Grain size estimation in polycrystals: solving the corpuscle problem using Maximum Likelihood Estimation

Dorian Depriester[a,*], Régis Kubler[a]

[a]Arts et Metiers Institute of Technology, MSMP, HESAM Université, F-13617 Aix-en-Provence, France

ARTICLE INFO

Keywords:
Microstructure
Grain size distribution
Corpuscle problem
Wicksell
Stereology
Maximum Likelihood

ABSTRACT

In materials science, the microstructures of materials are generally characterized by 2D observation (e.g. electron microscopy). For polycrystalline materials, such as crystalline rocks or ceramics, those observations can be used to measure the grain size distribution. However, the fact that grain sizes are measured in planar cuts introduces a statistical bias, since the real (3D) grain sizes cannot be directly measured. For almost spherical grains, this bias can be computed thanks to the so-called Wicksell's equation. This paper proposes a method, based on Maximum Likelihood Estimation (MLE) for *unfolding* the apparent 2D distribution. The efficiency of this method is extensively investigated in the special case of lognormal distribution. In this case, 10% uncertainty on the distribution parameters can be reached with only 580 empirical values.

Notations

D_n KS goodness-of-fit test againt Ω_n.

E Expectation of R.

f Parametric PDF.

F^{-1} Quantile function.

$\tilde{f}$ Wicksell transform of f.

$\tilde{F}$ CDF of the Wicksell transform of f.

$\tilde{F}_n$ Empirical CDF, computed from n values of r.

$\mathcal{L}_n$ Likelihood function, computed on Ω_n.

n Sample size.

*Corresponding author
dorian.depriester@ensam.eu (D. Depriester); regis.kubler@ensam.eu (R. Kubler)
ORCID(s): 0000-0002-2881-8942 (D. Depriester); 0000-0001-7781-5855 (R. Kubler)

Ω_n Sample, consisting in values $(r_1, r_2, ..., r_n)$.

r Equivalent radius in 2D.

R Equivalent radius in 3D.

R_{co} Cut-off value for histogram decomposition.

σ Shape parameter for lognormal distribution.

$\boldsymbol{\theta}$ Vector containing the parameters for a given parametric distribution.

$\hat{\boldsymbol{\theta}}$ Estimator, given by MLE.

1. Introduction

1.1. The corpuscle problem

The physical behaviour of polycrystalline materials (such as rocks, metals or ceramics) is largely related to the grain size distribution. For rocks, grain size affects the creep behaviour (Schmalholz and Duretz, 2017; Bose et al., 2018) and strain localization (Czaplińska et al., 2015). Thus, the characterization of those materials usually involves the measurement of such distribution in terms of mean value and spread as well. In some cases, this statistical description can be used to generate synthetic grain size distributions that replicate those of the material, in order to perform numerical simulations of the behaviour of polycrystalline aggregates. For instance, the Neper software (Quey et al., 2011) can be used to perform Finite Element Analysis (FEA) on polycrystalline aggregates, starting from grain size distributions on representative aggregates. The evaluation of 3D grain size distribution is also necessary when comparing samples (Cashman and Marsh, 1988, e.g.) or when studying the time-evolution of grain sizes, for instance during grain growth (Zöllner and Streitenberger, 2006). 3D imaging techniques, such as X-ray tomography or serial sectioning, allow to image the 3D geometry of grains, providing valuable information about the grain size distribution. Such techniques have become more and more popular for the last decades; however, it is still more common to characterize the microstructures from 2D observations, such as Optical Microscopy (OM), Scanning Electron Microscopy (SEM) or Electron Backscattered Diffraction (EBSD). Indeed, the latter are easier to use and are *a fortiori* cheaper than 3D imaging.

When the grains are polyhedrons (assuming convex geometry with planar grain boundaries), the apparent grains in 2D sections are convex polygons. The equivalent radius R of each grain is then estimated from its volume V so that:

$$R = \sqrt[3]{\frac{3V}{4\pi}} \tag{1}$$

whereas the equivalent radius r of an apparent grain is estimated from its area S:

$$r = \sqrt{\frac{S}{\pi}}\,. \tag{2}$$

The relationship between the distribution of r and R is usually referred to as the corpuscle problem. This issue falls in the general field of stereology. Sahagian and Proussevitch (1998) have numerically generated cross-sections of elementary shapes (spheres, ellipsoids and parallelepipeds) in order to build tabular relationships between the histogram of apparent sizes and the original size of the particle. The corpuscle problem can also be solved by spatial tessellation methods. For instance, Chiu et al. (2013, Chap. 10) established links between planar section and spatial tessellations. Using Laguerre–Voronoï tesselation, a method has been proposed in a previous work (Depriester and Kubler, 2019a) to evaluate the 3D grain size distribution from the size distribution of apparent polygons. However, this method only works when assuming that the underlying distribution is lognormal.

To solve the corpuscle problem in more general cases, it is usually assumed that grains can be considered as non overlapping spheres. This assumption seems reasonable for fully recrystallized materials since the average number of faces per grain are typically larger than 12 in metals and ceramics (Depriester and Kubler, 2019a, and references therein); this assumption is also valid for spherical inclusions embedded in a matrix, like quartz grains in mylonite (Lopez-Sanchez and Llana-Fúnez, 2016) or in permafrost (King et al., 1988), garnets in schist (Kretz, 1993), olivine in plagioclase/augite matrix (Farr et al., 2017), pores in basalt (Al-Harthi et al., 1999) or vesicular

basalts in volcanic rocks (Sahagian and Proussevitch, 1998). It is worth pointing that this approach also assumes that the grains are randomly distributed in space, hence it neglects the neighbouring or clustering effect which can occur under certain circumstances (e.g. Czaplińska et al., 2015). Sahagian and Proussevitch (1998) have shown that the spherical approximation leads to little bias when the grains have elliptic or parallelipedic shapes, as long as the aspect ratio is small (< 1.5) ; Farr et al. (2017) came to the same conclusion when using the method of moments on sections of cubes.

