Robust metrological characterization of circular profiles

L Greco\textsuperscript{1}, J Caja\textsuperscript{2} and P Maresca\textsuperscript{2*}

\textsuperscript{1} University Giustino Fortunato, Viale Raffaele Delcogliano, 12, 82100 Benevento, Italy
\textsuperscript{2} Universidad Politécnica de Madrid, Ronda de Valencia, 3, Madrid 28012, Spain

*Corresponding author: piera.maresca@upm.es

Abstract: Outliers are likely to corrupt data acquired by optical profiles and ignoring them can badly affect the output of the analysis. The occurrence of very few outliers can completely invalidate the fitting process and uncertainty evaluations. Outliers are unexpected values that often mask each other being not detectable by any pre-screening of the data. Therefore, there is the need to adopt a proper robust approach to fit circular profiles data and identify data inadequacies. Here, it is suggested to resort to impartial trimming to fit the centre and the radius of a circle. The optimization problem is based on a subset of the data after discarding that fraction of points with the largest distance from the fitted centre. Trimmed observation are likely to be outliers. The methodology is illustrated through both synthetic and real data.

Keywords: Circular profiles, Geometric distance, Optical profiles, Non-linear least squares, Trimming.

1. Introduction

The use of digital optical measuring systems is common place for carrying out dimensional measurements in micro- and nano-manufacturing, in particular for circular features, such as small holes, rounded notches, fillet radii. In order to characterize and classify circular profiles, a common requirement consists in the evaluation of the best fitting circumference to a set of data points, obtained using digital optical machines, along with its diameter and centre \cite{1}. The points correspond to the coordinates of the pixels detected on the edge of the circular profile. The resulting set of measured coordinates may contain several wrong measurements that can be generated from many different sources during the data gathering process: bad behaviour of the instrument, wrong measurements by the researcher, unusual conditions (dust, for instance) that are not properly taken into account \cite{2}. Here, such data inadequacies are denoted as outliers. Outliers corrupt data acquired by optical profiles and ignoring them can badly affect the output of the analysis. Outliers are unexpected values that do not share the pattern followed by the majority of the data and may not have any pattern at all. They can be isolated or clustered. The presence of very few outliers can completely invalidate the fitting process and uncertainty evaluations. One ideal approach towards outliers would be to first detect and then remove them. Actually, such a strategy has been pursued in \cite{1}, in which the authors developed a filtering process in order to eliminate pixels with an appearance probability lower than a specific valued depending on an uncertainty measure stemming from a classical fitting algorithm based on all the data. Unfortunately, quite often a pre-processing of the available data aimed to identify anomalous points is not successful. For instance, outliers could inflate the uncertainty around the fitted radius and the filtering process briefly described above could mask outliers, that, then, can survive the pre-processing step. Furthermore, a correct filtering process should be based on appropriate robust statistical tools able
to cope with outliers. In this paper, it is suggested to resort to proper robust methods, well designed to provide reliable estimates and uncertainties in the presence of outliers but also to lead to results not significantly different from those stemming from classical methods when data are not corrupted. The use of such techniques would avoid any self-developed filtering process. Furthermore, another fundamental feature of any robust procedure is that of providing effective tools for outlier detection based on the robust fit. The proposed methodology is based on the idea of impartial trimming: estimation and elimination of anomalous pixels is performed simultaneously. The technique shares same aspects with random sample consensus (RANSAC [3]). However, there are substantial differences that will be clarified throughout the paper. The paper is structured as follows. Some background on the problem of fitting a circle is first given in Section 2. The robust methodology is illustrated in Section 3, also with some synthetic examples and a small numerical study. Real data applications are discussed in Section 4.

