-estimator and has been recognized as one of the most competitive methods.

4. Proposed algorithms

In this section, we propose two variants of diagnostic and robust statistical approaches for local planar surface fitting in laser scanning 3D point cloud data. The methods use the multivariate outlier detection approach and the robust version of PCA.

4.1. Methods used in the proposed algorithms

The proposed algorithms, namely diagnostic PCA and robust PCA, use a robust mean vector (simply called the mean) and a robust covariance matrix to generate a robust distance that can be used for finding outliers and to determine the outlyingness measure in robust PCA. The workflow for the proposed algorithms is shown in Fig. 2.

Figure 2

4.1.1.Robust estimators of mean vector and covariance matrix

In a multivariate setting, we can represent the data set of n observations with m dimensions as a $P_{n\times m}$ matrix, $P = (p_1, \dots, p_n)^T$, with the ith observation $p_i = (p_{i1}, \dots, p_{im})$. The classical mean vector and covariance matrix are the two well-known measures for the location (or centre) and scatter of the data. Both measures have a breakdown point of 0%, meaning a small portion (even just one) of the outliers can completely break the estimates. The Stahel-Donoho approach (Stahel, 1981; Donoho, 1982) was the first high breakdown mean and covariance matrix based estimator used to determine the outlyingness of an observation by looking at all univariate projections. The Minimum Volume Ellipsoid (MVE) is another high breakdown robust estimator introduced by Rousseeuw (1984) that looks for an ellipsoid with the smallest volume that covers a subset of h observations, where $\frac{n}{2} \leq h < n$. The minimum volume ellipsoid has low efficiency due to its low rate of convergence. A better way is the Minimum Covariance Determinant (MCD) also introduced by Rousseeuw (1984). The MCD finds those h observations that have the covariance matrix with the smallest determinant. It has several advantages: (i) better statistical efficiency because it is asymptotically normal. An asymptotically normal

estimator is a consistent estimator whose distribution around the true parameter approaches a normal distribution as the sample size increases (Rousseeuw and Leroy, 1987). In statistics, consistency is evident when the sampling distribution of the estimator becomes increasingly concentrated at the true parameter value as the sample size increases; (ii) better accuracy; (iii) a bounded influence function, where the influence function of an estimator aims to describe the influence of objects upon the estimator (Hampel et al., 1986), with respect to infinitesimal perturbations; and (iv) a breakdown point of 50%, when $h = \lfloor (n + m + 1)/2 \rfloor$ (Rousseeuw and Driessen, 1999). In addition, the MCD is affine equivariant, which makes the estimator independent of the scale of the measurements (Hubert et al., 2012). In spite of its many advantages, it has been rarely used because it is computationally intensive. However MCD has been used as the foundation of the Fast-MCD (Rousseeuw and Driessen, 1999) and Deterministic MCD (Hubert et al., 2012). Fig. 3 shows the stages of the MCD algorithm.

Figure 3

The Fast-MCD (FMCD) is a fast resampling algorithm to efficiently estimate the MCD. It can handle tens of thousands of points. The key component is the *C-step*. For each *C*-step, the Mahalanobis Distances (MDs) are sorted in increasing order and the *h* points having the least MDs are selected. Then the mean and covariance matrix are computed for the *h*-points. Finally the MDs are calculated for all the points using the mean and covariance matrix. The algorithm starts by drawing random initial subsets of size (*m*+1) and performing the *C*-step on them, yielding consecutive subsets of size *h* (simply *h*-subsets) with decreasing determinant of the covariance matrix. To get an outlier-free initial subset of size (*m*+1), many initial random subsets need to be drawn, which is computationally intensive. Rousseeuw and Driessen (1999) fix the number of iterations at 500 to get a good sample and to keep the computation time to an acceptable level. Fixing the number of iterations mainly depends on sample size and assumes the number of outliers in a dataset is less than 50%. For minimizing computational time only two *C-steps* are applied to each initial subset. FMCD uses *selective iteration* and *nested extensions* (when *n* is larger, say $n > 600$) as two further steps to minimize the time. It then keeps the 10 results with the lowest determinants. From these 10 subsets, *C-steps* are performed until convergence to get the final *h*-subset. Convergence occurs when the determinant of the covariance matrix of the *h*-subset is either zero or the determinants of the covariance matrices from two consecutive *h*-subsets are equal. This *h*-subset is later used for determining the FMCD based robust mean vector and covariance matrix. The reader is referred to Rousseeuw and Driessen (1999) for further details about the FMCD.

Hubert et al. (2012) introduced a Deterministic algorithm for the MCD (DetMCD) to get robust location (mean vector) and scatter (covariance matrix). FMCD needs to draw many random (*m*+1)-subsets to obtain at least one outlier-free subset, but DetMCD starts from a few easily computed *h*-subsets and then applies the *C-step* until convergence. It uses the same iteration step but does not draw a random subset, rather it starts from only a few well-chosen initial estimators followed by the *C-steps*. DetMCD couples aspects of both the FMCD and the orthogonalized Gnanadesikan and Kettenring estimators (Maronna and Zamar, 2002). This algorithm is permutation invariant (the result does not depend on the order of the data) and is almost affine equivariant, whereas, FMCD is not permutation invariant. Hubert et al. (2012) claim that DetMCD is much faster than FMCD and at least as robust as FMCD. For more details about DetMCD, the reader is referred to Hubert et al. (2012).

4.1.2. Robust distance

Robust distance is used to find outliers in a local neighbourhood. A general technique for the identification of an outlier in univariate data is based on its distance from the centre of the data. In the case of multivariate data such as a 3D point cloud, this distance is not sufficient for outlier detection; the covariance matrix of the data has to be considered together with the centre. In the following discussion, we will refer to the diagrams in Fig. 4. We generate 30 points in two dimensions that have a linear pattern. We

deliberately change (deviate from the majority pattern) one point in Fig.4a and five points in Fig. 4b to generate single and multiple outliers in the data.

The Mahalanobis Distance (MD, Mahalanobis, 1936) is one of the most well-known distance measures that considers the covariance matrix as well as the mean vector. For an *m*-dimensional multivariate sample (*P*) of size *n*, for the *i*th observation p_i, the MD_i to the centre of the data set *P* is defined as:

$$\mathrm{MD}_i = \sqrt{(p_i - c)^T \Sigma^{-1} (p_i - c)}, \quad i = 1, \ldots, n \tag{9}$$

where *c* is the estimated centre (mean vector) and Σ is the covariance matrix of the sample. Rousseeuw and Driessen (1999) stated that although it is still quite easy to detect a single outlier by means of MD (see Fig. 4a), this approach no longer suffices for multiple outliers because of the masking effect (see Fig. 4b). Masking occurs when an outlying subset goes undetected because of the presence of another, usually adjacent, subset (Hadi and Simonoff, 1993). In Fig. 4b, in the presence of the new set of multiple outliers (at the top-left corner), the MD was unable to identify any outlier even the isolated one (at the right side of the majority points). That is the outliers are now classified or masked as inliers. It shows that the MD is not robust because of the sensitivity of the mean vector and covariance matrix to outliers. It is necessary to use a distance that is based on robust estimators of multivariate location and scatter (Rousseeuw and Leroy, 1987). Many authors use robust estimators to get a robust mean vector and covariance matrix and use them in MD such as in Eq. (9) to obtain a robust version, simply called the robust distance. A highly efficient estimator is the MCD based robust distance (Rousseeuw and van Zomeren, 1990; Rousseeuw and Driessen, 1999). We calculate two versions of robust distances derived from Eq. (9) using FMCD and the DetMCD based mean vectors and covariance matrices. These are called FRD and DetRD, and are defined for the *i*th point, respectively as:

$$\mathrm{FRD}_i = \sqrt{(p_i - c_{FMCD})^T \Sigma_{FMCD}^{-1} (p_i - c_{FMCD})}, \quad i = 1, \ldots, n \tag{10}$$

$$\mathrm{DetRD}_i = \sqrt{(p_i - c_{DetMCD})^T \Sigma_{DetMCD}^{-1} (p_i - c_{DetMCD})}, \quad i = 1, \ldots, n. \tag{11}$$

Rousseeuw and van Zomeren (1990) showed that robust distance follows a Chi-square (χ^2) distribution with *m* degrees of freedom (the number of variables). Although the cut-off value for identifying outliers is to some extent arbitrary and depends on the knowledge of the data, Rousseeuw and Zomeren (1990), and Rousseeuw and Driessen (1999) argue that the observations that have Mahalanobis distance or robust distance (FRD and DetRD) values larger than $\sqrt{(\chi^2_{m,0.975})}$ can be identified as outliers.

For comparison of the three methods, Figs. 4a and 4b show the constructed ellipses for MD, FRD and DetRD values. All the methods are successful in identifying a single outlier (Fig. 4a), but MD fails in the presence of multiple outliers and its ellipse shape and orientation are distracted by the outliers. This is an example of the well-known masking effect. FRD and DetRD successfully identify all five outlying points without the ellipse shape and orientation being affected (Fig. 4b).

Figure 4

4.2. Implementation

Two algorithms are proposed: (i) diagnostic PCA, and (ii) robust PCA. Diagnostics and robust statistics have the same objective of fitting a model that is resilient to outliers. The difference is in the order of the analysis stages. In diagnostic statistics, first the outliers are detected and deleted, and then the remainder of the data is fitted in the classical way. In robust statistics, first a model is fitted that does justice to the majority of observations and then the outliers that have large deviations (e.g. residuals) from the robust fit are identified.

4.2.1. Diagnostic PCA

The algorithm couples outlier diagnostics and classical PCA. First, we find outliers from the data, and then fit a planar surface using PCA to the cleaned data.

For local planar surface fitting, we need to find the local region or neighbourhood of an interest point p_i. For local neighbourhood based point cloud processing, data points from a local planar surface are sampled from within a local fixed radius (r) or within a local neighbourhood of k points. We use the well-known k-Nearest Neighbourhood (k-NN; Fig. 5b) (Samet, 2006) searching technique rather than the fixed distance neighbourhood (Fig. 5a) method, because k-NN is able to avoid the problem of point density variation. We know point density variation is a common phenomenon particularly when we are dealing with Mobile Laser Scanning (MLS) data because of the movement of the data acquisition sensors (or vehicles). A further advantage is that the same size of local neighbourhood produces local statistics (e.g. normal and curvature) of equal support. After fixing a local neighbourhood (Np_i), we find outliers in the neighbourhood using robust distance (FRD or DetRD). We then fit a plane using classical PCA to the cleaned data. The best-fit-plane is obtained by projecting all the inlier points onto the two PCs with the highest eigenvalues. The third PC is the normal to the fitted plane, and the elements of the third eigenvector (PC) are the estimated plane parameters.

Figure 5

The algorithm for diagnostic PCA based on Robust Distance based PCA (called RD-PCA) is described in Algorithm RD-PCA.

Algorithm RD-PCA

(i) Input: point cloud, neighbourhood size k, the size of h used in robust distance, and $\chi^2_{3,0.975} = 3.075$.

(ii) Determine the local neighbourhood Np_i of a point p_i consisting of its k nearest neighbours.

(iii) Calculate robust distance (FRD or DetRD) for each point in Np_i.

(iv) Classify the points in Np_i into inliers and outliers according to the respective FRD or DetRD values and the chosen Chi-square (χ^2) cut-off value.

(v) Perform PCA on the inlier matrix.

(vi) Arrange the three PCs according to their respective eigenvalues.

(vii) Find the two PCs that have the largest eigenvalues, and fit the plane by projecting the points onto the directions of the two PCs.

(viii) Output: normals, eigenvalues and the necessary statistics such as curvature.

The RD-PCA algorithm can be performed in two different ways: using FRD and DetRD in place of RD for finding outliers in the local neighbourhood. For both methods, after finding outliers, classical PCA is performed on the cleaned data set. We name FRD based diagnostics PCA, and DetRD based diagnostics PCA, as FRD-PCA and DetRD-PCA, respectively.