If a sphere of radius R is cut at random latitude, the radius of such a disk is $r \leq R$, as illustrated in Figure 1. From this figure, it comes that the probability of finding a disk whose radius r is

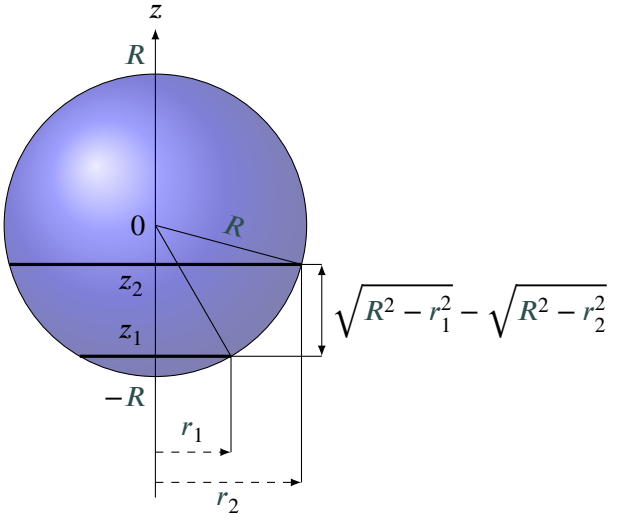


Figure 1: Representation of a sphere of radius R cut at random latitudes $z_1, z_2 \in [-R, R]$.

between r_1 and r_2 is (Sahagian and Proussevitch, 1998):

$$P\left(r_1 < r < r_2\right) = \frac{1}{R}\left(\sqrt{R^2 - r_1^2} - \sqrt{R^2 - r_2^2}\right) . \tag{3}$$

The most widely used technique to *unfold* the apparent 2D grain size distribution into the real (3D) one is the Saltykov method (Saltykov, 1967). It is an iterative method working with the finite histogram. Starting from the upper class (larger value for r), Eq. (3) is recursively used to evaluate the histogram for R. This method is simple to implement and provides accurate results, but it is

dependent on the number of bins used for describing the distribution. There is no "golden rule" for choosing such a number, but it usually ranges between 10 to 20 (Lopez-Sanchez and Llana-Fúnez, 2016; Depriester and Kubler, 2019b). A criterion for choosing it has been proposed in the latter reference.

In some cases, one may prefer working on a continuous distribution instead of a finite histogram. Lopez-Sanchez and Llana-Fúnez (2016) have proposed the so-called two-step method, which consists in fitting an underlying parametric distribution on the histogram given by the Saltykov method. In practice, they assumed that the underlying distribution was lognormal. In order to increase the accuracy, the present authors have recently proposed a set of "correction equations" for the two-step method (hence the three-step method) in the special case of lognormal distribution (Depriester and Kubler, 2019a).

Consider a medium composed of an infinite number of spheres whose radii R follow a distribution with Probability Density Function (PDF) $f(R)$; Wicksell (1925) has demonstrated that if a series of planar cuts of this medium are made at random latitudes, the PDF related to the apparent radii r of such cuts is:

$$\tilde{f}(r) = \frac{r}{E} \int_r^\infty \frac{f(R)}{\sqrt{R^2 - r^2}} \mathrm{d}R \tag{4}$$

where E denotes the expectation for R:

$$E = \int_0^\infty R f(R) \mathrm{d}R \,. \tag{5}$$

Obviously, the related Cumulative Density Function (CDF) for apparent radius is:

$$\tilde{F}(r) = \int_0^r \tilde{f}(\psi) \mathrm{d}\psi \,. \tag{6}$$

$\tilde{f}$ can be considered as the Wicksell transform of f. As a result, the corpuscle problem can be reduced to the following question: how to invert the Wicksell's equation (4)? In his original paper,

Wicksell has proposed the following solution for his equation:

$$f(R) = \frac{-2ER}{\pi} \int_R^\infty \frac{\mathrm{d}}{\mathrm{d}r} \left(\frac{\tilde{f}(r)}{r} \right) \frac{\mathrm{d}r}{\sqrt{r^2 - R^2}} \,. \tag{7}$$

In practice, one usually wants to find f based on a finite sample $\Omega_n = \left(r_1, r_2, ..., r_n\right)$ measured experimentally. In this case, the empirical CDF is defined as follows:

$$\tilde{F}_n(r) = \frac{\text{Number of elements in } \Omega_n \leq r}{n} \,. \tag{8}$$

Since $\tilde{F}_n$ is not continuous, it is not differentiable. As a result, $\tilde{f}$ in Eq. (7) cannot be evaluated without regularization (e.g kernel density estimation).

One way to avoid it is to assume that f comes from a given parametric distribution (e.g. normal or lognormal). In a previous work, Depriester and Kubler (2019b) have used the Minimum Distance Estimation (MDE), based on the transformed CDF $\tilde{F}$, to unfold the distribution. The results were similar to those obtained through the two-step method (Lopez-Sanchez and Llana-Fúnez, 2016).

