Maximum Likelihood instead of Least Squares in fracture analysis by means of a simple Excel sheet with VBA macro

Vincenzo Guerriero ( vincenzo.guerriero@univaq.it )
University of L’Aquila

Method Article

Keywords: Maximum Likelihood, Fracture statistical analysis, Monte Carlo Method, MS Excel, Visual Basic for Applications

Posted Date: June 28th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3042967/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Abstract

This short note illustrates a linear regression algorithm based on Maximum Likelihood Estimation, with related Excel sheet and VBA program, adapted to the case of fracture aperture data sets in which sampling of smallest values was problematic. The method has been tested using Monte Carlo simulations and exhibits surprisingly better convergence against Least Squares criterion. As the method is conceptually simple and, following the here illustrated indications, the relative spreadsheet is easily achievable, it may be routinely used instead of the Least Squares in fracture analysis, as well as all those numerous cases in geology and geophysics in which there are systematic biases at the lower limits of the sampling scale.

1. Introduction

In order to study statistical features of aperture values in stratabound fracture sets, Guerriero et al. (2015) carried out an accurate statistical investigation involving two Lower Cretaceous (Albian) carbonate successions, outcropping at Faito and Chianello Mts. (southern Italy). These successions, which had been selected as surface analogues of buried oil reservoirs in Val D’Agri (Basilicata, Italy), were previously well studied in terms of sedimentology and petrophysics (Giorgioni et al., 2016; Iannace et al., 2008 and references therein), as well as of geological structural settings (Guerriero et al., 2010, 2011, 2013; Vitale et al., 2012). It should be taken into account that, while stratabound fracture spacing has been extensively analysed (e.g., Odling et al., 1999, Bai and Pollard, 2000; Guerriero and Mazzoli, 2021 and references therein), their aperture is little studied. The statistical analysis carried out by Guerriero et al. (2015) pointed out that stratabound joint aperture depends on bed thickness, according to a linear function, characterized by a non-zero constant term. This may have important implications for the hydraulic and structural characterization of fractured rocks (Guerriero et al., 2015). Nevertheless, in that analysis, due to evident difficulties in field sampling of joint aperture, it was decided to set a minimum threshold value of 0.265 mm for them, so including into this single category all detected aperture values lesser than this limit. In several cases, this artifact may induce a marked tendency of the best fit line to intersect the ordinate axis near this threshold (Guerriero et al., 2015). As the constant term value identified by Guerriero et al. (2015) was just close to this limit, the suspicion arose that it could have been affected by an artifact. Therefore, it has been decided to repeat the analysis of those data using a more effective method, based on Maximum Likelihood Estimation (MLE). Illustrating the results of such analysis goes beyond the scope of this work and these will be presented in a companion paper to be published later. The aim of this paper is to describe the adopted statistical method, the details about its validation and effectiveness, as well as the utilized algorithm and software.

2. Methods

1.1. Linear regression by means of Maximum Likelihood Estimation
A simple probabilistic model for joint aperture value distribution, compatible with field fracture data is of Log Normal kind, in which the median is a linear function of mechanical bed thickness, and the standard deviation is constant (Guerriero et al., 2015). To estimate the aperture-bed thickness regression line parameters, in this work the MLE (e.g., Dekking et al., 2005) is used rather than that of Least Squares (LSM), in order to reduce biases related to small fracture sampling. Furthermore, in this analysis, residuals $r_j$ are calculated as difference between the logarithms of the observed and predicted values (Guerriero et al., 2015):

$$r_j = \ln z_j - \ln y_j ; \quad (1)$$

Let denote by $x_i$ the limit value between contiguous aperture classes $s_i$. In this instance $x_i$ is an intermediate value between $s_i$ and $s_{i+1}$, opportunebly chosen (Fig. 1). Under the hypothesis that aperture values exhibit Log Normal distribution $F_{mn,d}(x)$, whose mean is $\ln y$ and standard deviation is $d$ (calculated according to Eq. 1). Here $y$ is a linear function of bed thickness $T$, whose parameters are coefficient $m$ (mm/cm) and constant term $n$ (mm). Then the probability that a measured aperture value $S_i$ falls within the class $s_i$, here denoted by $p_{mn,d}(s_i)$, for $i > 1$, is:

$$\Pr(x_i < S < x_{i+1}) = p_{mn,d}(s_i) = F_{mn,d}(x_i) - F_{mn,d}(x_{i+1}); \quad i > 1; \quad (2)$$

Whilst, for the first aperture class:

$$\Pr(S < x_1) = p_{mn,d}(s_1) = F_{mn,d}(x_1); \quad i = 1. \quad (3)$$

Therefore, for a given sample $\{S_k, T_k\}$ the Likelihood Function $L(m,n,d)$ (e.g. Dekking et al., 2005), assumes the following form:

$$L(m,n,d) = p_{mn,d}(S_1) \cdot p_{mn,d}(S_2) \cdots p_{mn,d}(S_k) \cdots (4)$$

Searching for the maximum of this function on the space of the three parameters $m,n$ and $d$ (numerically), the maximum likelihood estimates of these three parameters are achieved.

