Trimmed Median PCA for Robust Plane Fitting

Suguru Miyagawa, Shin Yoshizawa, and Hideo Yokota

TOPCON Corp. and RIKEN, Japan

TRIMMED MEDIAN PCA FOR ROBUST PLANE FITTING

Suguru Miyagawa[1,2]

[1]R & D Division, TOPCON Corp.,
75-1 Hasunuma, Itabashi, Tokyo 174-8580, Japan
suguru.miyagawa@riken.jp

Shin Yoshizawa[2] *and Hideo Yokota*[2]

[2]Image Processing Research Team,
RIKEN Center for Advanced Photonics, RIKEN
2-1 Hirosawa, Wako, Saitama 351-0198, Japan
{shin, hyokota}@riken.jp

ABSTRACT

The paper proposes a robust plane fitting method for highly corrupted 3D point sets with significant outliers. Our method integrates a sampling strategy of robust regressions with median approximation of the covariation matrix. We numerically examined our method with popular conventional approaches in terms of speed and accuracy, and achieved higher accuracy for various distributions of outliers.

Index Terms— Geometric Median, Median Covariation Matrix, Plane Fitting, Robust Linear Regression

1. INTRODUCTION

Plane fitting is a fundamental computational tool and has many applications in image processing and computer vision [1, 2] for tasks such as range image segmentation [3] and 3D environment reconstructions [4, 5]. Least-squares (LS) fitting [6, 7] is the most popular method for estimating a plane from a point set; however, the accuracy of the traditional LS using vertical offsets, as shown in Fig. 1, strongly depends on the coordinate system. Alternatively, Principal Component Analysis (PCA) [8] (the eigenvector corresponding to smallest eigenvalue of a covariance matrix provides the estimated plane normal) is based on perpendicular offsets; its accuracy is independent of the coordinate system. Consequently, PCA has been widely used in range image analysis [9].

Unfortunately, the sensitivity of LS and PCA to outliers is well known. Hence statistical approaches such as robust regressions [10] with Random Sample Consensus (RANSAC) [11] have been well studied for the detection and removal of outliers [12, 13, 14]. Further, several extensions of RANSAC such as M-estimator Sample Consensus (MSAC) [15], which is based on maximum likelihood estimation, and Progressive Sample Consensus (PROSAC) [13], which utilizes similarity measurement, have been developed. The RANSAC family is a powerful paradigm for robust regression. It provides reliable estimation for highly corrupted data with significant outlier population [16]. In general, the residual range of inlying data points in the dataset must be known; however, in real-world data, such data points are unknown [11, 16].

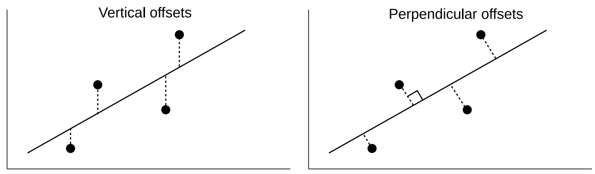


Fig. 1. Residuals of vertical (left) and perpendicular (right) offsets. The methods that utilize vertical offsets (parallel to an axis of the coordinate system) include LS [6, 7] and LTS [10], while a formulation of using perpendicular offsets leads PCA-based methods [9] such as MCMPCA [14].

Other popular robust regression approaches include Least Median of Squares (LMedS) [17] and Least Trimmed Squares (LTS) [10]; in these approaches, the *breakdown point*[1] is 0.5. That is, these methods can handle data containing up to 50% outliers with high accuracy. On the other hand, LS has a breakdown point of 0.0. LMedS is a generalized median minimizing medians of squared residuals. LTS is a sampling-based approach that has several advantages over LMedS, with the exception of computational speed [18]. To address this issue, FAST-LTS [19] was proposed, where a subset of input data was used for accelerating the computational speed of LTS. For a d-dimensional point $\mathbf{x} \in \mathbb{R}^d$ and its associated scalar $y \in \mathbb{R}$, consider an explicit representation of $(d+1)$-dimensional hyperplane defined by $y = \langle \theta, \mathbf{x} \rangle + b$ where $\theta \in \mathbb{R}^d$ is an uniformed parameter vector, $b \in \mathbb{R}$ is a constant, and $\langle \cdot, \cdot \rangle$ is the inner product. Assume that the squared residuals in the ascending order by $|r^2(\theta)|_{(1)} \leq |r^2(\theta)|_{(2)} \leq \cdots \leq |r^2(\theta)|_{(n)}$ for n samples $(y_i, \mathbf{x}_i) \in \mathbb{R}^{d+1}$, $i \in \{1, 2, ..., n\}$ where the residual is given by $r(\theta)_{(i)} = y_i - \langle \mathbf{x}_i, \theta \rangle$. Then, θ using LTS is given by the following Eq. (1):