Usually, parametric distributions are fitted on empirical data using Maximum Likelihood Estimation (MLE) rather than MDE. Let $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p)$ be the vector containing the p parameters for the considered distribution. Given a finite sample Ω_n and a parametric distribution whose PDF is f, MLE consists in finding the value for $\boldsymbol{\theta}$ which maximises the likelihood function, defined as follows:

$$\mathcal{L}_n(\boldsymbol{\theta}) = \prod_{i=1}^{n} f\left(r_i \,|\, \boldsymbol{\theta}\right) \,. \tag{9}$$

Such value of $\boldsymbol{\theta}$, denoted $\hat{\boldsymbol{\theta}}$ below, is usually referred as the estimator. However, it appears that MLE is only sparsely used in the literature for solving the corpuscle problem (Keiding and Jensen, 1972; Kong et al., 2005; Chan and Qin, 2016, e.g.).

Some authors use the moments of the distribution rather than the full empirical PDF or CDF (Kong et al., 2005; Farr et al., 2017) to solve the Wicksell's equation. Indeed, this method is easier to implement than MLE because the moments of the transformed distribution can be analytically

computed in some cases (see for instance appendix in Farr et al., 2017). Nevertheless, Kong et al. (2005) have concluded that MLE resulted in better accuracy (in terms of goodness-of-fit) than the method of moments.

When unfolding the 2D distribution, the confidence interval on the results is of great importance, regardless of the unfolding method. Still, only few authors investigate this issue. Obviously, this interval depends on the sample size (n): the higher, the better. It is usually assumed that at least 1000 unique values are required for performing the Saltykov method (Lopez-Sanchez and Llana-Fúnez, 2016). Depriester and Kubler (2019b) have concluded about the same requirement concerning MDE. For the specific case of a lognormal distribution, Farr et al. (2017) have shown that the method of moments has a decreasing accuracy with increasing distribution spread (σ hereafter, see Eq. (16)). For instance, they reported that the required number of values for reaching 10% uncertainty on the volume-weighted mean diameter ranged from 19 (when $\sigma = 0.22$) to 980 (when $\sigma = 0.58$). Farr et al. have also shown that the standard error when evaluating σ with the moment method was proportional to $n^{-0.5}$.

1.2. Wicksell transform of a finite histogram

Consider a discrete distribution whose PDF is given by a the finite histogram:

$$f_{\text{hist}}\left(R \mid \text{bins}\right) = \sum_{k=1}^{N} \text{Freq}^k f_{\text{uni}}\left(R \,\middle|\, R_{\min}^k, R_{\max}^k\right) \tag{10}$$

where $f_{\text{uni}}(\boldsymbol{R} \mid \boldsymbol{R}_{\min}^k, \boldsymbol{R}_{\max}^k)$ denotes the PDF associated to the uniform distribution between $R_{\min}^k$ and $R_{\max}^k$ and Freq^k is the relative frequency of the k-th bin. In a previous work, Depriester and Kubler (2019b) have shown that the Wicksell's equation (4) becomes:

$$\tilde{f}_{\text{hist}}\left(r \mid \text{bins}\right) = \frac{1}{E} \sum_{k=1}^{N} \text{Freq}^k E^k \tilde{f}_{\text{uni}}\left(r \,\middle|\, \boldsymbol{R}_{\min}^k, \boldsymbol{R}_{\max}^k\right) \tag{11}$$

where E^k is the mean value of class k (i.e. k-th mid-point) and $\tilde{f}_{\text{uni}}$ is the Wicksell transform of the uniform distribution:

$$\tilde{f}_{\text{uni}}\left(r \mid R_{\min}, R_{\max}\right) = \begin{cases} \frac{2}{R_{\max}^2 - R_{\min}^2} \log\left(\frac{R_{\max} + \sqrt{R_{\max}^2 - r^2}}{R_{\min} + \sqrt{R_{\min}^2 - r^2}}\right) & \text{if } r \leq R_{\min} \\ \frac{2}{R_{\max}^2 - R_{\min}^2} \log\left(\frac{R_{\max} + \sqrt{R_{\max}^2 - r^2}}{r}\right) & \text{if } R_{\min} < r < R_{\max} \\ 0 & \text{if } R_{\max} \leq r \end{cases} . \tag{12}$$

In Eq. (12), $R_{\min}$ and $R_{\max}$ denote the lower and upper bounds for uniform distribution, respectively. Likewise, the transformed CDF of a finite histogram is:

$$\tilde{F}_{\text{hist}}\left(r \mid \text{bins}\right) = \frac{1}{E} \sum_{k=1}^{N} \text{Freq}^k E^k \tilde{F}_{\text{uni}}\left(r \mid R_{\min}^k, R_{\max}^k\right) \tag{13}$$

with:

$$\tilde{F}_{\text{uni}}\left(r \mid R_{\min}, R_{\max}\right) = \begin{cases} 1 - \frac{\gamma(r) + r^2 \log\left(R_{\min} + \sqrt{R_{\min}^2 - r^2}\right) - R_{\min}\sqrt{R_{\min}^2 - r^2}}{R_{\max}^2 - R_{\min}^2} & \text{if } r \leq R_{\min} \\ 1 - \frac{\gamma(r) + r^2 \log(r)}{R_{\max}^2 - R_{\min}^2} & \text{if } R_{\min} < r < R_{\max} \\ 1 & \text{if } R_{\max} \leq r \end{cases} \tag{14}$$

and

$$\gamma\left(r\right) = R_{\max}\sqrt{R_{\max}^2 - r^2} - r^2 \log\left(R_{\max} + \sqrt{R_{\max}^2 - r^2}\right) . \tag{15}$$