2. Circular features identification by non-linear least squares

The metrological characterization of circular features can be performed according to several techniques. The reader is pointed to [1] for a complete review. Let \( x_1, x_2, \ldots, x_n \) be the set of sample coordinates, with \( x_i = (x_{i1}, x_{i2}) \). Let \( c_0 = (x_{01}, x_{02}) \) denote the centre and \( r_0 \) the radius of the ideal circle to whom the observed points belong. Here, the estimation of the circle’s parameters \( \theta = (c_0, r_0) \) from a set of noisy measurements according to non-linear least squares [4], is the focused. The mathematical definition of the problem states that the best fitting circle, that most closely approximates the real one, is determined by \( \hat{\theta} = (\hat{c}_0, \hat{r}_0) \) that minimizes the objective function:

\[
\sum_{i=1}^{n} \left[ d_i - r_0 \right]^2
\]

with \( d_i = d(x_i, x_{01}, x_{02}) = \left( (x_{i1} - x_{01})^2 + (x_{i2} - x_{02})^2 \right)^{1/2} = \|x_i - c_0\|_2 \)

In equation (1), the single terms are geometric distances. This non-linear optimization problem also coincides with maximum likelihood estimation under the assumption of Gaussian independent errors with null mean and variance \( \sigma^2 \) on both coordinates [5]. This assumption carries the need to estimate the extra parameter \( \sigma \) that defines the roundness error. Moreover, the uncertainty around the estimates is driven from the Hessian matrix. Let \( H(\hat{\theta}) \) denote the Hessian matrix evaluated at the maximum likelihood estimate. Then, the standard errors corresponding to parameters estimates are given by the square root of the diagonal elements of the inverse of the Hessian matrix, i.e.

\[ se(\hat{\theta}_j) = \sqrt{H(\hat{\theta})^{-1}} \], \( j = 1, 2, 3 \). Non-linear least squares fitting is optimal under a likelihood perspective, in terms of uncertainty evaluation. For what concerns the task of model fitting, optimization can be carried out according to standard techniques, such as simplex [6] or quasi-Newton methods [7].

In this paper, non-linear least squares is suitably initialized with \( \theta_0 = (\overline{x}_{10}, \overline{x}_{20}, \max(x_{i1} - \overline{x}_{10}, x_{i2} - \overline{x}_{20}) \).

The maximum likelihood estimate of \( \sigma^2 \) is \( \hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} \left[ \tilde{d}_i - \hat{r} \right]^2 \) with \( \tilde{d}_i = \|x_i - \hat{c}\| \).

3. Impartial trimming estimation

It well known that the occurrence of outliers can seriously affect maximum likelihood estimation. In particular, the presence of contaminated data may lead to biased estimates of both the centre and the radius and completely invalidate the process of circular features identification. In order to cope with the occurrence of outliers, robust methods are needed [8,9]. In this contribution, it is suggested to adjust non-linear least squares to deal with contamination according to the idea of impartial trimming [10].
The optimization problem is based on a subset of the data after discarding that fraction of points with the largest distance from the fitted centre. The subset of trimmed points is not known in advance but a result of the methodology. Estimation and filtering are performed simultaneously. In other words, it is minimized
\[ \sum_i [d_i - r_0]^2 \]
with respect to \( \theta = (x_{10}, x_{20}, r) \) and the index set \( I \), where \( I \) is the index set of not trimmed points \( d_i \leq d_j \), \( \forall i \in I \land j \notin I \). The optimization method requires an iterative algorithm. After a suitable initialization, for a fixed trimming level \( \phi \) selected in the interval \((0, 0.5] \), the algorithm alternates between a concentration step [11], in which distances are sorted in non-decreasing order and the largest are discarded, and nonlinear least squares on the non-trimmed set, until convergence. The algorithm can be described as follows:

1) based on current parameter values compute geometric distances in absolute value 
\[ g_i = g(x_{10}, x_{20}, x_{30}, r_0) = |d(x_{10}, x_{20}, x_{30}) - r_0| \],
2) sort distances in non-decreasing order \( g_{(1)} \leq g_{(2)} \leq \ldots \leq g_{(n)} \) and remove those \( n(1-\phi) \) points with the largest distances,
3) minimize the objective function (2) over \( I = (i_{(1)}, i_{(2)}, \ldots, i_{(n(1-\phi))}) \) and update parameters values to \( \theta_i = \arg\min \sum_i [d(x_{10}, x_{20}, x_{30}) - r_0]^2 \),
4) set \( \theta_0 = \theta_1 \) and iterate until \( \max(\theta_0 - \theta_1) < tol \) for a sufficiently small tolerance value \( tol \).