4.2.2. Robust principal component analysis

The robust version of PCA is for determining PCs that are only influenced by outliers to a small extent. Robust PCA (RPCA) algorithms can be categorized according to the dimensionality of the data. For 3D point cloud data, where the number of dimensions is considerably smaller than the number of observations, we are interested in an efficient low-dimensional method. Roughly they can be categorized into two classes: (i) methods that try to find a robust estimation of the covariance matrix (see Section 4.1.1), although these methods are sometimes limited in the case of insufficient data to robustly estimate a high-dimensional covariance matrix, (ii) Projection Pursuit (PP) based methods (Li and Chen, 1985) that try to maximize certain robust estimates of univariate variance to

obtain consecutive directions on which the data are projected. The PP based methods are qualitatively robust and inherit the robustness characteristics of the adopted estimators (Feng et al., 2012). Hubert et al. (2005) combined both the approaches and proposed a robust version of PCA, denoted RPCA. We choose this method because it yields accurate estimates of outlier free data sets, more robust estimates for contaminated data, is able to detect exact-fit situations, is location and orthogonal invariant, and has the further advantages of outlier diagnostics and classification (Hubert et al., 2005).

The RPCA (Hubert et al., 2005) involves the following steps. First, the data are pre-processed using SVD to make sure that the transformed data are lying in a subspace with dimensions less than the number of observations without loss of information. Reducing the data space to the affine subspace spanned by the n observations is especially useful when $m \geq n$, but even when $m < n$, the observations may span less than the whole m-dimensional space (Hubert et al., 2005). Second, a measure of outlyingness for each point is computed by projecting all the data points onto many univariate directions, each of which passes through two individual data points. In order to keep the computation time down, the data set is compressed to PCs defining potential directions. Then, each direction for a point p_i is evaluated for its corresponding value of outlyingness (Stahel, 1981; Donoho, 1982):

$$w_i = \arg\max_v \frac{|p_i v^T - c_{MCD}(p_i v^T)|}{\Sigma_{MCD}(p_i v^T)}, \qquad i = 1, \dots, n \tag{12}$$

where $p_i v^T$ denotes the projection of the ith observation onto the v direction, c_{MCD} and Σ_{MCD} are the MCD based mean vector and covariance matrix in univariate direction v. In the next step, an assumed portion ($h > n/2$) of observations with the smallest outlyingness values are used to construct a robust covariance matrix Σ_h. A larger h can give a more accurate RPCA but less is better for more robust results. Users can fix it according to their own objective and the nature of the data. For example, for our data sets, we use $h = \lceil 0.5 \times n \rceil$ in our algorithms. Then, the method projects all the remaining observations onto the d dimensional subspace spanned by the d largest eigenvectors of Σ_h, computes the mean vector and covariance matrix using a reweighted MCD estimator with weights based on the robust distance of every point. The eigenvectors of this covariance matrix from the reweighted observations are the final robust PCs, and the MCD mean vector serves as a robust mean vector.

The RPCA algorithm can find two types of outliers. One type is an orthogonal outlier that lies away from the subspace spanned by the first two PCs in the case of a plane, and is identified by a large Orthogonal Distance (OD), which is the distance between the observation p_i and its projection $\hat{p}_i$ in the d-dimensional PCA subspace. The orthogonal distance for p_i is defined as:

$$\mathrm{OD}_i = \|p_i - \hat{p}_i\| = \left\|p_i - \hat{\mu}_p - L t_i^T\right\|, \qquad i = 1, \dots, n \tag{13}$$

where $\hat{\mu}_p$ is the robust centre of the data, L is the robust loading (PC) matrix, which contains robust PCs as the columns in the matrix, and $t_i = (p_i - \hat{\mu}_p)L$ is the ith robust score. The other type of outlier is identified by the Score Distance (SD) that is measured within the PCA subspace, and is defined as:

$$\mathrm{SD}_i = \sqrt{\sum_{j=1}^{d} \left(t_{ij}^2 / l_j\right)}, \qquad i = 1, \dots, n \tag{14}$$

where l_j is the jth eigenvalue of the robust covariance matrix Σ_{MCD}, and t_{ij} is the ijth element of the score matrix:

$$T_{n,d} = \left(\mathrm{P}_{n,m} - 1_n c_{MCD}\right) L_{m,d}\,, \tag{15}$$

where $\mathrm{P}_{n,m}$ is the data matrix, 1_n is the column vector with all n components equal to 1, c_{MCD} is the robust centre, and $L_{m,d}$ is the matrix constructed by the robust PCs. OD and SD are shown in Fig. 6b. The cut-off value for the score distance is $\sqrt{(\chi^2_{d,0.975})}$, and for the orthogonal distance is a scaled version of χ^2. A scaled χ^2 ($g_1 \chi^2_{g_2}$) is a version of χ^2 which gives a good approximation of the unknown distribution of ODs, where g_1 and g_2 are two parameters estimated by the method of moments. The reader is referred to Hubert et al. (2005) for a more detailed description of RPCA.

In Fig. 6, we illustrate the orthogonal and score outliers based on 30 artificial points including six (or 20%) outliers projected onto the fitted plane. The points 25, 26 and 27 (green points) are essentially in the plane as their orthogonal distances are low although they are distant from the mean in the plane (score distance). In Fig. 6a, they are identified as good leverage points. Points 28, 29 and 30 (red points) exceed the cut-off value of orthogonal distance so are treated as orthogonal outliers. Projecting these points into the plane show their score distances (Fig. 6b). Note that point 30 has low score distance so would not be identified as an outlier without considering the orthogonal distance. In Fig. 6b, the points 28 and 29 have large orthogonal and score distances, and are treated as bad leverage points (Fig. 6a).

Figure 6

We perform the FMCD and DetMCD based robust PCA algorithms by plugging the FMCD and DetMCD based mean vector and covariance matrix for finding outlying cases into Eq. (12) and in the relevant places of the RPCA algorithm. In robust PCA, we can find orthogonal outliers in Eq. (13) and score outliers in Eq. (14) from the local neighbourhood (Np_i) for every point. The robust plane is obtained by projecting the regular or inlier points onto the two robust PCs for the local surface (Np_i). The third PC is the required robust normal. The resultant eigenvalues and their functions can then be used for estimating local parameters (e.g. curvature). The FMCD and DetMCD versions of RPCA are called FRPCA and DetRPCA, respectively.

5. Experimental results

In this section, we explore the methods and evaluate and compare the results of the new techniques with other methods using simulated and real data sets. Section 5.1 demonstrates and quantifies the abilities of the proposed techniques on simulated data to deal with the presence and effects of outliers, and compares them to the existing techniques LS, PCA, RANSAC, and MSAC. At the same time we show the comparative performance of the results from FMCD based FRD-PCA and FRPCA. Section 5.2 assesses the techniques on real data sets captured from MLS. It demonstrates the ability to more accurately perform common existing point cloud processing techniques in the presence of outliers. Such existing techniques include plane extraction, sharp feature preservation and segmentation.

For evaluation, we fit the planar surfaces by the different methods, and estimate normal and eigenvalue characteristics. To determine performance, we calculate three measures. The first is the bias (dihedral) angle θ (Wang et al., 2001) between the planes fitted to the data with and without outliers, defined as:

$$\theta = arccos|\hat{n}_1^T.\hat{n}_2|, \tag{16}$$

where $\hat{n}_1$ and $\hat{n}_2$ are the two unit normals from the fitted planes with and without outliers, respectively. To avoid the 180° ambiguity of the normal vectors we use the absolute value in Eq. (16). The second is the variation along the plane normal, which is defined by the least eigenvalue λ_o, as shown in Fig.7. The third is the surface variation (Pauly et al., 2002), a measure closely related to mean curvature (Sullivan, 2008), determined along the direction of the corresponding eigenvectors at the point p_i in a neighbourhood of size k that is defined as:

$$\sigma(p_i) = \frac{\lambda_0}{\lambda_0+\lambda_1+\lambda_2}, \qquad \lambda_0 \leq \lambda_1 \leq \lambda_2 \tag{17}$$

where λ_i $(i = 0, 1, 2)$ is the ith eigenvalue, and λ_2 and λ_1 are the two largest eigenvalues corresponding to the first two PCs.

Figure 7

5.1. Simulated data

Simulated data is used to demonstrate and evaluate some typical behaviours including (i) influence on the bias angle, which can be considered as the effect on the estimated plane parameters, (ii) effect on the bias angle of point density variation in different directions and surface roughness, and (iii) classification of points into inliers and outliers. Bias angles are estimated in terms of sample size and the percentage of outlier contamination. Statistical significance tests are used to check for any significant difference between the methods, to rank them, and to reduce the number of methods considered for effective comparison.

The artificial data sets used in this section are generated by randomly drawn points from two sets of multivariate 3D (x, y, z) Gaussian normal distributions. One set is used to generate points on a plane and the other set is for outlying points. Fig. 8 shows regular points are generally in the plane with some variation due to noise, and the outlying cases are far from the planar surface. The regular observations in 3D have means (3, 3, 3) and variances (7, 7, 0.01), and the outlying cases have means (8, 10, 12) and variances (7, 7, 1). We simulate the data sets with different sample sizes (n) and outlier percentages. Fig. 8 also shows the fitted planes for the three methods.

Figure 8

5.1.1. Plane fitting and bias angle evaluation

We simulate 1000 data sets (one example shown in Fig. 8) of 100 points including 20% outliers and fit them to get statistically representative results. We calculate different descriptive measures as shown in Table 1 for bias angles (in degrees). Results show that LS has the largest mean, median and standard deviation (Std. Dev.) for bias angles whereas DetRD-PCA has the smallest. This demonstrates that the bias angles from outlier resistant methods are the lowest for the best-fit-planes with and without outliers present.

To explore the effect of different percentages of outlier contamination on a fixed number of data points, we create 1000 datasets of 100 points with outlier contamination ranging from 1% to 40%. We fit the planes for every data set of 100 points with and without outliers, and calculate average bias angles. In Fig. 9, we plot the average bias angles versus outlier percentages for different methods. In Fig. 9a, it is clear that LS and PCA have very large average bias angles while robust methods have very low bias angles. Removing the LS and PCA results, Fig. 9b, we see that RANSAC and MSAC have worse bias angles compared to the more robust statistical methods (FRD-PCA, DetRD-PCA, FRPCA and DetRPCA). Fig. 9c shows the differences between the robust statistical methods. Results in Table 1 and Fig. 9c show that the two DetMCD based methods DetRD-PCA and DetRPCA generally have less average bias angles than their FMCD based counterparts FRD-PCA and FRPCA.

Table 1

Figure 9

To get more insight into the robustness of the results, we use boxplots (Fig. 10) in which the boxes enclose the middle half of the results, i.e. the length of a box is the interquartile range with the ends of the box at the 1st and 3rd quartiles. The line across the box is the position of the median, and the end of the whiskers shows the minimum and the maximum of the results. The '+' signs represent the outlying results. In Fig. 10, boxplots are created for different methods based on the bias angles (used in Table 1) from 1000 runs for the data of 100 simulated points including 20% outliers. It clearly shows significantly better robustness of the statistical robust methods than LS, PCA, RANSAC and MSAC, and further supports the results for the robust statistical methods shown in Fig. 9.

Figure 10

5.1.2. Sample size and outlier influence on bias angle

To explore the effect of sample size and different percentages of outlier contamination on the bias angle estimation, we generate data sets of various sample sizes (n = 20, 50 and 200) and outlier percentages (1% to 40%). We perform 1000 runs for each and every sample size and outlier percentage. Given the poor performance of LS and PCA (Table 1, and Figs. 9a and 10a), they are ignored in this analysis and we concentrate only on the robust methods.