$$\theta = \arg\min_{\theta} \sum_{i=1}^{h} |r^2(\theta)|_{(i)}, \quad h < n. \tag{1}$$

Although LTS-based methods incorporated a smart data trimming approach as described in Eq. (1) into regression analy-

[1]A robustness criteria in regression analysis [10] that measures the possible ratio of handling outliers in data without wrong estimations (higher values are preferred; 0.5 is maximum).

sis, they are sensitive to both its coordinate system and highly corrupted outliers because they employed LS fitting in their methods.

More recently, a generalization of median to a covariance matrix (MCM: median covariance matrix) was introduced in [20] and its PCA (MCMPCA) robustly estimates a plane from a point set corrupted with outliers. MCM is based on the so-called geometric median $\overline{\mathbf{m}}$ of a point set $\{\mathbf{x}_i \in \mathbb{R}^d\}$:

$$\overline{\mathbf{m}} = \arg\min_{\mathbf{u}\in\mathbb{R}^d} \mathbb{E}\left[||\mathbf{x}_i - \mathbf{u}|| - ||\mathbf{x}_i||\right], \tag{2}$$

where $||\cdot||$ and $\mathbb{E}[\,\cdot\,]$ are the associated norm and the expectation, respectively. In geometry, Eq. (2) has been well studied as a Fermat-Torricelli point of a triangle and also known as Weber's problem [21] minimizing transportation costs in location theory. Similar to the geometric median Eq. (2), MCM $\Gamma \in \mathbb{R}^{d\times d}$ is defined by

$$\Gamma = \arg\min_{V\in\mathbb{R}^{d\times d}} \mathbb{E}\left[||M - V||_F - ||M||_F\right],$$
$$M = (\mathbf{x}_i - \overline{\mathbf{m}})(\mathbf{x}_i - \overline{\mathbf{m}})^T \in \mathbb{R}^{d\times d},$$

where $||\cdot||_F$ is the Frobenius norm and $||M||_F$ (also $||\mathbf{x}_i||$ in Eq. (2)) is a regularization term [14]. Several nonlinear optimization techniques [22, 23] such as line search and stochastic gradient descent methods to numerically solve the geometric median Eq. (2) have been proposed. However, the absence of sophisticated sampling strategy in conventional methods makes difficult to take full advantage of the geometric median (breakdown point of 0.5) for plane estimations.

In this paper, we propose a novel and robust plane fitting method based on integrating a sampling strategy of robust regressions into the MCMPCA concept. Our method (Trimmed Median PCA) employs the weighted MCMPCA with a stochastic gradient descent method and the sampling framework of FAST-LTS in order to obtain both of their advantages. Our technical contributions also include the weighting and shuffling schemes to improve a conventional numerical solver of MCMPCA. We examined our method for various randomly generated 3D point sets including outliers in terms of fitting accuracy and computational speed, and compared with those for popular conventional methods.

2. TRIMMED MEDIAN PCA

Consider d-dimensional point set $\{\mathbf{x}_i \in \mathbb{R}^d\}$, $i \in \{1, 2, ..., n\}$. First, the set is regularized using Median Absolute Deviation (MAD) [24, 25]. Then, the sampling framework of FAST-LTS is applied as described in Sec. 2.1 to obtain the final best fitting plane $\mathcal{P}$; however, instead of the classical LS fitting used in the original FAST-LTS and LTS, the weighted MCMPCA is used for estimating the planes of the sampled subset. In addition, the residuals used in Sec. 2.1 was calculated by perpendicular offsets from estimated planes instead of the vertical offsets in order to avoid coordinate system dependency. Fig. 2 illustrates the flowchart of our method. The remainder of this section describes its components.