After each iteration data points are more concentrated and close to the fitted centre, due to the trimming of the most distant points. Then, at each iteration the objective function (2) evaluated at the current estimates decreases. The roundness error can be fitted based on the circle's parameters estimates at convergence. It would be tempting to evaluate it on the resulting trimmed set but this strategy would determine an underestimation. Then, it is necessary to inflate the final estimate by a consistency factor [11]. The robust fit clearly depends on the selected level of trimming. The larger the trimming level the more robust is the final estimate. A trimming level larger than the (unknown) actual rate of contamination in the data at hand leads to some efficiency loss. On the contrary, when the level of trimming is smaller, then the procedure lacks of robustness to the occurrence of outliers. Therefore, the selection of the trimming level is one crucial aspect of the methodology. One solution is represented by the application of the methodology for subsequent decreasing levels starting from the upper bound 0.5. This monitoring approach may be computationally expensive but is very informative regarding the nature of the outliers [8]. In order to increase efficiency with respect to the initially chosen trimming level, one could consider a re-weighting step in which those observations whose fitted geometric distance is larger than a specified threshold value is only discarded. Here, such cut-off value is found adaptively as \( Q_3 + k(Q_3 - Q_1) \), where \( Q_1, Q_3 \) are the first and third quartile of the distribution of distances at convergence. A common choice is \( k = 1.5 \), as it usually happens in the construction of box-plots.

Another important computational aspect is represented by the selection of initial values for the unknown parameters. An effective strategy is represented by subsampling: an initial solution is found by running nonlinear least squares on a random subsample. The size of the subsample needs to be small in order to low the chance to include outliers and get useful starting values. Here, the subsamples of size three, that is the minimum number of points that allow to fit a circle, are considered. The algorithm is particularly robust with respect to the selection of the initial subset. However, practitioners are advised
to run the algorithm from several starting points, in order to avoid it to be trapped in local or even spurious solutions. Once several solutions have been found, one could select the solution leading to the lowest objective function (2) evaluated at the final estimates. Trimmed values are likely to be outliers. However, it may happen that some genuine observations are wrongly declared outliers and that some outlying points are not correctly identified. In the first case, there is a swamping effect, whereas, in the latter case there is a masking effect. An ideal trimming approach should not contemplate both swamping and masking errors. However, this is objective cannot be reached. Actually, a high level of trimming could determine a low masking but a large swamping, and the converse is true in the case of a low level of trimming. Therefore, it is necessary to set the trimming level in order to get a suitable trade-off between swamping and masking. One should reduce masking to the larger extent while keeping a low swamping. Actually, the use of a relatively large trimming level can aid to control masking, whereas the reweighting step allows to low swamping.

The proposed impartially trimmed non-linear least squares algorithm is based on an initial trimming level and the final estimate can be refined according to a reweighting step to gain some efficiency. At each step, a fixed rate of observations is removed from the fitting process. The same is true in RANSAC but the trimmed set is defined according to a certain fixed threshold rather than on a trimming level and there is not any safe strategy to select such threshold.

The impartially trimmed non-linear least squares algorithm returns the set I of non-trimmed points on which the fit is evaluated. Even if tempting, it is not possible to measure uncertainty around the fitted circle's parameters according to the inverse of the Hessian matrix evaluated at the robust estimate from (2), otherwise one would underestimate variability. Actually, a correct evaluation of uncertainty requires to take into account also the intrinsic variability due to trimming. Therefore, a fair uncertainty evaluation could be pursued according to a suitable bootstrap approach: data are re-sampled with replacement to obtain several bootstrap samples and the robust fit is evaluated over each of them. Then, uncertainty can be assessed through the variance of the estimates obtained over the bootstrap replicates.

3.1. Synthetic examples
Let us consider some synthetic example on simulated data in order to illustrate the methodology and its benefits in the presence of contamination in the circular data. The coordinates have been drawn independently from a pair of normal variates centred on the unit circle with standard error 0.05. Then, three outliers’ constellations have been taken into account, that are characterized by translated, scattered and clustered outliers, respectively. In the first case a fraction of points has been shifted by 0.25 along both coordinates, in the second case, outliers are uniformly distributed over the square region $[-1.5, 1.5] \times [-1.5, 1.5]$, in the third case, outliers are clustered around a different circular structure centred on one point of the true circumference.