Results for average bias angles (in degrees) from 1000 data sets are shown in Fig.11. For a small sample of size 20, Fig. 11a, we see that RANSAC, MSAC and DetRPCA give inconsistent results for outlier percentages of 25% and more. The robust statistical methods give better results (i.e. smaller bias angles) than RANSAC and MSAC for increasing sample sizes. For every sample size, DetMCD based methods perform better than the respective FMCD based methods, meaning DetRD-PCA and DetRPCA will produce more accurate results than FRD-PCA and FRPCA, respectively. DetRD-PCA has the smallest bias angle for every sample size and outlier percentage. Fig. 11a also shows that even for low point density and in the presence of a high percentage of outliers, DetRD-PCA performs better than the others.

Figure 11

5.1.3. Statistical significance test

Table 1 shows there is much variability in, and sometimes little difference between, the average bias angles from different methods. We explore the results to determine if there is any statistically significant departure between the relevant pairs of methods. Since the bias angle values do not follow the so-called normality assumption, we perform the non-parametric 'Wilcoxon Signed Rank' statistical significance test (Sheskin, 2004) based on the information from Table 1 and the relevant bias angles from 1000 runs. This test procedure is equivalent to the parametric 'dependent t-test' (Sheskin, 2004), which verifies the difference between two medians (in column 7, Table 1) from two different methods (i.e. populations), in columns 1 and 2 in Table 2. We test the null hypothesis (H_o) with respect to the alternative hypothesis (H_a):

- H_o= *there is no significant difference between two medians from two different methods.*
- H_a= *there is some difference between two medians, that means, the two methods perform significantly different.*

Table 2 shows the results from the 'Wilcoxon Signed Rank' test obtained by using the SPSS® software package. We perform the test at the 5% level of significance. Therefore, we may reject H_o if the calculated p-value (column 3 in Table 2) is less than 0.05, otherwise we may retain H_o. We see only three pairs (i) RANSAC, MSAC (ii) FRD-PCA, DetRD-PCA and (iii) FRPCA, DetRPCA retain H_o, i.e. for the three pairs there is no statistically significant differences between the methods, because respective significant values exceed the assigned significance level (0.05). Therefore, based on the test results, we may reach the decision: the methods in the three pairs perform similarly to each other. The methods in the rest of the pairs have significant differences, because for those pairs we may reject H_o, i.e. one method significantly performs better than the other in the pair. For example, PCA is better than LS, and RANSAC is better than PCA, because for these cases the decisions are: rejected H_o (between the pairs they have significant median difference), and at the same time from Table 1 we get $\text{median (LS)} > \text{median (PCA)}$, and $\text{median (PCA)} > \text{median (RANSAC)}$. Similarly, results from Table 1 and Table 2 illustrate that robust statistical methods perform significantly better than RANSAC and MSAC.

Table 2

In the remainder of this paper, for brevity, we just consider PCA, RANSAC, MSAC, and the deterministic MCD based DetRD-PCA and DetRPCA as the robust statistical methods for comparison and performance evaluation.

5.1.4. Point density variation

To study the effect of point density variations on bias angle, we create data sets with different variations in surface directions. Point density is defined as the number of points that occur in a specific unit volume. To generate data sets of different density, we keep the data size the same but change the variances of the Gaussian distribution from where the data have been drawn randomly. That is, a large variance in the point distribution gives low point density and vice versa. The size of the volume is considered in the surface directions (x and y). The rows of Table 3 contain six sets of variance combinations for regular (R) and outlier (O) data in the x and y directions. We simulate 1000 sets of 100 points with 20% outliers for every variation (I to VI, Table 3). For example, Class I has variances of 3 for regular and 3 for outlier in both x and y directions. The other classes have increased variances and decreased point density. Other parameters are the same as for the previous experiments. Fig. 12a shows that PCA produces larger bias angles than the robust methods, and Fig. 12b shows that both the DetRD-PCA and DetRPCA methods have smaller bias angles than RANSAC and MSAC. In spite of the changes in point density, robust statistical methods produce more consistent results and the performance of DetRD-PCA is better than the others.

Surface roughness may influence the planar surface fitting methods and can change the estimates (Nurunnabi et al. 2012a). We calculate the bias angles for different methods for similar data generated as described in the previous experiments with different z-variances (0.001, 0.01, 0.02, 0.05 and 0.1). With increasing z-variances, results in Figs. 12c and 12d show that DetRD-PCA and DetRPCA perform significantly better, and produce more consistent results than PCA, RANSAC and MSAC.

Table 3

Figure 12

5.1.5. Classification into regular and outlying observations

The robust methods (RANSAC, MSAC, DetRD-PCA and DetRPCA) can be considered as classifiers, because they have the ability to group data into inliers and outliers. To show their performance as classifiers, we generate data sets of 100 points with 20% outliers i.e. 80 inliers and 20 outliers. We run the experiment 100 times and calculate the number of correctly identified outliers and inliers for each of the 100 runs. Fig. 13 shows histograms of the number of inliers identified over all the runs. It shows that most of the time DetRD-PCA and DetRPCA identify allmost all inliers correctly as the histograms for the two new methods are grouped around the 80 inlier point. RANSAC and MSAC identify low percentages (around 20 to 40 out of the 80 inliers), i.e. they falsely show the majority of inliers as outliers. This is the well-known swamping effect. Swamping occurs when good observations are incorrectly identified as outliers because of the presence of another, usually remote, subset of observations (Hadi and Siminoff, 1993). The swamping effect can be considered as the False Positive Rate (FPR) in classification. To evaluate the performance of the classification, we calculate the True Positive Rate (TPR), also known as Sensitivity, the True Negative Rate (TNR), FPR and 'Accuracy' defined by Sokolva et al. (2006):

- $\text{TPR} = \frac{number\ of\ outliers\ correctly\ identified}{total\ number\ of\ outliers} \times 100,$
- $\text{TNR} = \frac{number\ of\ inliers\ correctly\ identified}{total\ number\ of\ inliers} \times 100,$
- $\text{FPR} = \frac{number\ of\ inliers\ identified\ as\ outliers}{total\ number\ of\ inliers} \times 100,$

• Accuracy $= \frac{number\ of\ correctly\ identified\ outliers + number\ of\ correctly\ identified\ inliers}{total\ number\ of\ points} \times 100.$

Results in Table 4 show that RANSAC and MSAC correctly identify outliers but they misclassify inliers as outliers at a very high rate, i.e. RANSAC and MSAC are highly affected by the swamping phenomenon. RANSAC performs slightly better than MSAC in terms of swamping and accuracy. DetRD-PCA and DetRPCA have more than 97% accuracy and less than 4% swamping rate.

Figure 13

Table 4

5.2. Laser scanning data

In this section, results are presented for the plane fitting methods on Mobile Laser Scanning (MLS) data. The data were captured using a system developed by a local survey company. The MLS data has been collected by a vehicle moving at typical traffic speeds. The system's rotating laser collects points along the transport corridor measuring the distance to every object within a 30 metre range of the scanner. The data has been post-processed into (*x*,*y*,*z*) coordinates and has a positional accuracy of approximate 0.015m and a point precision of 0.006m.

This section illustrates that the geometrical features (normal, the least eigenvalue and curvature) estimated by the developed methods can make significant improvements over existing methods and algorithms used for point cloud processing (e.g. reduce over and under segmentation). The performances of the estimated plane parameters and saliency features from the methods are evaluated for the applications (i) plane fitting, (ii) sharp feature preservation and surface edge detection, and (iii) segmentation. Segmentation is the process of labelling a point cloud into a number of homogeneous regions, which is useful for surface reconstruction, object detection and modelling. We use two algorithms: (i) for sharp feature extraction, which is introduced as a classification algorithm (separation of points into border-line points, edge/corner points and surface points) (Nurunnabi et al., 2012b), and (ii) region growing based segmentation (Nurunnabi et al., 2012c). The algorithms are briefly described as follows.

Classification: The classification algorithm estimates λ_o (the least eigenvalue) values for all the points in the data based on their local neighbourhood (*k*-NN), and the *i*th point is identified as an edge/corner points if:

$$\lambda_o > \bar{\lambda}_o + a\,\sigma\,(\lambda_o)\,, \tag{18}$$

where $\bar{\lambda}_o$ and $\sigma\,(\lambda_o)$ are the mean and standard deviation of λ_o, respectively, and a = 1 or 2 or 3 based on knowledge of the data.

Segmentation: Generally, region growing based segmentation algorithms begin by searching for a seed point, assuming that the chosen seed point selection gives better segmentation results. We choose the first seed point as the one with the lowest curvature value (i.e. surface variation, as defined in Eq. (17)) and then grow a region using local surface point proximity (distance between two points) and the coherence criteria (e.g. normal) based on the *k*-nearest neighbourhood Np_i of the *i*th seed point p_i. The algorithm considers the Orthogonal Distance (OD) for every neighbouring point of the *i*th seed point to its best-fit-plane, Euclidian Distance (ED) between the seed point p_i and one of its neighbours p_j, and the angle difference θ between the seed points p_i and p_j defined in Eq. (16), which is calculated depending on the unit normals at p_i and p_j. The region grows with the seed point p_i and one of its neighbours p_j, if they have OD, ED, and θ less than their respective pre-assigned thresholds. The process of segmentation continues with further seed points until all the points in the point cloud have been processed. The regions that have a size, in terms of the number of points, greater than the minimum number of points (R_{min}) will be considered as the final segments for the data.

5.2.1. Plane fitting

The MLS data set in Fig. 14a contains a road scene including a lamp post along with a road sign (indicated by boxes), which looks unclear because of the presence of vegetation around it. We name this data set the '**road scene**' data set. We extract the sign (Fig. 14b, front view and Fig. 14c, side view). This data may be regarded as a planar surface. We see some points created by vegetation in Fig. 14c that are not on the plane and can be considered as outliers. Fig. 15a shows the original points and the plane fitted using PCA. The points that are magenta in color are identified as the points on the fitted plane by PCA, and the points that are green in color show the original positions of the points. Fig. 15c shows the fitted plane contains outliers projected onto the 2D approximation, and the planar surface was not correctly estimated by PCA. These outliers are those to the right of the diagram. This means the outliers appear as inliers in the PCA determined plane, which clearly shows the masking effect caused by the presence of multiple outliers. Fig. 15b shows the fitted plane (in magenta) using DetRD-PCA. Many more points are now correctly identified as outliers and the plane is a better fit. The points classified as part of the extracted plane using DetRD-PCA are shown in Fig. 15d, which matches those points to the left of Fig. 15c.

Figure 14

Figure 15

5.2.2. Sharp feature preservation

A more accurate plane fit will produce more accurate surface normals. Reliable and accurate normals are required to detect and recover sharp features (i.e. lines, edges, corners) (Li et al., 2010). Many algorithms have been developed for sharp feature preservation (Fleishman et al., 2005; Li et al., 2010; Weber et al., 2012). This task is not trivial because of the possible presence of outliers/noise in the data. We will show that our plane fitting algorithms can produce reliable and robust normals, and are better for applications of sharp feature preserving. Sharp features can delineate surface patches and are useful for accurate surface reconstruction.

The normals on or near sharp features become overly smooth mainly because of two reasons: (i) neighbourhood points may be present locally from two or more different surfaces (Fig. 16a), and (ii) presence of outliers/noise (Fig. 16c) in the local neighbourhood. The main strength of the robust statistical methods used in the new algorithms is that they automatically disregard outliers in a neighbourhood and consider the majority of points those are most consistent with themselves. Hence, the fitted plane would be the best-fit-plane for the region (portion) of the majority of points without outliers and the estimated normal represents the surface from which the majority of points came from. In Fig. 16a and Fig. 16c, the non-robust (PCA) method regards all the points for plane fitting in the local neighbourhood (dotted circle), and hence misrepresents the normal and smooths out the sharp features. In Fig. 16b and Fig. 16d robust/diagnostic methods (e.g., DetRD-PCA) consider the majority of points (magenta), ignore the outliers, and fit a plane and normal with the correct orientation. The robust normals (magenta) are correctly estimated on an edge (Fig. 16e) and a corner (Fig. 16f) but non-robust PCA fails to do so. Fig. 16g shows that for a small amount of MLS point cloud data, the orientation of the normals show that PCA makes a sharp edge into a smooth surface, whereas the DetRD-PCA (Fig. 16h) clearly separates the two regions.