MAD regularization: To avoid the size influence of data distribution, $\{\mathbf{x}_i\}$ is regularized with MAD [24, 25] which is defined by

$$\mathrm{MAD} = \mathrm{median}[||\mathbf{x}_1 - \mathbf{m}_e||, ..., ||\mathbf{x}_n - \mathbf{m}_e||],$$

where $\mathbf{m}_e \in \mathbb{R}^d$ is an element-wise median of $\{\mathbf{x}_i\}$. Each point is rescaled by the MAD as $\mathbf{x}_i \leftarrow (\mathbf{x}_i - \mathbf{m}_e)/\mathrm{MAD}$.

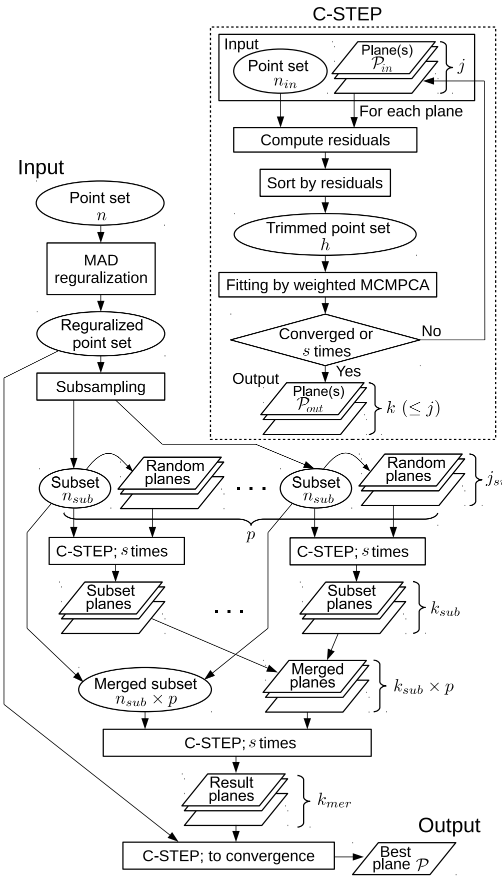


Fig. 2. Flowchart of our trimmed median PCA.

2.1. Sampling Strategy

The basic structure of our sampling strategy follows the framework of FAST-LTS; first, p subsets are chosen from $\{\mathbf{x}_i\}$, where each subset comprises randomly sampled n_{sub} points. Then, each subset is subjected to a robust plane estimation procedure, referred to as C-STEP [19]. C-STEP inputs n_{in} points and j planes (usually calculated form this n_{in} points), and generates $k\,(\leq\, j)$ planes as the output, as described in Fig. 2. The final plane $\mathcal{P}$ is also estimated from the MAD regularized points and the merged subset planes by using C-STEP.

Sampling and merging subsets: The Fisher-Yates method [26] which is an efficient data shuffle algorithm is applied to $\{\mathbf{x}_i\}$, and subsequently p subsets are sampled by classifying successive n_{sub} points as a subset. Merging in Fig. 2 is simply a collection of the data (point and plane subsets).

C-STEP: The C-STEP in FAST-LTS is a basically original LTS estimating the planes by using the residuals and trimming data set described in Eq. (1). Each output plane $\mathcal{P}_{out}$ is estimated using the input point set, and each input plane $\mathcal{P}_{in}$ (the case of $k = j$) or small number of $\mathcal{P}_{out}$ are selected according to their residuals (the case of $k < j$). In contrast to FAST-LTS, the weighted MCMPCA and perpendicular offsets are employed in our method for fitting planes and residual calculation, respectively.

2.2. Weighted MCMPCA

Our weighted MCM is estimated using the stochastic gradient descent method based on [14] with the weighting and shuffling schemes. First, the geometric median $\overline{\mathbf{m}}$ is obtained by iteratively applying the following procedure, where one iteration consists of Eqs. (3) and (4) for all $i \in \{1, 2, ..., n\}$.