1.3. The lognormal distribution

The lognormal distribution is defined such that its PDF is:

$$f_{\log\mathcal{N}}\left(R \mid \mu, \sigma\right) = \frac{1}{R\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(\ln R - \mu\right)^2}{2\sigma^2}\right) \tag{16}$$

where σ and μ are the shape and scale parameters of the lognormal distribution, respectively. Let E be the expectation; it comes:

$$E = \exp\left(\mu + \frac{\sigma^2}{2}\right) . \tag{17}$$

In various science domains, the particle size approximately follows a lognormal distribution. Indeed, Table 1 provides a couple of such examples reported in the literature. It is worth mentioning that, based on the original work by Rhines and Patterson (1982), this hypothesis is now usually assumed for grain sizes in recrystallized materials (Heilbronner and Bruhn, 1998; Berger et al., 2011; Lopez-Sanchez and Llana-Funez, 2015; Lopez-Sanchez and Llana-Fúnez, 2016, e.g). Table 1 also shows some typical values for the scale parameter σ. It appears that, except for bubble sizes in wheat dough, $[0.2, 1]$ can be considered as a satisfactory range for σ.

Table 1
Examples of lognormal size distribution and typical values for the shape parameter σ.

Material	σ	Reference
Bacillus subtilis bacteria	0.24	(Yamamoto and Wakita, 2016)
Grains in 99.998% aluminum	0.29 to 0.44	(Rhines and Patterson, 1982)
Quartz grains in mylonite	0.46 to 0.53	(Lopez-Sanchez and Llana-Fúnez, 2016)
Grains in Alpha titanium	0.45 to 0.57	(Conrad et al., 1985)
Olivine grains in basalt	0.59	(Farr et al., 2017)
Grains in IN100 superalloy	0.64	(Tucker et al., 2012)
Pores in soils	0.52 to 1.15	(Kosugi and Hopmans, 1998)
Pores in cement paste	0.90 to 0.95	(Holly et al., 1993)
Bubbles in wheat flour dough	1.62 to 1.79	(Bellido et al., 2006)

1.4. Aims of this work

It appears that MLE is rarely used in the literature for solving the corpuscle problem, despite its popularity in statistical applications and its robustness, compared to the methods based on histograms (like Saltykov). This is probably because the Wicksell transform (4) is not straightforward to evaluate. Therefore, the aim of this work is first to propose an efficient way to numerically compute the Wicksell transform of a given distribution, circumventing the improper integral. This allows for evaluating the efficiency of MLE in terms of robustness and confidence interval for the

resulting estimator. Since a lognormal distribution is often met in corpuscle problems, this special case is investigated. Thus, this paper aims to investigate the accuracy of MLE in this case.

This paper first proposes an efficient method for computing the Wicksell transforms (CDF and PDF) of a continuous distribution, based on histogram decomposition of the underlying distribution. Then, an algorithm is proposed to rapidly generate synthetic data representing the process of random sectioning of spheres whose radii follow a given distribution. The robustness of this algorithm, with respect to asymptotic theory, is shown. Those synthetic data are used to evaluate the accuracy of MLE in the special case of a lognormal distribution. This analysis allows to estimate the required number of empirical values (n) for performing MLE; this also provides a tool for estimating the uncertainty on the estimators given by MLE. The efficiency of MLE applied on polyhedrons (instead of spheres) is also investigated.

The methods introduced in this paper are finally applied to two real-world datasets, taken from the literature. Thus, the 3D grain size distributions of two materials (namely uranium dioxide and mylonite) are estimated by MLE. The results are compared to those from other stereology techniques (two-step method and MDE).

2. Numerical methods

This section describes the numerical methods used for computing the Wicksell transforms of parametric distributions, circumventing the improper integral in Eq. (4). It also introduces a method for numerically generating datasets corresponding to the corpuscle problem, that is emulating the cross-sectioning principle of spheres whose radii follow a given parametric distribution. The framework in which all the methods have been implemented is also described.

2.1. Convert the continuous distribution into finite histogram

Let f be a continuous PDF. It can be discretized into an histogram consisting in a series of N bins such that:

$$f(R) \approx f_{\text{hist}}(R \mid \text{bins}) \tag{18}$$

where f_{hist} denotes the PDF of the finite histogram, as introduced in Eq. (10). The histogram approximation (18) is made here by constant-quantile decomposition, neglecting the last quantile, such that the bounds of the k-th bin are:

$$R^k_{\text{min}} = F^{-1}\left(\frac{k-1}{N+1}\right), \tag{19a}$$

$$R^k_{\text{max}} = F^{-1}\left(\frac{k}{N+1}\right). \tag{19b}$$

In Eqs. (19), F^{-1} denotes the inverse of CDF (also known as the quantile function) for the underlying distribution. Figure 2 illustrates such decomposition. For the sake of clarity, only 19 bins