Also two scenarios are considered. In the first, coordinates are measured all around the circle, whereas in the second, measurements are concentrated on the arc $[0, \pi/2]$. The sample size equal to 500 and 100, respectively is set. In both scenarios the rate of contamination is 20%. In each synthetic example, the circle according to the proposed impartially trimmed non-linear least squares algorithm with a trimming level 25% and reweighting based on $k = 1.5$, is fitted. Figure 1 displays the fitted circles in the scenario with points all around the circumference. It is possible note the two main features of the proposed technique. From the one hand, the methodology allows to fit a circle after eliminating the effect of outliers on the fit. Actually, the robust fitted circle cannot be distinguished from the true model, whereas the circle fitted by the classical non robust technique exhibits a non-negligible bias. On the other hand, the robust strategy allows to identify a set of potential outliers, whose nature may be worth of further investigation. In a similar fashion, figure 2 gives the results on the samples with points along the arc $[0, \pi/2]$. The robust technique still performs satisfactory and provide both accurate fits and outliers detection. On the contrary, the classical method suffers from severe bias.
Figure 1. Synthetic example. 500 points on $[0, 2\pi]$ with 20% contamination. Trimmed non-linear least squares fit (red solid line), non-linear least squares fit (dashed black line), true circumference (black solid line). Clockwise from top right: translated (a), scattered (b) and clustered outliers (c). The plain black circles give the trimmed points after reweighting.

Figure 2. Synthetic example. 100 points on the arc $[0, \pi/2]$ with 20% contamination. Trimmed non-linear least squares fit (red solid line), non-linear least squares fit (dashed black line), true circumference (black solid line). Clockwise from top right: translated (a), scattered (b) and clustered outliers (c). The plain black circles give the trimmed points after reweighting.
3.2. Numerical studies
In order to investigate the accuracy of the proposed technique a small numerical study is run. 100 points on the unit circle over different arcs that measures 60, 90, 120, 180, 360 degrees are considered. Two scenarios have been assumed. In the first, the data are not contaminated, whereas in the second 20 % contamination is added, shifting 20 points by 0.25 along both coordinates. The entries in table 1 give mean values averaged over 500 replications. It worth to note that differences are negligible when the data are not prone to contamination, whereas the accuracy provided by the trimmed approach is superior under contamination for all considered circular sectors.

Table 1. Mean values from trimmed NLS with 25 % trimming level and classical NLS.

| α   | r      | x₁₀    | x₂₀    | r      | x₁₀    | x₂₀    |
|-----|--------|--------|--------|--------|--------|--------|
|     |        |        |        |        |        |        |
| 60  | 1.287  | 0.288  | 0.163  | 1.456  | 0.468  | 0.270  |
| 90  | 1.156  | 0.168  | 0.126  | 1.223  | 0.175  | 0.178  |
| 120 | 1.196  | 0.109  | 0.183  | 1.385  | 0.204  | 0.356  |
| 180 | 1.002  | 0.001  | -0.001 | 1.057  | 0.003  | 0.057  |
| 360 | 1.002  | 0.000  | 0.000  | 1.002  | 0.000  | 0.000  |
|     |        |        |        |        |        |        |
|     |        |        |        | 20 % contamination |
| 60  | 1.001  | -0.007 | -0.005 | 0.548  | 1.048  | 0.618  |
| 90  | 1.022  | 0.012  | 0.013  | 1.135  | 0.504  | 0.494  |
| 120 | 1.080  | 0.035  | 0.072  | 1.510  | 0.420  | 0.699  |
| 180 | 0.993  | 0.003  | 0.015  | 1.244  | 0.062  | 0.299  |
| 360 | 1.001  | 0.003  | 0.005  | 1.006  | 0.050  | 0.051  |

4. Real data example
A ring gauge of nominal diameter 1.497 mm and a piece of LEGO [1] is considered. The cylindrical standard and the LEGO piece have been measured using the digital optical machine with 4.5x magnification and using diffused back-light (manufacturer: TESA, model: VISIO 300, resolution: 1 µm, magnification: 0.8x ÷ 4.5x, vision software: PC-Dmis, U(k - 2) = 2 µm)). Both elements were previously calibrated by an accredited laboratory. The ring pattern was calibrated using a Mitutoyo QuicK Vision digital optical machine and the value 1.496 ± 0.003 mm for the diameter is obtained. The Lego piece was measured using a CMM DEA FDG/SCIROCCO, leading to 4.886 ± 0.070 mm for the diameter. The estimates from the proposed methodology and classical maximum likelihood estimation through non-linear least squares are given in table 2.