$$\mathbf{m}_{i+1} = \mathbf{m}_i + w_i \gamma_i^{(m)} \frac{\mathbf{x}_i - \mathbf{m}_i}{||\mathbf{x}_i - \mathbf{m}_i||}, \quad (3)$$

$$\overline{\mathbf{m}}_{i+1} = \overline{\mathbf{m}}_i - \frac{1}{\xi + i + 1}(\overline{\mathbf{m}}_i - \mathbf{m}_{i+1}), \quad (4)$$

where $\overline{\mathbf{m}}_1 = \mathbf{m}_1 = \mathbf{m}_e$, $\xi \in \{0, 1, 2...\} \times n$ is an iterator number incremented at each iteration, $w_i \in \mathbb{R}$ is a weight, $\gamma_i^{(m)} = c_m/(\xi + i)^\alpha$ is a coefficient of gradient descent process, and c_m and α are the user-specified parameters. For each iteration, $\{\mathbf{x}_i\}$ is shuffled by Fisher-Yates method and reassigned $i \to 1$, $\mathbf{m}_{n+1} \to \mathbf{m}_1$, and $\overline{\mathbf{m}}_{n+1} \to \overline{\mathbf{m}}_1$. The iteration is converged numerically when $||\overline{\mathbf{m}}_1 - \overline{\mathbf{m}}_{n+1}|| < \varepsilon$. Here the weight w_i is also updated by each iteration such that the following equation with $\overline{\mathbf{m}}_1 \to \overline{\mathbf{m}}$ is employed.

$$w_i \equiv \begin{cases} \left(1 - \left(\frac{||\mathbf{x}_i - \overline{\mathbf{m}}||}{c}\right)^2\right)^2 & \text{if} \quad \frac{||\mathbf{x}_i - \overline{\mathbf{m}}||}{c} < 1, \\ 0 & \text{Otherwise,} \end{cases}$$

where c is a user-specified parameter. This weighting function is inspired by the so-called τ scale weight [27] which increases importance of the point $\mathbf{x}_i$ if it is close to $\overline{\mathbf{m}}$ during our gradient descent process.

Next, the weighted MCM Γ is also estimated by the following iterative process where one iteration consists of Eqs. (5) and (6) for all $i \in \{1, 2, ..., n\}$.

$$\Gamma_{i+1} = \Gamma_i + w_i \gamma_i \frac{(\mathbf{x}_i - \overline{\mathbf{m}})(\mathbf{x}_i - \overline{\mathbf{m}})^T - \Gamma_i}{||(\mathbf{x}_i - \overline{\mathbf{m}})(\mathbf{x}_i - \overline{\mathbf{m}})^T - \Gamma_i||_F}, \quad (5)$$

$$\overline{\Gamma}_{i+1} = \overline{\Gamma}_i - \frac{1}{\xi + i + 1}\left(\overline{\Gamma}_i - \Gamma_{i+1}\right), \quad (6)$$

where $\overline{\Gamma}_1 = \Gamma_1 = (\mathbf{x}_1 - \overline{\mathbf{m}})(\mathbf{x}_1 - \overline{\mathbf{m}})^T$, $\gamma_i = c_\gamma/(\xi + i)^\alpha$ is a coefficient of gradient descent process, c_γ is a user-specified parameter, and $\overline{\mathbf{m}}$ is the geometric median estimated by Eqs. (3) and (4). For each iteration, $\{\mathbf{x}_i\}$ is shuffled by Fisher-Yates method and reassigned $i \to 1$, $\Gamma_{n+1} \to \Gamma_1$, and $\overline{\Gamma}_{n+1} \to \overline{\Gamma}_1$ in our method, where the method of [14] only one iteration is performed in both $\overline{\mathbf{m}}$ and $\overline{\Gamma}$ without iterative random sampling such as shuffling. The iteration is converged numerically when $||\overline{\Gamma}_1 - \overline{\Gamma}_{n+1}||_F < \varepsilon$, and then $\overline{\Gamma}_{n+1} \to \Gamma$. The resulting plane is defined by the plane passing through $\overline{\mathbf{m}}$ whose unit normal vector is given by PCA of Γ.