Figure 16

To further show the performance for sharp feature recovery, we pick two small sets of real MLS data, acquired by a vehicle based laser scanner. One data set contains part of a road, kerb and footpath (Fig. 17a) and consists of 13,698 points. We call this the '**road-kerb-footpath**' data set, and the other data set is a part of a roof crown extracted from a roadside building (Fig. 17b), called the '**crown**' data set, containing 3,017 points. The '**road-kerb-footpath**' data set consists of edges, and the '**crown**' data set is a polyhedron that consists of edges, corners and bilinear surfaces with common edges. We know the angle of the tangent planes for

bilinear surfaces varies along the edges. The case of varying angles in sharp features is important in real data sets and could cause problems for feature detecting and reconstructing systems using global sets of parameters (Weber et al., 2012). To extract the sharp features, we use the algorithm in Nurunnabi et al. (2012b) fitting the planes for every point in the cloud with a local neighbourhood of size $k = 30$. We calculate the least eigenvalues (λ_o) and classify the points into inliers and outliers according to Eq. (18), where $a = 1$. Results are in Figs. 18 and 19 for the 'road-kerb-footpath' data set and the 'crown' data set, respectively.

The results for the two data sets show that PCA fails to recover the edge/corner points. Although RANSAC and MSAC are robust methods, they do not successfully classify the surfaces, edges and corners. Many surface points (e.g. in regions I, II and III of the 'road-kerb-footpath' data set) appear as edge points. Figs. 18 and 19 show that the DetRD-PCA and DetRPCA methods are more accurate than PCA, RANSAC and MSAC. Figs. 19d and 19e show that DetRD-PCA and DetRPCA efficiently recover sharp features for the 'crown' data set in the presence of bilinear surfaces.

Figure 17
Figure 18
Figure 19

5.2.3. Segmentation

We evaluate the resultant normals and curvatures (defined in Eq. (17)) obtained by existing methods: PCA, RANSAC and MSAC; and new methods: DetRD-PCA and DetRPCA, and compare them for segmentation using region growing. To see the robustness for the estimated curvatures from the different methods used in seed point selection for the segmentation algorithm, boxplots are generated for the curvatures obtained from the 'crown' data set with neighbourhood size 50 (see Fig. 20). We see DetRD-PCA and DetRPCA produce more robust curvatures than PCA, RANSAC and MSAC.

Figure 20

We use the segmentation algorithm (Nurunnabi et al., 2012c) described earlier that uses curvature and normals. The segmentation results from the different methods are evaluated using two MLS data sets consisting of planar and non-planar complex object surfaces.

Dataset 1. 'Crown' data set

To evaluate the segmentation algorithm, we consider the 'crown' data set (Fig. 17b) that consists of 12 different regions (segments). We set the required parameters: $k = 50$, angle threshold $\theta_{th} = 5°$, and minimum region size $R_{min} = 10$. Segmentation results are in Fig. 21 and summarized in Table 5. Figs. 21a and 21b show that PCA and RANSAC give similar results and failed to segment all the surface segments properly. Table 5 shows that both PCA and RANSAC have only two Proper Segments (PS) with seven and eight Over Segments (OS), respectively, and four Under Segments (US) in each. A Proper Segment is identified as a true segment from manually determined ground truth i.e. one segment describes a single feature such as the wall of a house that is one planar surface.. An Over Segment is where one true segment is broken into two or more separate segments, and an Under Segment is where more than one true segments are wrongly grouped together as one segment. Although MSAC performs better than RANSAC, it still has six OS and three US. Using the normals and curvature from the proposed robust statistical methods the same segmentation algorithm performs very well. DetRD-PCA (Fig. 21d) and DetRPCA (Fig. 21e) based segmentations both have eleven PS, one OS and zero US.

Figure 21

Dataset 2. 'Traffic furniture' data set

Our second data set (Fig. 22a) is also MLS point cloud data representing road side objects including a lamp post, sign posts and road ground surfaces. We call this the 'traffic furniture' data set that consists of 23,306 points and includes 12 different planar and

non-planar complex object surfaces. One surface highlighted in the box inset in Fig. 22(a) is of a cylinder that joins seamlessly to an approximately toroidal surface. We set parameters for the segmentation algorithm: $k = 50$, $\theta_{th} = 15°$, and $R_{min} = 10$. Fig. 22 shows the quality of the segmentation results from different methods. In Table 5, the results for the 'traffic furniture' data set show that the segmentation based on PCA, RANSAC and MSAC are not accurate, as they are influenced by over segmentation. RANSAC and MSAC have nine PS with five and three OS, respectively. Figs. 22e, and 22f show that the two sets of DetMCD based segmentation results from DetRD-PCA and DetRPCA, respectively, are accurate without any OS and US occuring. That means the normals and curvatures estimated from the proposed diagnostic and robust methods, which are used in the segmentation purposes, are more reliable, robust and accurate than for the other methods.

Figure 22

Table 5

5.3. Computational speed and effort

We know that for many algorithms, there is a trade-off between computational speed and accuracy of the results. In this paper, we have given priority to accuracy and robustness of the results. We now consider the speed of computation. It has been demonstrated in the previous sections that the robust methods produce significantly better results than the classical methods in terms of accuracy and the ability to deal well with the effects of outliers. In this section, we compare the computational speed only for the robust methods as these are the only methods that give acceptable results.

A major issue for the MCD algorithm is that it is computationally intensive. Both the FMCD and DetMCD algorithms are developed to increase the computational efficiency of the MCD algorithm without loss of accuracy and robustness of the estimators. Hubert et al. (2012) demonstrated and showed the computational efficiency of DetMCD over FMCD. We investigate the computational efficiency empirically for the devised algorithms that use DetMCD estimators and compare with FMCD based methods along with RANSAC and MSAC using existing MATLAB® functions.

To evaluate the computational speed of the proposed algorithms for plane fitting, we simulate data sets as for the previous experiments in Section 5.1 of different sample sizes 20, 50, 100, 500, 1000 and 10000 with 20% outliers. We simulate each of the data sets 1000 times. Results in Table 6 are the times (in seconds) for average plane fitting calculated by using the MATLAB® profile function. Results in Table 6 show that every variant of the DetMCD based method is significantly faster than the respective FMCD based method. For example, for a sample size of 50, FRD-PCA takes 0.815 Sec. and DetRD-PCA takes 0.027 Sec., which is 30 times faster, whereas RANSAC takes 0.205 Sec., which is 7.59 times slower than DetRD-PCA. For a sample size of 10000, DetRD-PCA fits a plane in 0.447 Sec, which is 2.56 and 5.34 times faster than FRD-PCA (1.146 Sec.) and RANSAC (2.389 Sec.), respectively. MSAC takes a little more time than RANSAC, and DetRPCA takes more time than DetRD-PCA. Therefore, it shows that the new methods are faster than the others for a large range of data set sizes. The algorithms are much faster for small sample sizes, which is an advantage as they will reduce the time for any local neighbourhood based point cloud processing tasks, where the local saliency features (e.g., normals and curvatures) are used.

Table 6

From a theoretical point of view, the evaluation of computational effort is not trivial for the new robust statistical algorithms. To find out more about the computational effort of RANSAC, FMCD and DetMCD algorithms, readers are referred to Zuliani (2011), Rousseeuw and Driessen (1999) and Hubert et al. (2012), respectively. We implement all the algorithms using existing MATLAB® functions assuming that they are efficiently implemented. RANSAC and MSAC algorithms used Zuliani's RANSAC toolbox (see

Zuliani, 2011) and the necessary functions for FMCD and DetMCD based algorithms are performed using the MATLAB® library for robust analysis (see Hubert et al., 2005; Hubert et al., 2012).

6. Conclusions

This paper introduces two variants of Deterministic MCD based diagnostic PCA (DetRD-PCA) and robust PCA (DetRPCA) algorithms for fitting planar surfaces in 3D laser scanning point cloud data. Experiments based on simulated and real mobile laser scanning data sets show that the new techniques outperform classical methods (LS and PCA) and are more robust than RANSAC, MSAC and Fast-MCD based methods (FRD-PCA and FRPCA). Results from a statistical significance test (Wilcoxon Signed Rank test) show that the new algorithms are significantly more accurate than LS, PCA, RANSAC and MSAC. The new methods give better results in terms of (i) different percentage of outlier contamination, (ii) size of the data, (iii) point density variation, and (iv) classification of data into inliers and outliers. To quantify, e.g. for accuracy of plane fitting, for a sample of 100 points with 20% outliers, the proposed methods DetRD-PCA and DetRPCA have bias angles (angle between the two planes fitted to the data with and without outliers) of 0.20^{o} and 0.24^{o}, whereas, LS, PCA and RANSAC have bias angles 52.49^{o}, 39.55^{o} and 0.79^{o}, respectively. In terms of speed, DetRD-PCA takes 0.033s on average for plane fitting, which is approximately 6.5, 25.4 and 25.8 times faster than RANSAC, FRD-PCA and FRPCA, respectively. The methods classify outliers and inliers and can reduce masking and swamping effects. By contrast, RANSAC and MSAC misclassify inliers and outliers and are highly affected by the swamping phenomenon. Hence the resultant planes from RANSAC and MSAC are ill-fitted. The proposed algorithms are significantly faster than RANSAC, MSAC, and their robust counterparts FRD-PCA and FRPCA. The normals and curvatures estimated from the new methods are more accurate and robust than the others. Results, using the normals and curvatures from the algorithms in the experiments based on MLS data (for planar and non-planar incomplete complex object surfaces) for plane fitting, sharp feature preservation/recovery and segmentation tasks are more accurate and robust. Using the robust and accurate normals and curvature values it is possible to reduce over and/or under segmentation in a region growing based segmentation process. Overall results show that the DetRD-PCA and DetRPCA are comparable to each other. We observe that DetRPCA gives inconsistent results for a small sample size when combined with a high percentage of outlier contamination. It is also demonstrated that DetRD-PCA performs better than DetRPCA for low point density and a high percentage of outliers. The proposed methods have the potential for improved surface reconstruction, registration and 3D modeling, as well as for other applications.

The new algorithms are similar to many other robust techniques in that they are not suitable when there is more than 50% outliers and/or noise present. Future research will investigate non-planar and non-smooth surface extraction and fitting tasks.

Acknowledgments

We are grateful to McMullan Nolan and partners for the supply or the real mobile mapping laser data.

References

Alexa, M., Behr, J., Cohen-Or, D., Fleishman, S., Levin, D., Silva, C. T., 2001. Point set surfaces. In: Proceedings of the 12th IEEE conference on Visualization, San Diego, California, USA, 21–26 October, pp. 21–28.

Bae, K.–H., Belton, D., Lichti, D. D., 2005. A framework for position uncertainty of unorganised three-dimensional point clouds from near-monostatic laser scanners using covariance analysis. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences 36(3/W19), 7–12.

Boulch, A., Marlet, R., 2012. Fast and robust normal estimation for point clouds with sharp features. Computer Graphics Forum 31(5), 1765–1774.

Belton, D., 2008. Classification and segmentation of 3D terrestrial laser scanner point cloud. PhD thesis, Department of Spatial Sciences, Curtin University of Technology, Australia.

Boulaassal, H., Landes, T., Grussenmeyer, P., Tarsha-Kurdi, F., 2007. Automatic segmentation of building facades using terrestrial laser data. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences 36(3/W52), 12–14.

Choi, S., Kim, T., Yu, W., 2009. Performance evaluation of RANSAC family. In: Proceedings of British Machine Vision Conference, London, UK, 7–10 September, pp. 1–12.

Crosilla, F, Visintini, D, Sepic, F., 2009. Reliable automatic classification and segmentation of laser point clouds by statistical analysis of surface curvature values. Applied Geomatics 1, 17–30.

Deschaud, J-E., Goulette, F., 2010. A fast and accurate plane detection algorithm for large noisy point clouds using filtered normals and voxel growing. In: Proceedings of the 5th International Symposium on 3DPVT, Paris, France, 17–20 May.