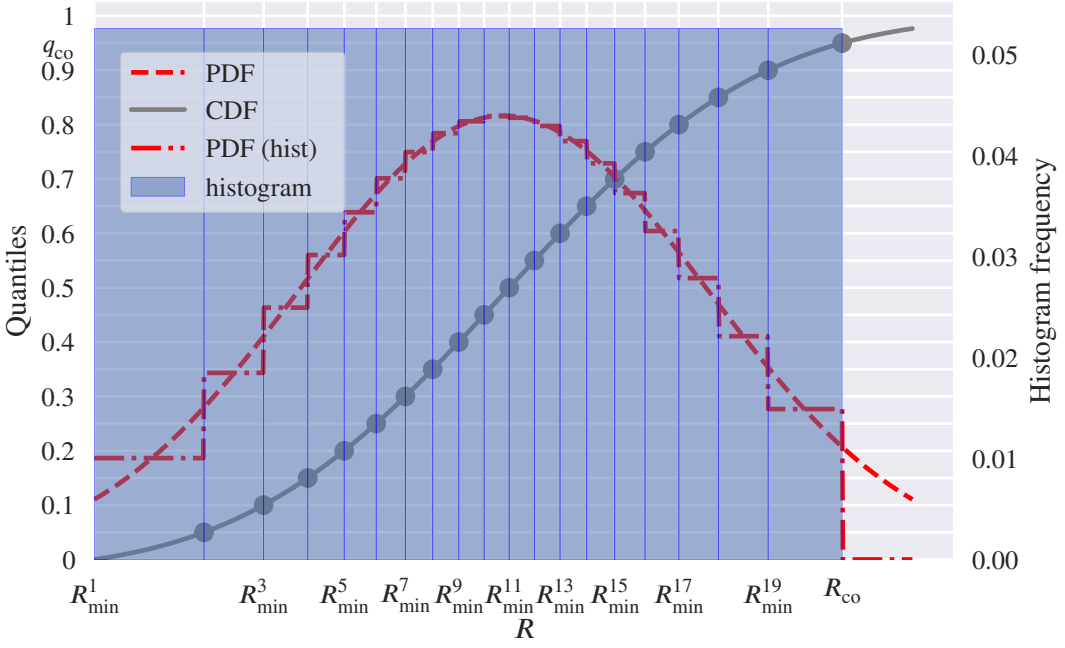


Figure 2: Schematic representation of the histogram approximation of a positive normal distribution: only 19 bins are used here, thus each quantile of the binned domain is $1/20$. R^k_{min} denotes the lower bound for the k-th bin, as defined in Eq. (19a) whereas R_{co} is the cut-off value at $q_{\text{co}} = 19/20$. The PDF associated to histogram approximation (10) is also shown (dash-dotted curve).

are used in this figure. Neglecting the last quantile is the same as assuming that $F(R) \approx 1$ and

$f(R) \approx 0$ if $R > R_{\text{co}}$, where R_{co} is the cut-off value, so that:

$$R_{\text{co}} = F^{-1}\left(q_{\text{co}}\right) \quad \text{with } q_{\text{co}} = \frac{N}{N+1}\,. \tag{20}$$

The normalized frequencies of the approximated histogram are:

$$\text{Freq}^k = \frac{\Delta F^k}{\sum_{k=1}^{N} \Delta F^k} \tag{21}$$

with:

$$\Delta F^k = F\left(R^k_{\text{max}}\right) - F\left(R^k_{\text{min}}\right) \tag{22}$$

Eqs. (19), (21) and (22) lead to:

$$\text{Freq}^k = \frac{1}{N} \tag{23}$$

Figure 2 also shows the resulting histogram, whose upper limits are given by the values associated to each quantile (denoted R^k_{max} here). The constant quantile decomposition results in bins with constant frequencies and variable widths. This avoids round-off errors near the upper-tail of the underlying distribution.

Under the approximation (18), the Wicksell transforms of the underlying distribution are defined so that:

$$\tilde{f}(r) \approx \begin{cases} \tilde{f}_{\text{hist}}\left(r \mid \text{bins}\right) & \text{if } r < R_{\text{co}} \\ 0 & \text{otherwise} \end{cases}, \tag{24a}$$

$$\tilde{F}(r) \approx \begin{cases} \tilde{F}_{\text{hist}}\left(r \mid \text{bins}\right) & \text{if } r < R_{\text{co}} \\ 1 & \text{otherwise} \end{cases}. \tag{24b}$$

2.2. Efficient sampling of a test distribution

For numerical generation of Random Variables (RVs) following a given distribution, the inverse transform sampling method is usually used (Devroye, 2006, e.g.). This method requires to compute the quantile function of the distribution. Here, computing the CDF is computationally slow because Eq. (13) cannot be analytically inverted; thus, computing the quantile function of the folded distribution ($\tilde{F}^{-1}$) requires a numerical solver. As a result, the inverse transform sampling method would be highly time consuming. Instead, the generation of RVs can be made in a more physical way. The method proposed here is to numerically emulate cross-sectioning of aligned spheres whose radii follow the underlying distribution, as illustrated in Figure 3. The aim of such

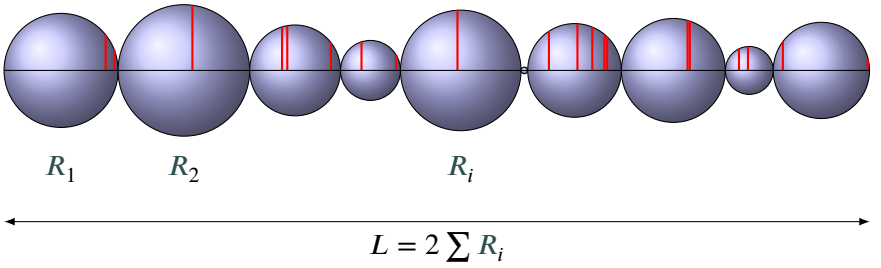


Figure 3: Schematic representation of the proposed algorithm for the generation of random variables for the corpuscle problem: a series of spheres whose radii follow the underlying distribution are packed. Then, virtual sections (red) are made at random positions, thus passing through random spheres.

an algorithm is to take into account the fact that, in a real aggregate, the larger the sphere, the more likely it can be randomly cut. Thus, the proposed algorithm for generating n RVs, assuming a given underlying distribution, can be summed up as follows:

1. generate a set of m spheres whose radii follow the underlying distribution,
2. "pack" them along a single axis, then compute the cumulative length of such axis (denoted L below),
3. pick n random values from uniform distribution between 0 and L,
4. make virtual sections of those coordinates,
5. locate which spheres are cut,

6. compute the length of each cut using the Pythagorean theorem:

$$r_i = \sqrt{R_i^2 - \left(c_i - x_i\right)^2} \tag{25}$$

where x_i denotes the coordinate of the i-th virtual section and R_i and c_i denote the radius and the center coordinate of the sphere at this location.