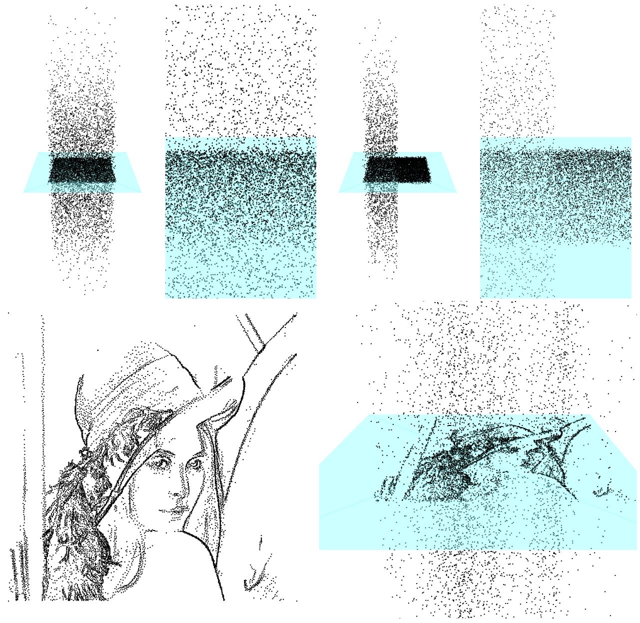
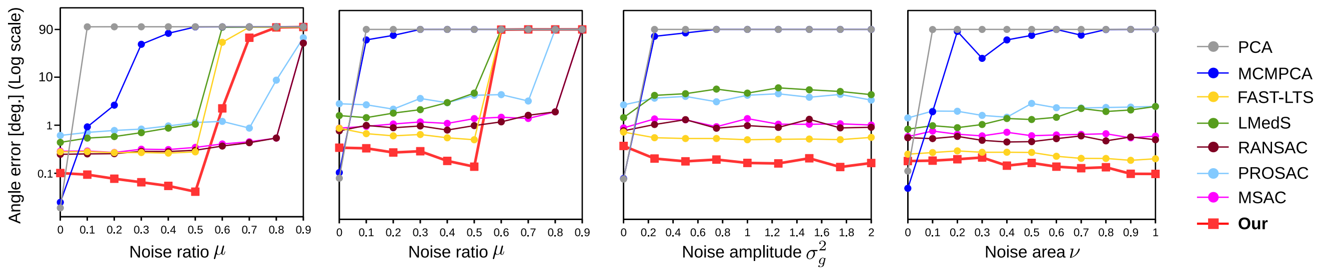


Fig. 3. Input point set examples. Uniform (top) and non-uniform (bottom) point sets corrupted with both additive noise and outliers are used in our numerical experiments.

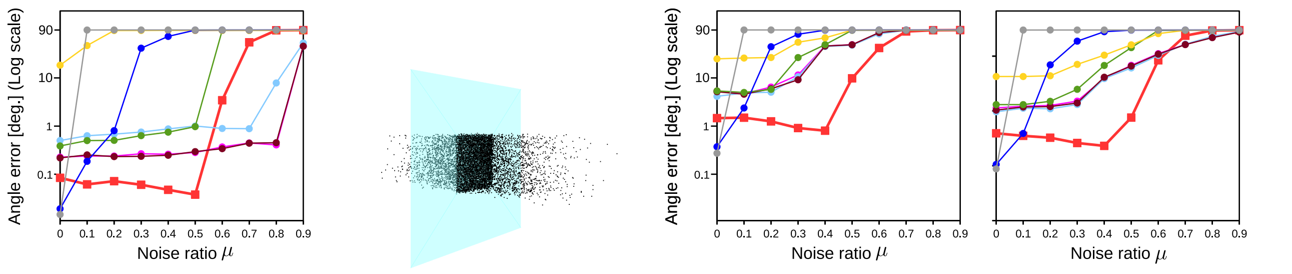
3. COMPARISON AND EVALUATION

All numerical experiments in this paper were performed on a Core-i7 5960X (3.0 GHz 8 core, no parallelization was used) PC with 128 GB RAM and 64-bit OS.

Methods and parameters: We compared our method with the popular conventional methods: PCA [8], LMedS [17], RANSAC [11], PROSAC [13], and MSAC [15] by using the point cloud library (PCL) [28], and MCMPCA [14] and FAST-LTS [19] implemented based on the R package Gmedian [29] and MATLAB library LIBRA [30], respectively. We employed the recommended parameters of PCL, Gmedian, LIBRA, and MCMPCA [14] in these conventional methods. We also employed the recommended parameters on MCMPCA [14] and FAST-LTS [19] for our method such that $n_{sub} = 300$, $j_{sub} = 100$, $k_{sub} = k_{mer} = 10$, $s = 2$,



(a): Ave. (b): Max. (c): Max. (d): Max.