Donoho, D. L., 1982. Breakdown properties of multivariate location estimators. Qualifying paper, Harvard University, Boston.

Duda, R. O., Hart, P.E., 1972. Use of Hough transformation to detect lines and curves in pictures. Communication of the ACM 15(1), 11–15.

Feng, J., Xu, H., Yan, S., 2012. Robust PCA in high-dimension: A deterministic approach. In: Proceedings of the 29th International Conference on Machine Learning, Edinburgh, Scotland, UK, 26 June –1 July, pp. 249 – 256.

Fischler, M. A., Bolles, R. C., 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM 24, 381–395.

Fleishman,S., Cohen-Or,D., Silva,C.,2005. Robust moving least-squares fitting with sharp features. ACM Transaction on Graphics 24(3), 544 –552.

Hadi, A. S., Simonoff, J. S., 1993. Procedures for the identification of outliers. Journal of the American Statistical Association 88, 1264 –1272.

Hampel, F., Ronchetti, E., Rousseeuw, P. J., Stahel, W., 1986. Robust Statistics: The approach based on influence functions. John Wiley, NY.

Hoppe, H., De Rose, T., Duchamp, T., 1992. Surface reconstruction from unorganized points. ACM Transaction Computer Graphics 26(2), 71–78.

Huber, P. J., 1981. Robust Statistics. John Wiley, New York.

Hubert, M., Rousseeuw,P.J., Branden, K.V., 2005. ROBPCA: A new approach to robust principal component analysis. Technometrics 47(1), 64–79.

Hubert, M., Rousseeuw, P. J., Verdonck, T., 2012. A deterministic algorithm for robust scatter and location. Journal of Computational and Graphical Statistics 21(3), 618 – 637.

Huffel, S. V., Vandewalle, J., 1991. The Total Least Squares Problem: Computational Aspects and Analysis, SIAM, Philadelphia, PA, USA.

Joliffe, I. T., 1986. Principal Component Analysis. Springer, NY, USA.

Kanatani, K., 1996. Statistical Optimization for Geometric Computation: Theory and Practice. Elsevier Science, Amsterdam.

Kwon, S.-W., Bosche, F., Kim, C., Haas, C. T., Liapi, K. A., 2004. Fitting range data to primitives for rapid local 3D modeling using sparse range point clouds. Automation in Construction 13, 67– 81.

Klasing, L., Althoff, D., Wollherr, D., Buss, M., 2009. Comparison of surface normal estimation methods for range sensing applications. In: Proceedings of the IEEE International Conference on Robotics and Automation, Kobe, Japan, 12–17 May, pp. 3206 –3211.

Lay, D. C., 2012. Linear Algebra and Its Applications. Fourth Edition. Pearson, Boston, USA.

Levin, D., 2003. Mesh-independent surface interpolation. Geometric Modeling for Scientific Visualization, Springer, Berlin, Heidelberg, 37– 49.

Li, G., Chen, Z., 1985. Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo. Journal of the American Statistical Association 80(391), 759 –766.

Li, B., Schnabel, R., Klein, R., Cheng, Z., Dang, G., Jin, S., 2010. Robust normal estimation for point clouds with sharp features. Computers & Graphics 34, 94 –106.

Mahalanobis, P. C., 1936. On the generalized distance in statistics. In: Proceedings of the National Institute of Science in India, vol. 12, pp. 49 –55.

Maronna, R. A., Zamar, R. H., 2002. Robust estimates of location and dispersion for high-dimensional datasets. Technometrics 44(4), 307– 317.

Masuda, H., Tanaka, I., Enomoto, M., 2013. Reliable surface extraction from point-clouds using scanner-dependent parameters. Computer Aided Design and Applications 10(2), 265 –277.

Mitra, N. J., Nguyen, A., 2003. Estimating surface normals in noisy point cloud data. In: Proceedings of the 19th ACM Symposium on Computational Geometry, San Diego, California, USA, 8 –10 June, pp. 322 – 328.

Nurunnabi, A., Belton, D., West, G., 2012a. Diagnostic-robust statistical analysis for local surface fitting in 3d point cloud data. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 1–3, Melbourne, Australia, pp. 269 –274.

Nurunnabi, A., Belton, D., West, G., 2012b. Robust segmentation for multiple planar surface extraction in laser scanning 3D point cloud data. In: Proceedings of the 21st International Conference on Pattern Recognition, Tsukuba Science City, Japan, 11–15 November, pp. 1367–1370.

Nurunnabi, A., Belton, D., West, G., 2012c. Robust segmentation in laser scanning 3D point cloud data. In: Proceedings of the Digital Image Computing: Techniques and Applications, Fremantle, Australia, 3–5 December, pp. 1– 8.

Pauly, M., Gross M., Kobbelt, L. P., 2002. Efficient simplification of point sample surface. In: Proceeding of the Conference on Visualization, Washington, D.C., 27 October – 1 November, pp. 163 –170.

Rabbani, T., 2006. Automatic reconstruction of industrial installations using point clouds and images. PhD Thesis, NCG, Nederlandse Commissievoor Geodesie, Netherlands Geodetic Commission, Delft, The Netherlands.

Raguram, R., Frahm, J. M., Pollefeys, M., 2008. A comparative analysis of RANSAC techniques leading to adaptive real-time random sample consensus. In: Proceeding of the 10th European Conference on Computer Vision, Marseille, France, Part II, 12–18 October, pp. 500 –513.

Rousseeuw, P. J., 1984. Least median of squares regression. Journal of the American Statistical Association 79(388), 871 – 880.

Rousseeuw, P. J., Leroy, A., 1987. Robust Regression and Outlier Detection. John Wiley, NY.

Rousseeuw, P. J., Driessen, K. V., 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41(3), 212 – 223.

Rousseeuw, P. J., van Zomeren, B. C., 1990. Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association 85(411), 633 – 639.

Samet, H., 2006. Foundations of Multidimensional and Metric Data Structures. Morgan Kaufmann, San Francisco, USA.

Schnabel, R., Wahl, R., Klein, R., 2007. Efficient RANSAC for point-cloud shape detection. Computer Graphics Forum, Blackwell Publishing, 26 (2), 214 – 226.

Schölkopf, B., Smola, A., Müller, K-R., 1997. Kernel principal component analysis. In: Proceedings of the 7th International Conference on Artificial Neural Networks, Lausanne, Switzerland, 8 –10 October, pp. 583 – 588.

Sheung, H., Wang, C. C., 2009. Robust mesh reconstruction from unoriented noisy points. In: Proceedings of the SIAM/ACM Joint Conference on Geometric and Physical Modeling, San Francisco, California, USA, 4 – 8 October, pp. 13 – 24.

Sheskin, D. J., 2004. Handbook of Parametric and Nonparametric Statistical Procedures. Third Edition, Chapman and Hall/CRC, USA.

Sokolova, M., Japkowicz, N., Szpakowicz, S., 2006. Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. In: Proceedings of the 19th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, 4– 8 Dec., pp. 1015–1021.

Sotoodeh, S., 2006. Outlier detection in laser scanner point clouds. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences 36 (Part 5), 297– 302.

Soudarissanane, S., Lindenbergh, R., Menenti, M., Teunissen, P., 2011. Scanning geometry: Influencing factor on the quality of terrestrial laser scanning points. ISPRS Journal of Photogrammetry and Remote Sensing 66 (4), 389 –399.

Stahel, W. A., 1981. Robust Estimation: Infinitesimal optimality and covariance matrix estimators. PhD Thesis, Department of Mathematics, Eidgenössische Technische Hochschule (ETH), Zurich, Switzerland.

Stahel, W., Weisberg, S., 1991. Direction in robust statistics and diagnostics. Part II, The IMA Volume in Mathematics and its Applications 34, Springer-Verlag, NY.

Sullivan, J. M., 2008. Curvature measures for discrete surfaces. In Desbrun, M., Grinspun, E., Schroder, P., Wardetzky, M. (Eds.), Discrete Differential Geometry: An Applied Introduction, SIGGRAPH Asia 2008 Course Notes, pp. 10–13. http:// ddg.cs.columbia.edu/ SIGGRAPH ASIA08/ Siggraph Asia2008DDGCourse.pdf (Accessed 10 March, 2014).

Tamal, K. D., Gang, L., Sun, J., 2005. Normal estimation for point cloud: A comparison study for a Voronoi based method. In: Proceedings of the Eurographics Symposium on Point-Based Graphics. NY, USA, 20 – 21 June, pp. 39 – 46.

Tarsha-Kurdi, F., Landes, T., Grussenmeyer, P., 2007. Hough-transform and extended RANSAC algorithms for automatic detection of 3D building roof planes from LiDAR data. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 36(Part 3/ W52), 407– 412.

Torr, P. H. S., Zisserman, A., 2000. MLESAC: A new robust estimator with application to estimating image geometry. Journal of Computer Vision and Image Understanding 78(1), 138 –156.

Teunissen, P. J. G., 2000. Testing Theory: An Introduction. Delft University Press, Delft.

Vosselman, G., Klein, R., 2010. Visualization and Structuring of Point Clouds. In Vosselman, G., Maas, H. –G. (Eds.) Airborne and Terrestrial Laser Scanning, Whittles Publishing, Scotland, UK.

Vosselman, G., Gorte, B. G. H., Sithole, G., Rabbani, T., 2004. Recognizing structure in laser scanner point clouds. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 46 (8), 33–38.

Wang, C., Tanahashi, H., Hirayu, H., 2001. Comparison of local plane fitting methods for range data. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai, HI, USA, 8–14 December, vol. 1, pp. 663 – 669.

Weber, C., Hahmann, S., Hagen, H., Bonneau, G-P., 2012. Sharp feature preserving MLS surface reconstruction based on local feature line approximations. Graphical Models 74(6), 335–345.

Yoon, M., Lee, Y., Lee, S., Ivrissimtzis, I.Seidel, H-P., 2007. Surface and normal ensembles for surface reconstruction. Computer-Aided Design 39 (5), 408 – 420.

Zuliani, M., 2011. RANSAC for Dummies, http://vision.ece.ucsb.edu/~zuliani/Research/RANSAC4Dummies.pdf (Accessed 05 March, 2012).

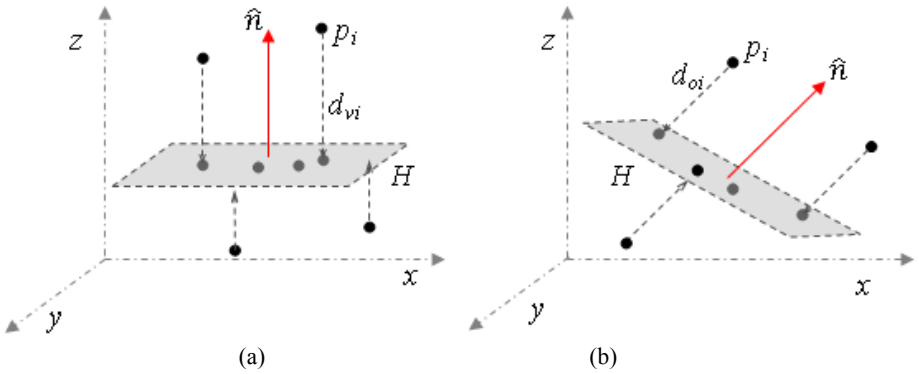


Fig. 1. Fitted plane and estimated normal: (a) least squares, and (b) total least squares.

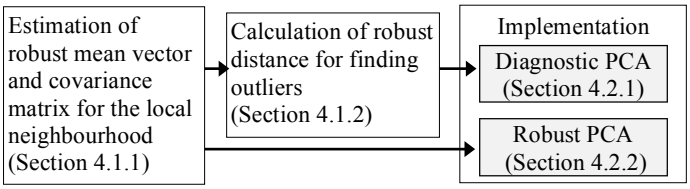


Fig. 2. Work flow for the proposed algorithms.