Table 2. Fitted diameter from trimmed NLS with 25% trimming level and classical NLS.

|                  | Rewighted trimmed NLS | Raw trimmed NLS | Classical NLS |
|------------------|-----------------------|----------------|---------------|
|                  |                       |                |               |
| Ring gauge       | 1.4979                | 1.4994         | 1.4958        |
| Ring gauge, first 500 | 1.4762                | 1.4824         | 1.4485        |
| LEGO piece       | 4.8952                | 4.8965         | 4.8879        |
| LEGO piece, first 1000 | 5.0048                | 5.0044         | 4.7180        |

The number of measured coordinates for the ring gauge are 2989 all along the circumference. Also trimmed NLS and NLS to the case in which only the first 500 measurements have been used that are concentrated on one angular sector only, are applied. The number of points acquired for the Lego piece is 9313. The case with only the first 1000 points has been also analysed. The entries in table 1 give the fitted diameters. In all cases, the trimmed approach returns more accurate estimates. The authors note that the advantage over the classical approach is overwhelming when data from a circular sector is only used. Actually, estimates from NLS fall outside a 95 % tolerance intervals of expanded uncertainty.
Figures 3 and 4 display the coordinates measurements with the trimmed points. It is important to remark the ability of the procedure to detect anomalous points simultaneously with the fitting procedure.

![Figure 3. Ring gauge. Coordinates measurements with trimmed points (blue + red) and trimmed points after reweighting (red), for all the data (a) and the first 500 points (b).](image)

![Figure 4. Lego piece. Coordinates measurements with trimmed points (blue + red) and trimmed points after reweighting (red), for all the data (a) and the first 1000 points (b).](image)

5. Conclusions
A technique to fit circular profiles in the presence of possible outliers was proposed. The proposed methodology performs satisfactory: it prevents fitting from severe bias under contamination and allows to detect outliers through the inspection of the trimming set. Computational time always lies in a feasible range. The re-weighting step avoids the final fit to be dependent on the initial level of trimming. However, it is advised to run the procedure for different trimming level and compare the results.

Acknowledgements
This work has been carried out within the framework of Project PID2019-110199GB-I00 “Mediciones En proceso mediante Tecnologías de Reconocimiento de Imágenes orientadas a su aplicación en entornos de producción industrial (METRI)”, belonging to the call 2019 of the Proyectos de I+D corresponding to the Plan Estatal de Investigación Científica y Técnica y de Innovación 2017-2020.

References
[1] Caja J, Maresca P, Gomez E, Barajas C and Berzal M 2014 Metrological characterization of interior circular features using digital optical machines. Calculation models and application
scope Precision Engineering 38 pp 36–47
[2] Yang C, Caja C and Gómez E 2018 Comparison of methods for outlier identification in surface characterization Measurement 117 pp 312–325
[3] Fischler M and Bolles R 1981 Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography Commun. Assoc. Comp. Mach. 24 pp 381–395
[4] Coope I 1993 Circle fitting by linear and non-linear least squares Journal of Optimization Theory and Applications 76 pp 381–388
[5] Chan Y T, Lee B H and Thomas S M 2005 Approximate Maximum Likelihood Estimation of Circle Parameters J Optim Theory Appl 125 pp 723–734
[6] Nelder J A and Mead R 1965 A simplex algorithm for function minimization Computer Journal 7 pp 308–313
[7] Byrd R H, Lu P, Nocedal J and Zhu C 1995 A limited memory algorithm for bound constrained optimizatin SIAM Journal on Scientific Computing 16 pp 1190–1208
[8] Farcomeni A and Greco L 2015 Frobust methods for data reduction (New York: CRC Press)
[9] Greco L, Maresca P and Caja J 2019 Big Data and Advanced Analytics in Industry 4.0: a comparative analysis across the European Union Procedia Manufacturing 41 pp 383–390
[10] Cuesta-Albertos J A, Gordaliza A and Matrán C 1997 Trimmed k-means: an attempt to robustify quantizers Ann. Statist. 25 (2) pp 553–555
[11] Rousseeuw P J, Van Driessen K 1999 A fast algorithm for the minimum covariance determinant estimator Technometrics 4 (3) pp 212–223