Trial and error has shown that best accuracy (see section 2.5) is achieved if the two following inequalities are true:

$$m \geq 10\,000 \tag{26a}$$

$$m \geq 10n \tag{26b}$$

where n is the size of the generated sample.

2.3. Python code

A Python class, named `Wicksell_transform`, has been developed as a subclass of `rv_continuous`, which is part of the `scipy.stats` module (Virtanen et al., 2020). The associated code is accessible along with the Wicksell-py project[1].

2.4. Maximum Likelihood

MLE was performed using the built-in `rv_continuous.fit` method on the transformed PDF ($\tilde{f}$). In `scipy.stats`, MLE actually consists in minimizing the log-likelihood, that is finding the estimator $\hat{\boldsymbol{\theta}}$ so that:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{Arg\,min}} \left(-\log \mathcal{L}_n(\boldsymbol{\theta})\right) \tag{27}$$

where $\log \mathcal{L}_n$ is the natural logarithm of the likelihood function, as defined in Eq. (9). It becomes:

$$-\log \mathcal{L}_n = -\sum_{i=1}^{n} \log \left[\tilde{f}\left(r_i \,\middle|\, \boldsymbol{\theta}\right)\right] \tag{28}$$

[1] https://github.com/DorianDepriester/Wicksell-py

This allows the use of standard minimizing techniques and avoids floating point underflow (multiplication of near-zero floating-point numbers quickly results in 0). Here, the default optimizer for the `fit` method in `scipy.stats` has been used. It is actually the `fmin` function, as part of the `optimizer` module. `fmin` uses the well-known downhill simplex algorithm (Nelder and Mead, 1965).

Based on the definition for log-likelihood (28), it is clear that null value for the transformed PDF should be avoided. As a consequence, the cut-off value R_{co}, as previously defined in Eq. (20), becomes:

$$R_{\text{co}} = \max\left\{F^{-1}\left(q_{\text{co}}\right), \beta r_{\max}\right\} \tag{29}$$

where $r_{\max}$ denotes the maximum value in empirical sample Ω_n and β is a coefficient slightly larger than 1. Eq. (29) ensures that $\tilde{f}(r_i) \neq 0$ for all r_i in Ω_n, according to Eq. (24a). In this work, we use $\beta = 1.001$. Figure 4 schematically illustrates the algorithm used in this paper to perform MLE on a sample Ω_n. In this figure, $\boldsymbol{\theta}_0$ denotes the so-called initial guess for the estimator. In this work, it was equal to the initial guess for the underlying distribution f. This may be a poor idea from a computational point a view (the estimator may be far from the initial guess), but this ensures that $\boldsymbol{\theta}_0$ lies in the validity domain for the parameters (e.g. $\sigma > 0$ and $E > 0$ for lognormal). Depriester and Kubler (2019b) have shown that the value for the initial guess has almost no influence on the estimator given by MDE. In order to speed-up convergence, one could also use a method of moments to evaluate the initial guess, as Kong et al. (2005) did.

2.5. Assessments of the proposed methods

In order to assess the consistency between the RVs generator (proposed in section 2.2) and the continuous CDF (computed as detailed in section 2.1), the underlying lognormal distribution has been investigated. Let n be the size of the generated sample. For a given value of n, 100 samples have been generated using random values for σ (ranging in $[0.1, 1]$) and such that the expectation (E) equals 1, that is $\mu = -\sigma^2/2$, according to Eq. (17).

In each case, the corresponding PDF has been compared against the generated sample using the