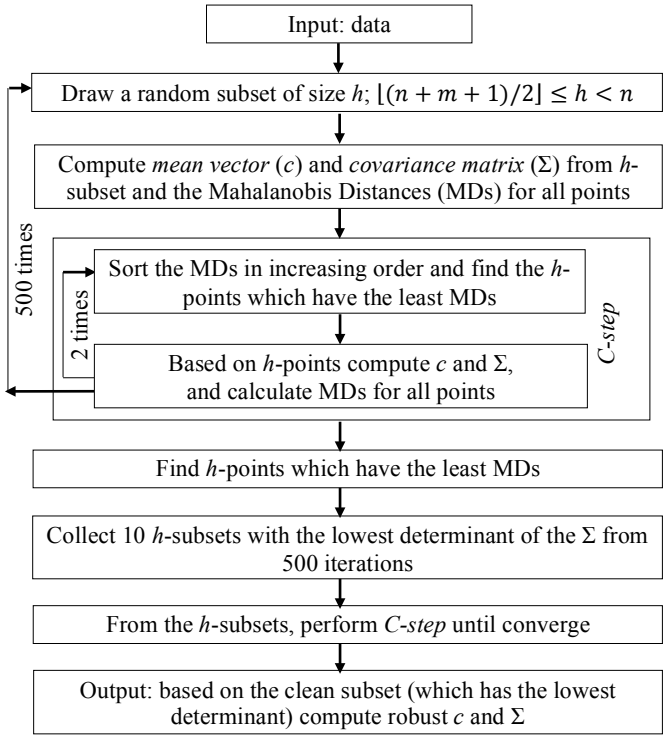


Fig. 3. Minimum Covariance Determinant (MCD) algorithm; n = sample size, and m = the data dimension.

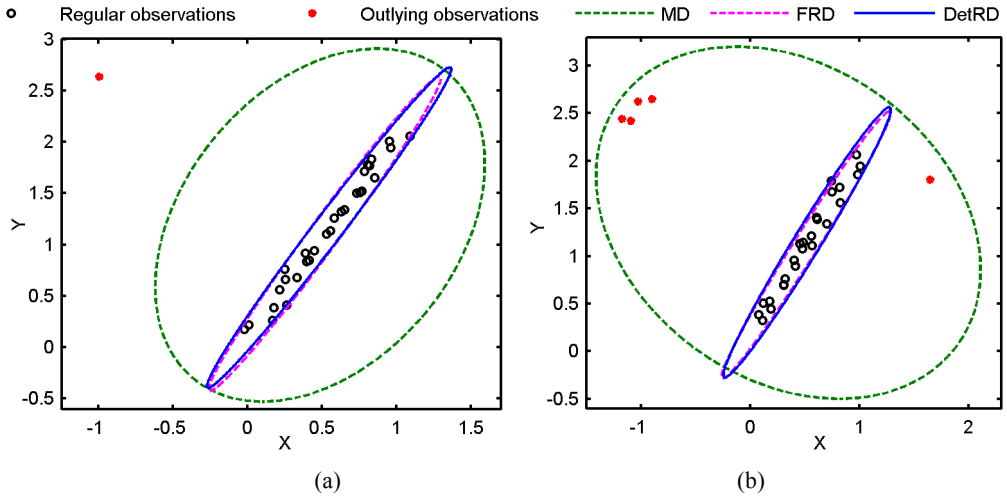


Fig. 4. Outlier (red point) detection by MD, FRD and DetRD; in the presence of (a) single outlier, and (b) multiple and clustered outliers.

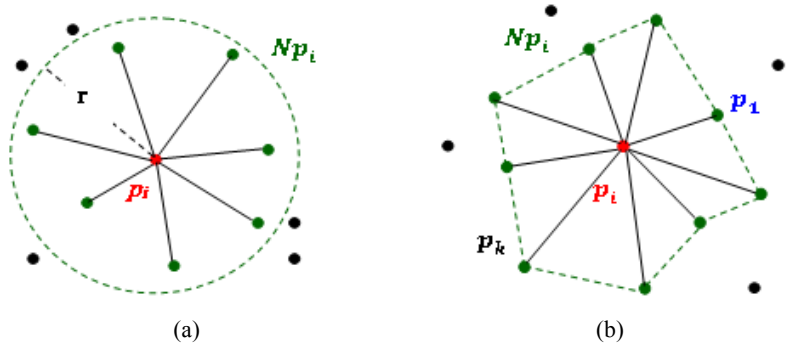


Fig. 5. Local neighbourhood for p_i: (a) fixed distance neighbourhood, and (b) *k*-nearest neighbourhood.

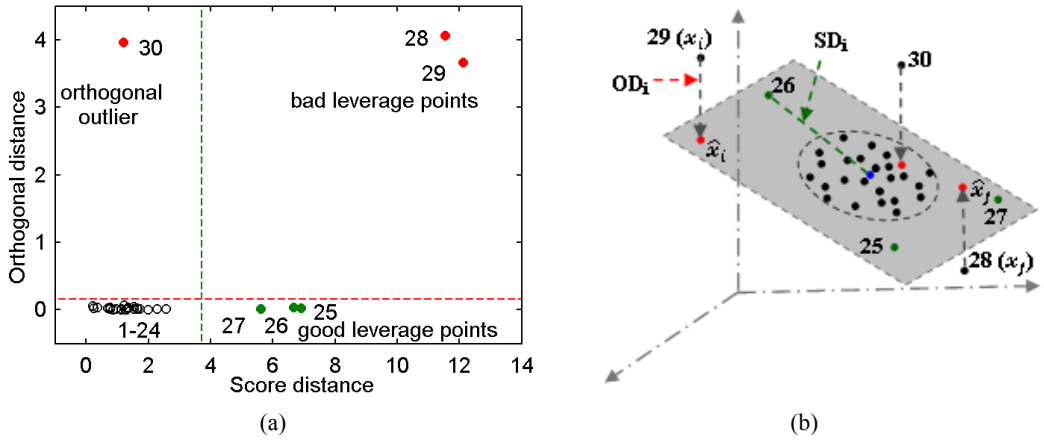


Fig. 6. Outlier detection: (a) diagnostic plot of orthogonal distance versus score distance, and (b) fitted plane. Green points are distant in terms of score and red points are orthogonal outliers.

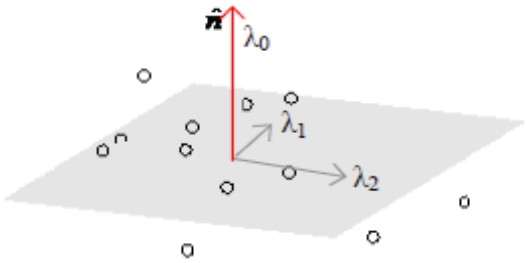


Fig. 7. Point variations along the plane normal ($\hat{n}$) and the first two PCs.

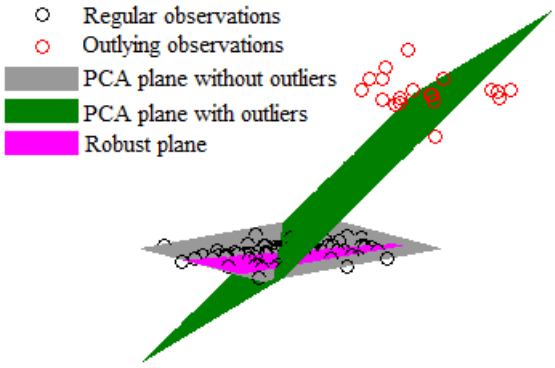
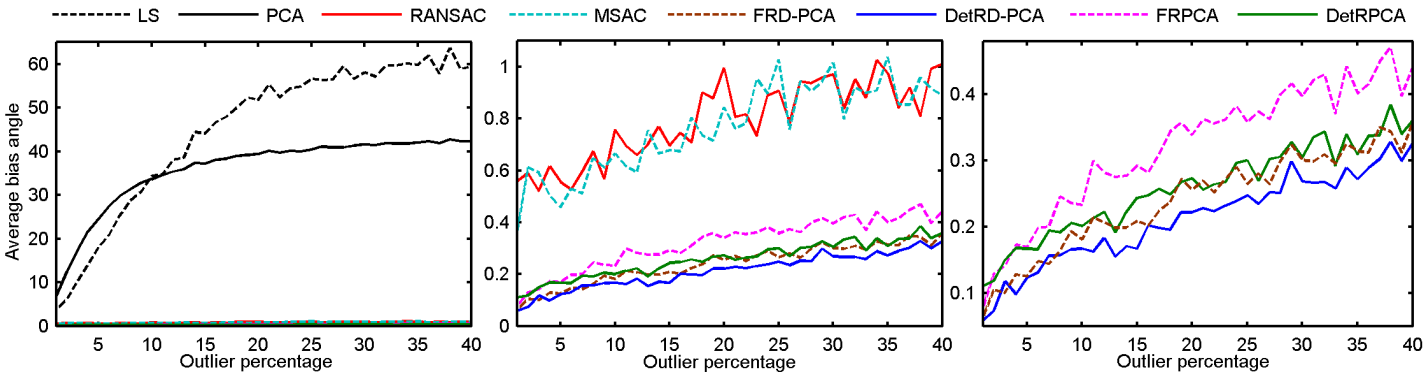
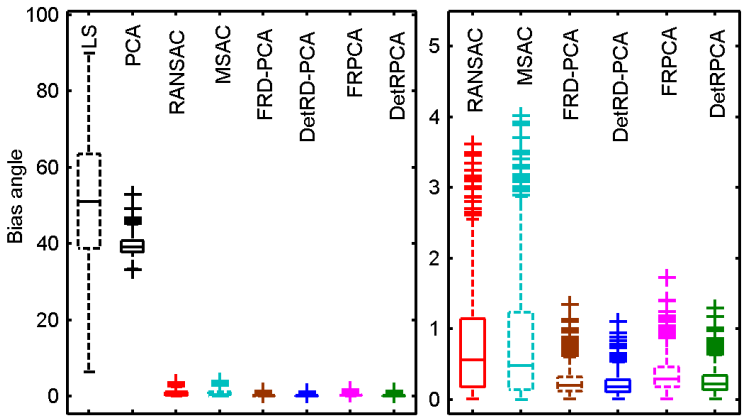


Fig. 8. Artificial dataset of 100 points including 20% outliers; outliers influence on the fitted planes using PCA and a robust method.



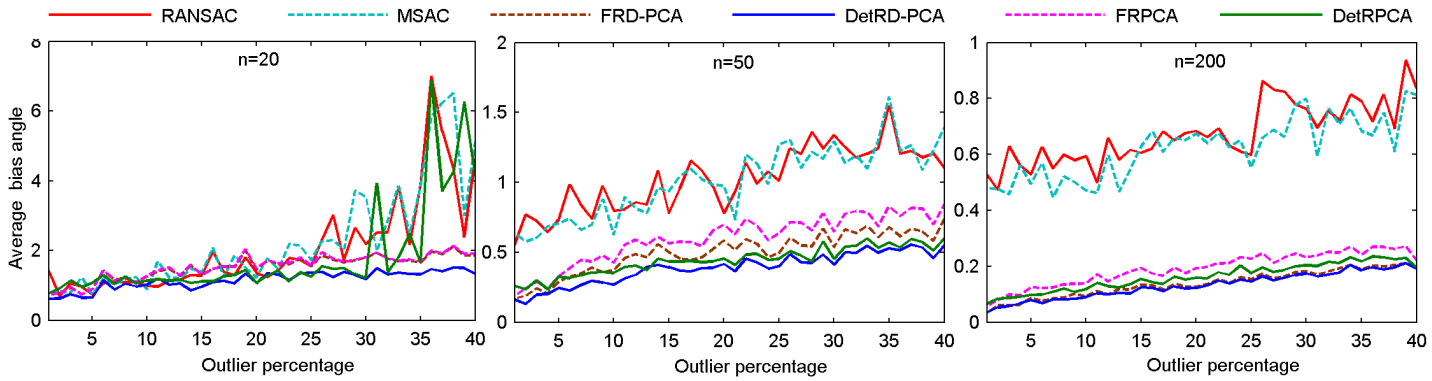
(a) (b) (c)

Fig. 9. Line diagrams for average bias angles versus outlier percentage; n =100, and outlier percentage = 1% – 40% : (a) all methods, (b) all methods except LS and PCA, and (c) robust statistical methods (FRD-PCA, DetRD-PCA, FRPCA and DetRPCA).