Should be noted as according to the LSM, for each aperture value belonging to the 0.265 mm class, it is stated that this is substantially close to the theoretical one (false statement) whereas, according to the MLE (i.e, according to Eq. 3), it is stated that the observed value is less than or equal to the value $x_f$ (true statement). This is the substantial difference between MLE and LSM.

### 1.2. The Excel sheet and VBA program

The Excel folder utilized in this work includes three sheets: MLE, LSM and Results. Figure 1 illustrates in detail the sheet MLE; with respect to this latter the sheet LSM is different only in column K cells, whose formula calculate the square of residual in the adjacent cell in column J (). The first 30 rows, which are not depicted here, include a header illustrating some user instructions. The routines, written in Visual
Basic for Applications (VBA), which utilize this folder to analyse data and carry out Monte Carlo simulations are:

- **Sub Maximize()** and **Sub Minimize_LS()**: analyse a data set by maximizing or minimizing an object function, which is Log likelihood for the former and residual standard deviation for the latter.

- **Sub Simul_Apert_Data()**: based on field data (#7 in Fig. 1) and on model true values of \( m, n \) and \( d \) (#2), it produces a 35 item data set by (i) resampling thickness data and (ii) producing, for each thickness, a random aperture value. Then, it identifies which class it belongs to and its limits (#6).

- **Sub Simul_100()**: For each triplet of true values \( m, n \) and \( d \), it produces 100 simulated data sets and analyses each one by means of **Sub Maximize()**, in sheet MLE, and **Sub Minimize_LS()**, in sheet LSM. Then it saves the estimated values in columns S, T and U.

- **Sub Monte_Carlo()**: varies the \( n \) true value in the range 0.05–4.75 mm, and for each value produces simulations by calling **Sub Simul_100()**, then saves the related results (#11) in sheet Results (Fig. 2).

The core of the calculation method in the **MLE** sheet is in formulas in column \( K \), in which for each aperture value (in column \( E \), #6) the probability logarithm that it falls within the range to which it belongs, is calculated as follows:

\[
\ln(\text{LOGNORMDIST}(G60;\ln(I55*C60+I57);J48)-\text{LOGNORMDIST}(F60;\ln(I55*C60+I57);J48))
\]

The formula \( \text{LOGNORMDIST}(G60;\ln(I55*C60+I57);J48) \) provides the probability that an aperture value is lesser or equal to the limit in cell G60, when its median value is a linear function of thickness (term \( \ln(I55*C60+I57) \)) and its standard deviation is the value \( d \) in cell J48. The difference between these distributions furnishes the probability that an aperture value lies within the range limited by cells G60-F60. The sum of logarithms of these probabilities returns the Log likelihood (e.g., Dekking et al., 2005) in cell J55. The routine **Sub Maximize()** starts calculations by assigning likely initial values to parameters \( m, n \) and \( d \) achieved by Excel least squares functions in cells J46:L46 (#4) by means of instructions such as:

\[
\text{Range("I55") = Range("K46")}
\]

\[
\text{Range("I57") = Range("L46")}
\]

\[
\text{Range("J48") = Range("J46")}
\]

then, it iteratively adjusts the values of parameters \( m, n \) and \( d \), in cells I55, I57 and J48 respectively (#3), according to a simple steeper descent algorithm, until the maximum of the Log Likelihood (cell J55) is reached.

1.3. Validation by means of Monte Carlo simulation

A series of Monte Carlo simulations was carried out to verify MLE effectiveness and to make a comparison with the LSM, with parameter \( n \) (which was critical in our analysis) varying in the range 0.05
From sheet MLE, the Sub Monte_Carlo() is called, which assigns, as “true” values from which simulated data are produced, \( m = 0.3, \ d = 0.2 \) and varies \( n \), starting from 0.05 to 0.475 with step of 0.025. For each \( n \) value, 35 thickness values are chosen from field data by means of random numbers (Sub Simul_Apert_Data()). For each thickness value, an aperture one is produced by means of a random number and of the function in cell F52:

\[
\text{LOGINV(E52;LN(H41*G52 + H43);I48)}
\]

which returns the inverse Log Normal distribution of: (1) random number, (2) mean as logarithm of a linear function with \( m \) value from cell H41 and \( n \) from H43 and (3) standard deviation from cell I48. Then, this subroutine individuates the class and related limits, to which this value belongs, and saves these in columns E, F and G. Hence, the Sub Maximize() is called, which maximize the Log Likelihood. Then, the simulated data set is copied within sheet LSM, and the best fit line is calculated, by minimizing the sum of square of residuals, calculated according to Eq. 1 (Sub Minimize_LS()).