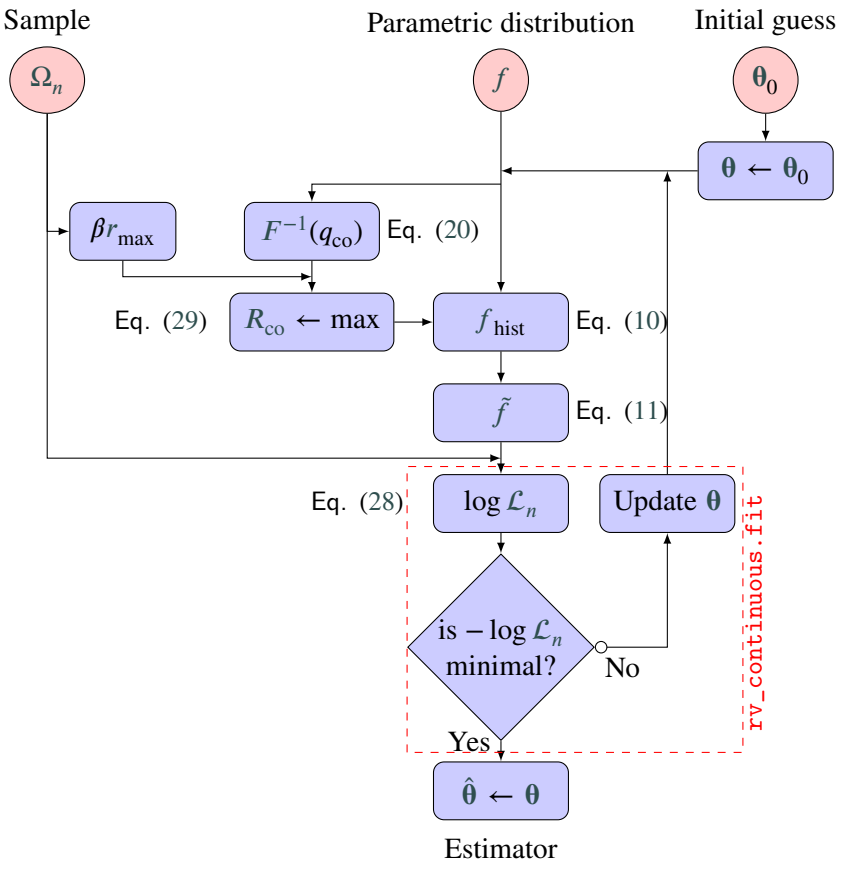


Figure 4: Schematic representation of MLE: assuming a parametric PDF f, MLE aims to find the estimator $\hat{\boldsymbol{\theta}}$ which minimizes the log-likelihood function. The red dashed block indicates the part of the algorithm which runs through the `rv_continuous.fit` method, provided by the SciPy's `stats` module (Virtanen et al., 2020); the way $\boldsymbol{\theta}$ is updated relies on the `optimizer.fmin` function. The shape of $\boldsymbol{\theta}$ depends on the choice made for the parametric function (See Table 4 in appendix for examples).

Kolmogorov–Smirnov (KS) goodness-of-fit test (Gibbons and Chakraborti, 2014, Chap. 4). This statistic is defined as follows:

$$D_n = \max_{r \in [0,\infty)} \left| \tilde{F}(r) - \tilde{F}_n \right| \tag{30}$$

where $\tilde{F}_n$ denotes the empirical CDF, as defined in Eq. (8). For each value of n, the mean and standard deviation for D_n have been computed. The results are illustrated in Figure 5. It is clear that the mean value for D_n converges toward 0 when n increases, starting from 0.25 (at $n = 10$) to 0.0114 (at $n = 5000$). The range of D_n decreases as well (see error bars). Theoretically, the null

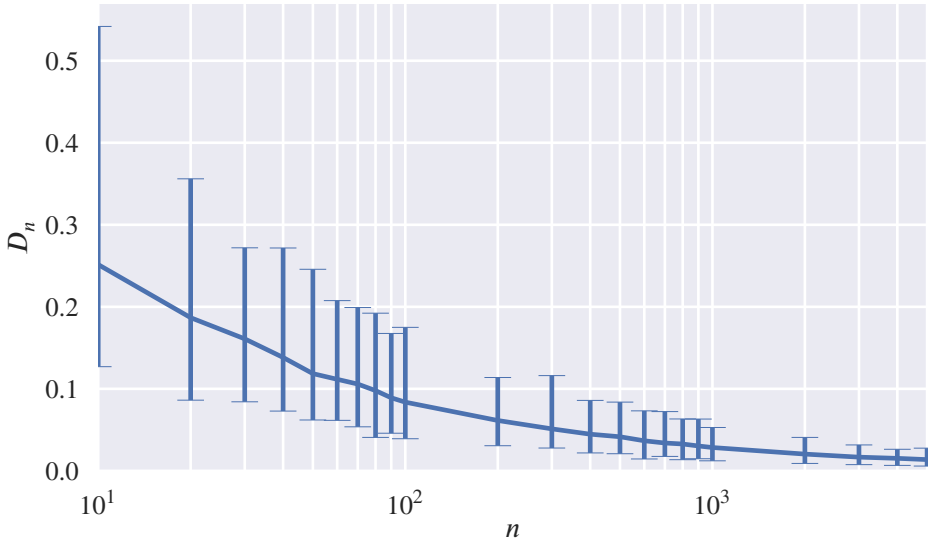


Figure 5: Evolution of the KS statistics as a function of n: the curve shows the mean values whereas the error bars indicate the extremal values.

hypothesis is rejected at significance level α if (Gibbons and Chakraborti, 2014, Chap. 4):

$$D_n > \frac{d_\alpha}{\sqrt{n}} \tag{31}$$

where d_α is a factor depending on the significance level and taken from tabular data (Smirnov, 1948). For instance, we have $d_{0.1} = 1.224$ and $d_{0.2} = 1.073$.

Let $D_{n,\alpha}$ be the empirical value for which the significance level is reached (that is the value such that a fraction α of the population of D_n is greater than this value). Figure 6 illustrates the evolution of $D_{n,\alpha}$, to be compared with the theoretical critical values given by Eq. (31) for $\alpha = 0.1$. It appears that the computed values are consistent with the asymptotic theory. This indicates that the algorithm for RVs is unbiased.

In order to investigate the accuracy of MLE, depending on the sample size, MLEs have been performed on each sample defined above. In each case, the results from MLE (in terms of estimators) have been compared with those used for RV generation. For each fit, the corresponding value for the expectation has been computed. Hereafter, $\hat{\sigma}$ and $\hat{E}$ denote the expectation and the shape