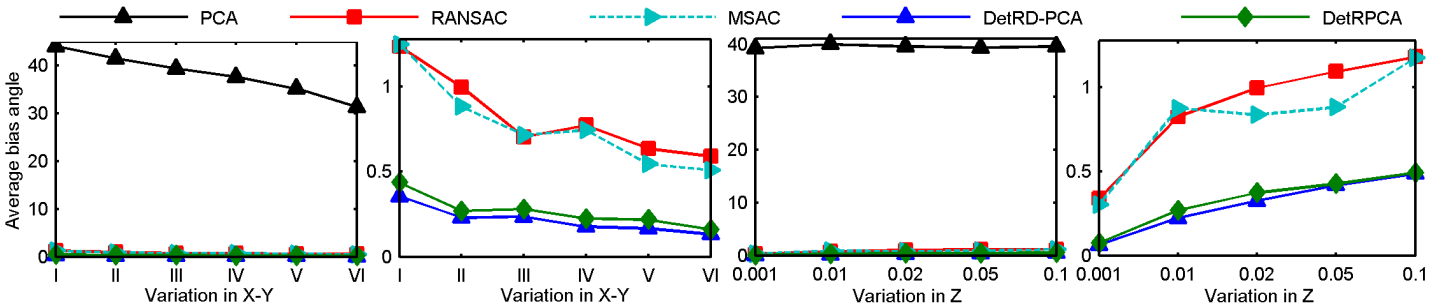
(a) (b)

Fig. 10. The presented boxplots are exploring the robustness of the descriptive statistics (e.g., median and quartile range) graphically; the boxplots are for bias angles from 1000 runs; n =100 and outlier percentage = 20: (a) all methods, and (b) all methods except LS and PCA.



(a) (b) (c)

Fig. 11. Line diagrams for average bias angles versus outlier percentages (1% to 40%): (a) n = 20, (b) n = 50, and (c) n = 200.



(a) (b) (c) (d)

Fig. 12. Average bias angle with respect to point density variation in x-y: (a) all methods, and (b) robust methods; average bias angle with respect to z-variation: (c) all methods, and (d) robust methods.

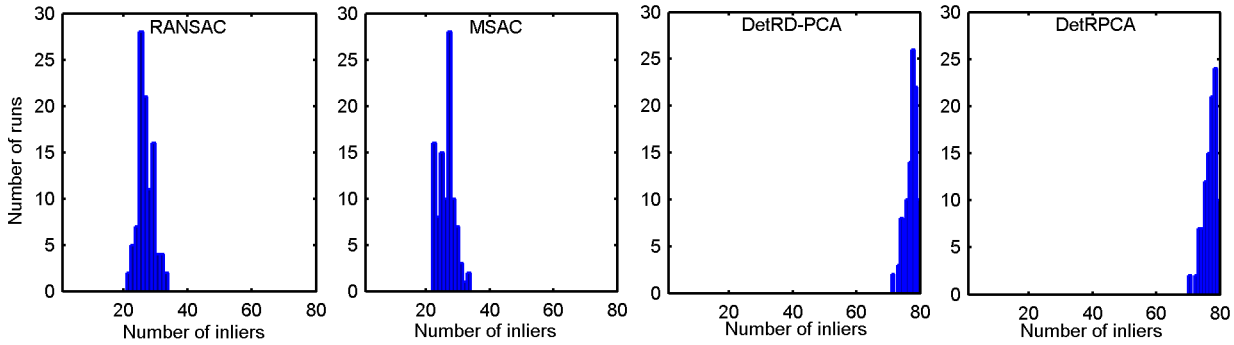


Fig. 13. Histograms for the number of run versus number of correctly identified inliers from different robust methods.

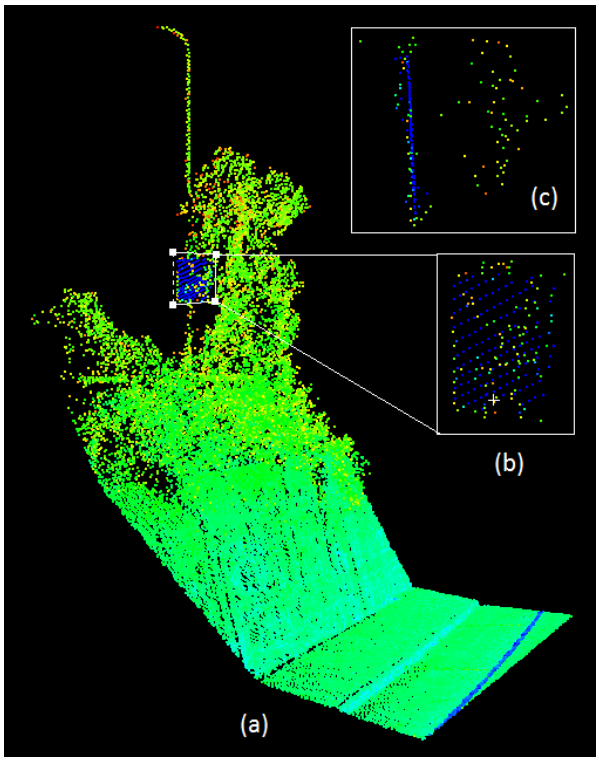


Fig. 14. (a) 'Road scene' data set with a road sign in the zoomed-in windows: (b) front view, and (c) side view.

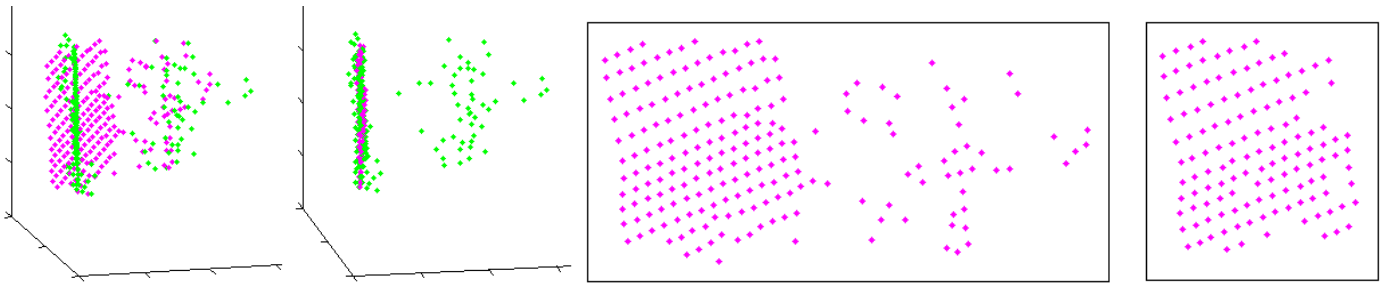


Fig. 15. Fitted plane (magenta) orientation by (a) PCA, and (b) robust method; fitted/extracted plane: (c) PCA, and (d) robust method.

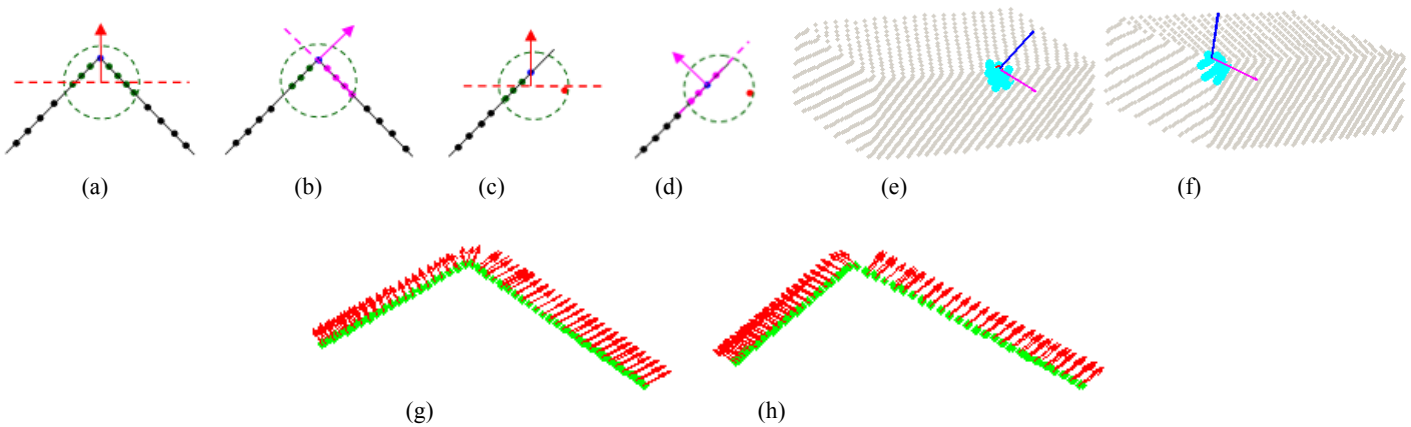


Fig. 16. Neighbouring points in the dashed green circle are from two planar regions: (a) PCA plane (red dotted line) and normal (red arrow), and (b) robust plane (magenta dotted line) and normal (magenta arrow); neighbouring points in a circle include a noise point (red dot): (c) PCA plane (red dotted line) and normal (red arrow) (d) robust plane (magenta dotted line) and normal (magenta arrow); PCA normals are blue and robust normals are magenta, cyan points are the local neighbouring points: (e) normals on an edge point, and (f) normals on a corner point; normals on sharp region: (g) PCA, and (h) robust method.

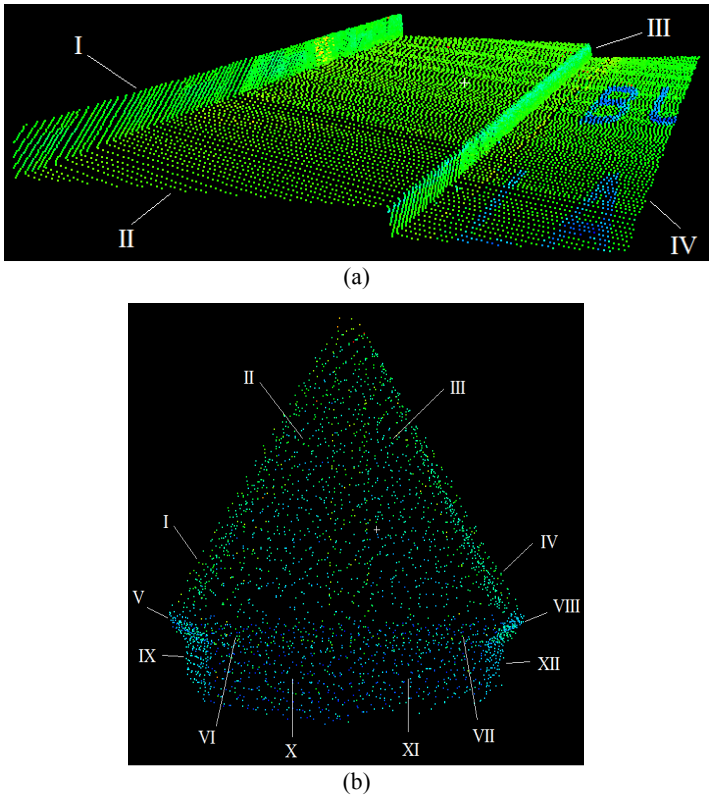
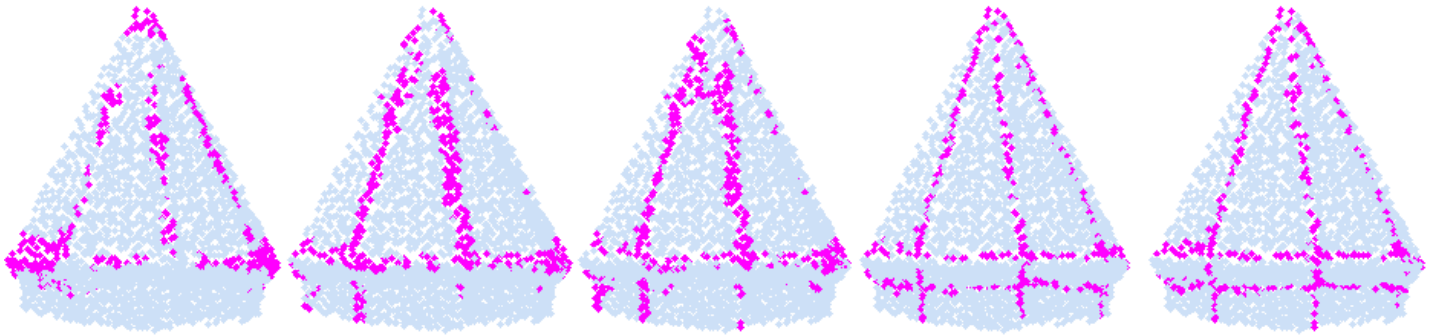


Fig. 17. MLS point cloud data sets: (a) 'road-kerb-footpath' data set, and (b) 'crown' data set.