(e): Ave. (f): an input of (e). (g): Ave. $\sigma_a^2 = 0.1$. (h): Ave. $\sigma_a^2 \in [0.01, 0.1]$.

Fig. 4. Accuracy comparisons for varying noise ratio μ (a,b,e,g,h), amplitude σ_g (c), and area ν (d). Average (a,e,g,h) and maximum (b,c,d) angle errors are examined where (a-e): $\sigma_a^2 = 0.01$, $\sigma_g^2 = 1.0$ and $\nu = 1.0$ excepting (c), and (c): $\nu \in [0, 1]$.

PCA [8]	MCMPCA [14]	FAST-LTS [19]	LMedS [17]	RANSAC [11]	PROSAC [13]	MSAC [15]	Our
1.24×10^8	4.30×10^6	6.77×10^4	2.86×10^5	3.52×10^4	1.43×10^5	3.50×10^4	1.64×10^4

Fig. 5. Average computational speed for all experiments via a number of points per second.

$p = 5$, and $h = (n + d + 1)/2$ in Sec. 2.1 as well as $c = 3$, $c_\gamma = c_m = 2$, $\alpha = 0.75$, and $\varepsilon = 0.01$ in Sec. 2.2 where j_{sub}, k_{sub}, and k_{mer} are the numbers of initial, estimated, and resulting subset planes described in Fig. 2, respectively.

Input point sets and their noises: For the inputs in our numerical experiments, we employed the 3D point sets uniformly and non-uniformly distributed on xy-plane and their rotation as shown in Figs. 3 and 4 (f), respectively. Each point set consists of 10^4 points (we omitted the results obtained using other point numbers during the evaluation of accuracy because it only affects the computational speed). We employed white Gaussian distribution with standard deviations σ_a and σ_g for additive noise (element-wise perturbation) and outliers (the direction of unit normal of xy-plane), respectively. The magnitude of noise is proportional to the diagonal length of the point set region in the xy-plane. The outliers were randomly selected with probability μ from ν area ratio of $\{\mathbf{x}_i\}$. The images in the top of Fig. 3 present the examples of full ($\nu = 1.0$, left) and half ($\nu = 0.5$, right) area ratio of outliers. For every parameter setting combination of σ_g^2, σ_a^2, ν, and μ, we examined 10 randomly sampled point sets.

Accuracy and timings: In Figs. 4 and 5, the accuracy and computational time of our method are compared with that of conventional methods via the angle error between ideal and estimated plane normals. Our method achieved the highest accuracy, where $\mu \in [0.1, 0.5]$ as shown in Fig. 4 (a), (b), and (e). Also, varying the coordinate system, noise amplitude σ_g, and area of corruption ν as in Fig. 4 (e), (f), (c), and (d) do not change the accuracy ranking results. It can be clearly seen that FAST-LTS is sensitive to rotation in Fig. 4 (e) and (f). For $0.5 < \mu$, the RANSAC family is better when its range of inlying data points is small ($\sigma_a^2 = 0.01$). On the other hand, if the range is much larger than the parameter of RANSAC family, then our method obtains better estimation for $0.5 < \mu$, as shown in Fig. 4 (g) and (h). Although our method is computationally slower than the conventional methods, the order of our computational speed is similar to that of FAST-LTS, RANSAC, and MSAC as indicated in Fig. 5.

4. CONCLUSION

In this study, we have proposed a novel and robust plane fitting method for 3D point sets corrupted with outliers. Our method integrates the advantages of FAST-LTS and MCMPCA, and achieved high accuracy for various outlier distributions. In addition, our method is not sensitive to the coordinate system, amplitude of outliers, inlier point range, and data scale. We believe that our method is applicable to other linear regression problems in image processing; range image segmentation is a promising future work.

5. REFERENCES

[1] H. Richard and A. Zisserman, *Multiple view geometry in computer vision*, Cambridge University Press, 2003.

[2] S. Milan, H. Vaclav, and B. Roger, *Image processing, analysis, and machine vision*, Cengage Learning, 2014.

[3] J. Deschaud and F. Goulette, "A fast and accurate plane detection algorithm for large noisy point clouds using filtered normals and voxel growing," in *3DPVT*, 2010.