For each \( n \) value, this procedure is repeated 100 times, then \( (m, n, d) \) are saved in columns U, T and S respectively. After 100 iterations, the average and standard deviation of these column values are saved into the sheet Results. Then, \( n \) is incremented of 0.025 mm.

### 3. Results discussion and concluding remarks

Figure 3 shows the results of Monte Carlo simulations in which the MLE and modified LSM are compared. The MLE shows a surprising effectiveness in the \( n \) range values 0.05–0.25, just where the LSM fails. In calculating the coefficient and constant term, the MLE estimator shows modest deviation from the true value as early as \( n = 0.05 \) and converges for \( n \geq 0.1 \). Also in estimating the standard deviation \( d \) of Log aperture value (i.e., the standard deviation of residuals according to Eq. 1) the MLE exhibits significant better convergence against LSM, over the whole studied range. Therefore, MLE is particularly effective in analyzing fracture data sets in which field measurement of minor fracture aperture are problematic. As this linear regression method can be easily performed with an Excel spreadsheet, it may be routinely used in fracture analysis and in various experimental situations where biases occur at the lower limit of the sampling scale.

### Declarations

#### Funding

Ministry of Economic Development (MiSE), Italy, Grant Id: C19C20000520004

#### Conflicts of interest/Competing interests

No, I declare that the authors have no competing interests as defined by Springer, or other interests that might be perceived to influence the results and/or discussion reported in this paper.
Availability of data and material (data transparency)

Name of the code/library: MonteCarlo_MLE.xlsm

Contact: vincenzo.guerriero@univaq.it; vincenzo.guerriero@unina.it

Program language: Visual Basic, Applications Edition (VBA)

Software required: MS Office or equivalent

Program size: 82 KB

The source codes are available for downloading at the link: https://github.com/vincenzo-guerriero/MonteCarlo_MLE.git

Authors' contributions

Author individual contribution are as follows. Conceptualization: Guerriero V.; Data curation: Guerriero V.; Formal analysis: Guerriero V.; Investigation: Guerriero V.; Methodology: Guerriero V.; Software: Guerriero V.; Supervision: Guerriero V.; Validation: Guerriero V.; Visualization: Guerriero V.; Roles/Writing - original draft: Guerriero V.; Writing - review & editing: Guerriero V.

Acknowledgements

The research leading to these results has received funding from the Italian Ministry of Economic Development (MiSE) under the project “SICURA – CASA INTELLIGENTE DELLE TECNOLOGIE PER LA SICUREZZA CUP C19C20000520004 - Piano di investimenti per la diffusione della banda ultra larga FSC 2014-2020”

References


Figures
Figure 1

Spreadsheet utilized for simulations. Yellow cells with red font denote variables to be adjusted in order to maximize the object function. Dark blue cells with yellow font denote parameter true values. #1: aperture classes and related limits. #2: True $m$, $n$ and $d$ values of the model simulating fracture data. #3: Estimated $m$, $n$ and $d$ values, by MLE. #4: Estimated $m$, $n$ and $d$ values, by Excel linear regression functions. #5: Excel function LOGINV() to produce a single random aperture value, starting from a bed thickness value. #6: Simulated data set; from left: resampled bed thickness value, fracture aperture class and its upper and lower limits. #7: Field bed thickness data. #8: Likelihood data; this is the core of this calculus method. From the left: expected aperture value (from linear law), residual of Log aperture, according to Eq. 1, and Log of probability calculated by means of Excel function LOGNORMDIST() (see main text). #9: Object function; in sheet MLE, it is the Log Likelihood, in sheet LSM it is the sum of square of residuals. #10: Data to build up probability plots of residuals. Use of these plots in fracture analysis lies outside the scope of this paper and is illustrated by Mazzoli et al., (2009). #11: Monte Carlo simulation output data. Three columns on the left include each one 100 estimated values of $m$, $n$ and $d$; on the right side average and standard deviation of these columns.
Spreadsheet “Results”. The diagrams on the left side point out as MLE estimators exhibit modest deviations over the whole analysed range of true $n$ values and excellent convergence for $n>0.1$, whereas LSM ones converge only for $n>0.3$, showing large deviations elsewhere.

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.

- [Highlights.docx](#)