(a) (b) (c) (d) (e)

Fig. 18. Edge points (in magenta) recovery for 'road-kerb-footpath' data set: (a) PCA, (b) RANSAC, (c) MSAC, (d) DetRD-PCA, and (e) DetRPCA.



(a) (b) (c) (d) (e)

Fig. 19. Edge and corner points (in magenta) recovery for 'crown' data set: (a) PCA, (b) RANSAC, (c) MSAC, (d) DetRD-PCA, and (e) DetRPCA.

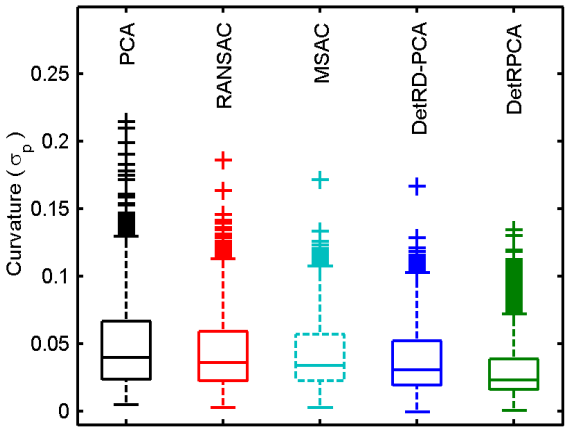
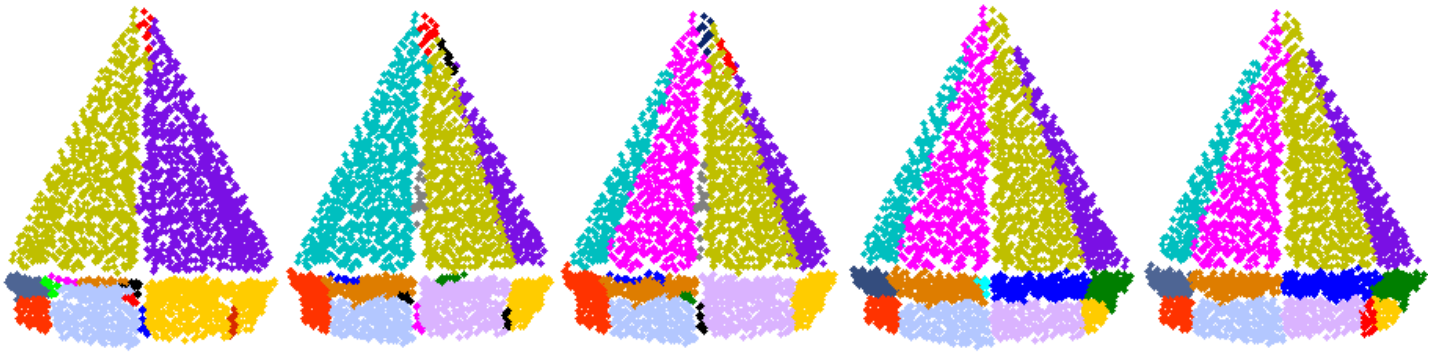
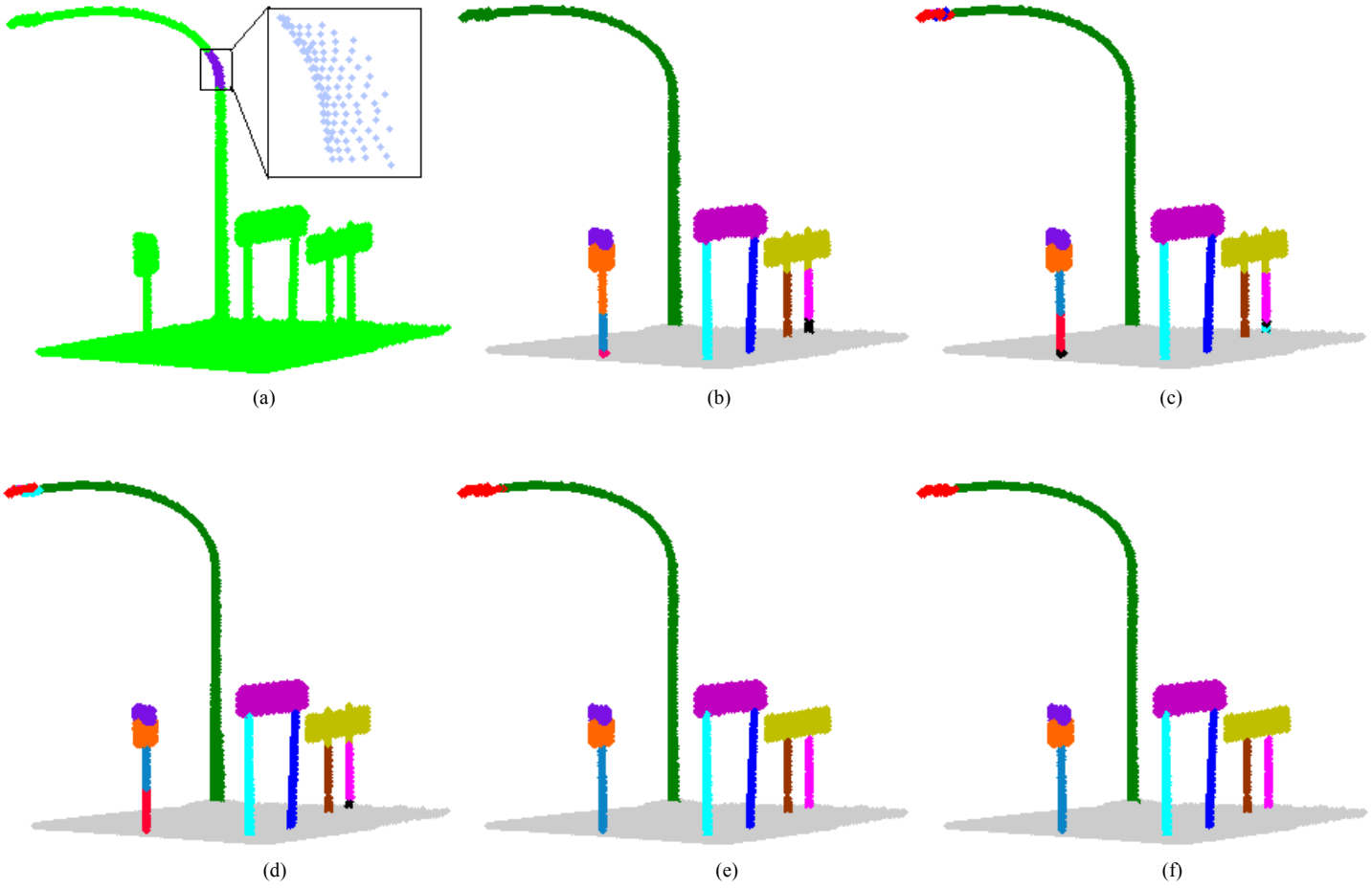


Fig. 20. Boxplots of curvature values for 'crown' data set.



(a) (b) (c) (d) (e)

Fig. 21. Segmentation results for 'crown' data set: (a) PCA, (b) RANSAC, (c) MSAC, (d) DetRD-PCA, and (e) DetRPCA.



(a) (b) (c)

(d) (e) (f)

Fig. 22. (a) 'Traffic furniture' data set, segmentation results: (b) PCA, (c) RANSAC, (d) MSAC, (e) DetRD-PCA, and (f) DetRPCA.

Table 1
Descriptive measures for bias angles (in degrees) from different methods.

Methods	Mean	95% Confidence interval of mean		Minimum	Maximum	Median	Std. Dev.
		Lower bound	Upper bound				
LS	52.497	51.470	53.523	9.220	89.648	52.859	16.540
PCA	39.554	39.409	39.699	33.249	51.881	39.436	2.333
RANSAC	0.799	0.752	0.846	0.006	4.027	0.573	0.758
MSAC	0.798	0.747	0.849	0.002	4.199	0.527	0.819
FRD-PCA	0.206	0.196	0.212	0.003	0.870	0.178	0.128
DetRD-PCA	0.204	0.195	0.212	0.005	1.099	0.175	0.121
FRPCA	0.245	0.236	0.253	0.014	1.033	0.222	0.144
DetRPCA	0.240	0.239	0.259	0.002	1.172	0.218	0.142

Table 2
Statistical significance test.

Methods		Significance (p-value)	Decision
LS	PCA	0.0000	Reject H_o
RANSAC	MSAC	0.4390	Retain H_o
FRD-PCA	DetRD-PCA	0.9210	Retain H_o
FRPCA	DetRPCA	0.6820	Retain H_o
PCA	RANSAC	0.0000	Reject H_o
PCA	MSAC	0.0000	Reject H_o
MSAC	DetRD-PCA	0.0000	Reject H_o
MSAC	DetRPCA	0.0000	Reject H_o
DetRD-PCA	DetRPCA	0.0000	Reject H_o
RANSAC	DetRD-PCA	0.0000	Reject H_o
RANSAC	DetRPCA	0.0000	Reject H_o

Table 3
Variances for regular (R) and outlier (O) data.

Variance	I	II	III	IV	V	VI
x (R,O)	(3,3)	(5,5)	(7,7)	(9,9)	(11,11)	(15,15)
y (R,O)	(3,3)	(5,5)	(7,7)	(9,9)	(11,11)	(15,15)

Table 4
Classification performance.

	RANSAC	MSAC	DetRD-PCA	DetRPCA
TPR (Sensitivity)	100.00	100.00	100.00	100.00
FPR (Swamping)	66.13	66.90	3.35	3.36
Accuracy	47.10	46.48	97.32	97.31

Table 5
Segmentation results for 'crown' and 'traffic furniture' data sets.

Methods	'Crown' data set				'Traffic furniture' data set			
	TS	PS	OS	US	TS	PS	OS	US
PCA	14	2	7	4	13	7	3	2
RANSAC	16	2	8	4	17	9	5	0
MSAC	15	3	6	3	15	9	3	0
DetRD-PCA	13	11	1	0	12	12	0	0
DetRPCA	13	11	1	0	12	12	0	0

TS = Total Segments, PS = Proper Segments, OS = Over Segments, and US = Under Segments.

Table 6
Plane fitting time (in seconds).

Sample size	RANSAC	MSAC	FRD-PCA	DetRD-PCA	FRPCA	DetRPCA
20	0.203	0.209	0.821	0.025	0.820	0.028
50	0.205	0.212	0.815	0.027	0.817	0.031
100	0.214	0.227	0.839	0.033	0.853	0.038
500	0.318	0.342	0.968	0.059	0.968	0.066
1,000	0.436	0.467	1.046	0.136	1.043	0.142
10,000	2.389	2.538	1.146	0.447	1.241	0.575