[4] J. Huang, A. Dai, L. Guibas, and M. Nießner, "3DLite: Towards commodity 3D scanning for content creation," *ACM TOG*, vol. 36, no. 6, pp. 203, 2017.

[5] A. Ulusoy, M. Black, and A. Geiger, "Semantic multi-view stereo: Jointy estimating objects and voxels," in *IEEE CVPR*, 2017, pp. 2414–2423.

[6] C. Gauss, "Theoria motus corporum coelestium in sectionibus conicis solem ambientium," *Perthes and Besser*, 1809.

[7] A. Legendre, "Nouvelle méthodes pour la détermination des orbites des comètes," *Courcier*, 1805.

[8] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philos. Mag.*, vol. 2, no. 11, pp. 559–572, 1901.

[9] H. Badino, D. Huber, Y. Park, and T. Kanade, "Fast and accurate computation of surface normals from range images," in *IEEE ICRA*, 2011, pp. 3084–3091.

[10] P. Rousseeuw, *Robust Regression and Outlier Detection*, Wiley - Interscience, 2003.

[11] M. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[12] R. Wilcox, *Introduction to Robust Estimation and Hypothesis Testing*, Academic Press, 2011.

[13] O. Chum and J. Matas, "Matching with PROSAC - progressive sample consensus," in *IEEE CVPR*, 2005, vol. 1, pp. 220–226.

[14] H. Cardot and A. Godichon-Baggioni, "Fast estimation of the median covariation matrix with application to online robust principal components analysis," *TEST*, vol. 26, pp. 461–480, 2017.

[15] P. Torr and A. Zisserman, "MLESAC: A new robust estimator with application to estimating image geometry," *Comput. Vis. Image. Underst.*, vol. 78, no. 1, pp. 138–156, 2000.

[16] R. Raguram, O. Chum, M. Pollefeys, J. Matas, and J. Frahm, "USAC: A universal framework for random sample consensus," *IEEE TPAMI*, vol. 35, no. 8, pp. 2022–2038, 2013.

[17] P. Rousseeuw, "Least median of squares regression," *J. Am. Stat. Assoc.*, vol. 79, no. 388, pp. 871–880, 1984.

[18] L. Li, "An algorithm for computing exact least-trimmed squares estimate of simple linear regression with constraints," *Comput. Stat. Data Anal.*, vol. 48, no. 4, pp. 717–734, 2005.

[19] P. Rousseeuw and K. Van Driessen, "Computing LTS regression for large data sets," *Data Min. Knowl. Discov.*, vol. 12, no. 1, pp. 29–45, 2006.

[20] H. Cardot, P. Cénac, and P. Zitt, "Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm," *Bernoulli*, vol. 19, no. 1, pp. 18–43, 2013.

[21] G. Wesolowsky, "The Weber problem: History and perspective," *Location Science*, vol. 1, no. 1, pp. 5–23, 1993.

[22] H. Fritz, P. Filzmoser, and C. Croux, "A comparison of algorithms for the multivariate L_1-median," *Comput. Stat.*, vol. 27, no. 3, pp. 393–410, 2012.

[23] M. Cohen, Y. Lee, G. Miller, J. Pachocki, and A. Sidford, "Geometric median in nearly linear time," in *ACM STOC*, 2016, pp. 9–21.

[24] U. Gather and T. Hilker, "A note on Tyler's modification of the MAD for the Stahel-Donoho estimator," *Ann. Stat.*, vol. 25, no. 5, pp. 2024–2026, 1997.

[25] R. Maronna and R. Zamar, "Robust estimates of location and dispersion for high-dimensional datasets," *Technometrics*, vol. 44, no. 4, pp. 307–317, 2002.

[26] R. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, Oliver & Boyd, 1938.

[27] V. Yohai and R. Zamar, "High breakdown-point estimates regression by means of the minimization of an efficient scale," *J. Am. Stat. Assoc.*, vol. 83, no. 402, pp. 406–413, 1988.

[28] R. Rusu and S. Cousins, "3D is here: Point cloud library (PCL)," in *IEEE ICRA*, 2011, pp. 1–4.

[29] H. Cardot, "Package 'Gmedian'," *The R Foundation*, 2017.

[30] S. Verboven and M. Hubert, "Matlab library LIBRA," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 509–515, 2010.