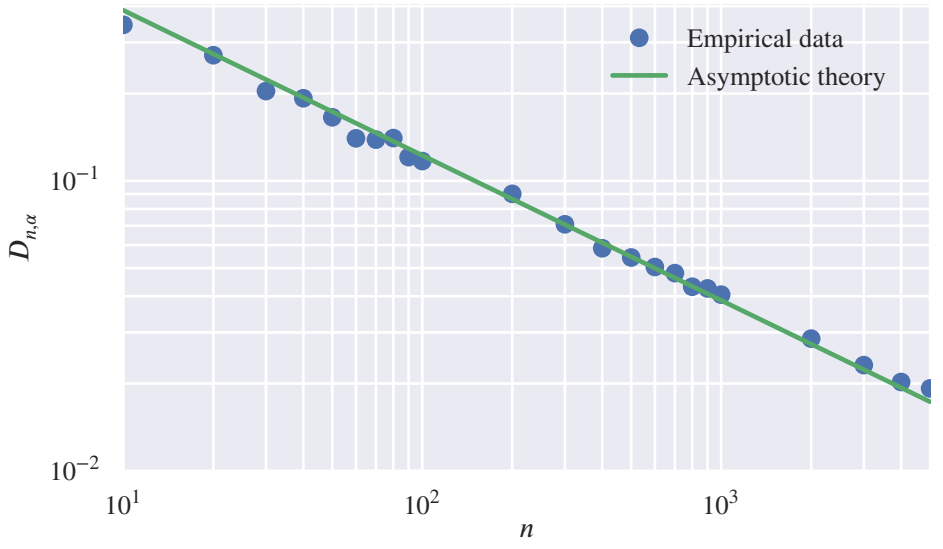


Figure 6: Critical values for D_n at significance level $\alpha = 0.1$, as function of n: comparison between empirical data and asymptotic theory (31).

parameter returned by MLE, respectively. The values of $\hat{E}$ are partially illustrated in Figure 7. It appears that the resulting expectation tends toward unity when n is large or when σ is small. Conversely, it is clear that the larger σ, the larger the dispersion (that is the weaker certainty). In order to estimate the uncertainty when fitting, the envelope of scatter data has been estimated in each case (see red lines in Figure 7). Those envelopes bound 95% of the experimental values such that:

$$1 - \bar{\epsilon}_E \leq \hat{E} \leq 1 + \bar{\epsilon}_E \tag{32}$$

with:

$$\bar{\epsilon}_E = \frac{3.003\sigma}{\sqrt{n}}\,. \tag{33}$$

The square root of n in the denominator of $\bar{\epsilon}_E$ resembles the expression of standard error when computing a sample mean (Altman and Bland, 2005) although it is a completely different approach here (we try to estimate the error on the expectation of the underlying distribution from a folded sample); still, the expression proposed in Eq. (33) for $\bar{\epsilon}_E$ appears to be a good approximation, according to Figure 7, and this square root relationship is similar to that reported by Farr et al.

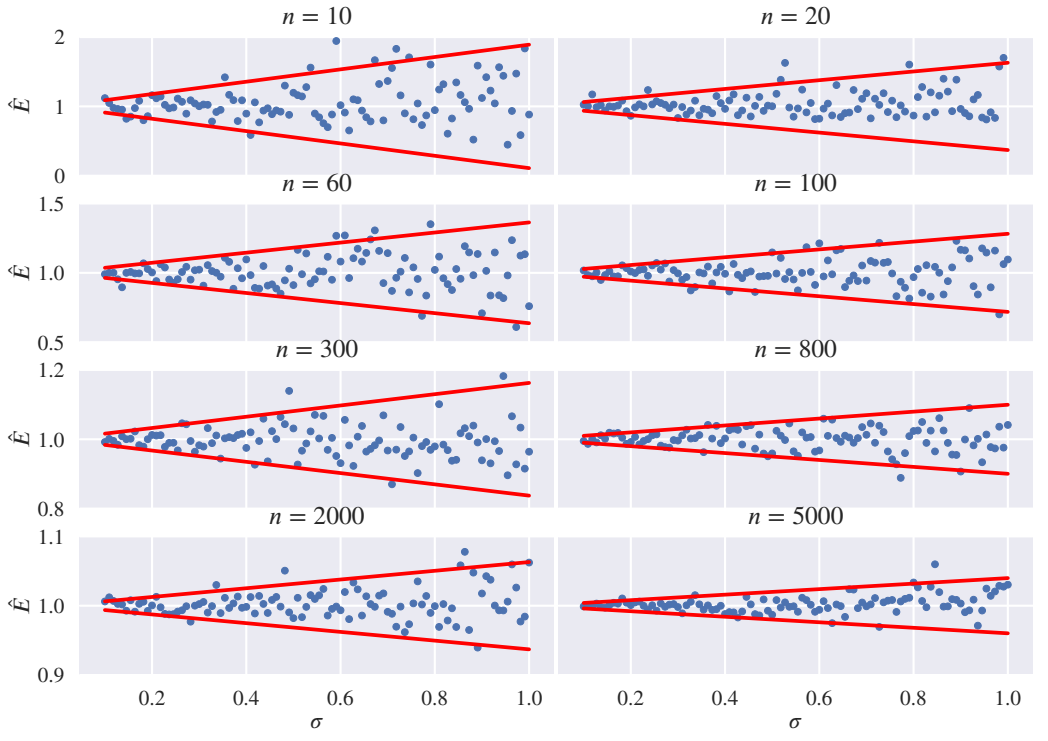


Figure 7: Results from fitting, depending on the sample size (n) and the shape parameter (σ): expectation for the lognormal distribution (E). The envelopes (red lines) are approximations, with respect to Eq. (33).

(2017). As a reminder, $E = 1$ here; thus, a scale factor can be used to generalize Eq. (32) for any value of E, leading to:

$$1 - \frac{3.003\sigma}{\sqrt{n}} \leq \frac{\hat{E}}{E} \leq 1 + \frac{3.003\sigma}{\sqrt{n}} \tag{34}$$

We now define the relative error related to shape parameter as follows:

$$\epsilon_\sigma = \frac{\hat{\sigma} - \sigma}{\sigma} \tag{35}$$

Figure 8 illustrates the evolution of ϵ_σ as functions of n and σ. It appears that, except for $n = 10$, the dispersion is independent of σ. Thus, the envelopes (red lines in Figure 8) are defined as function

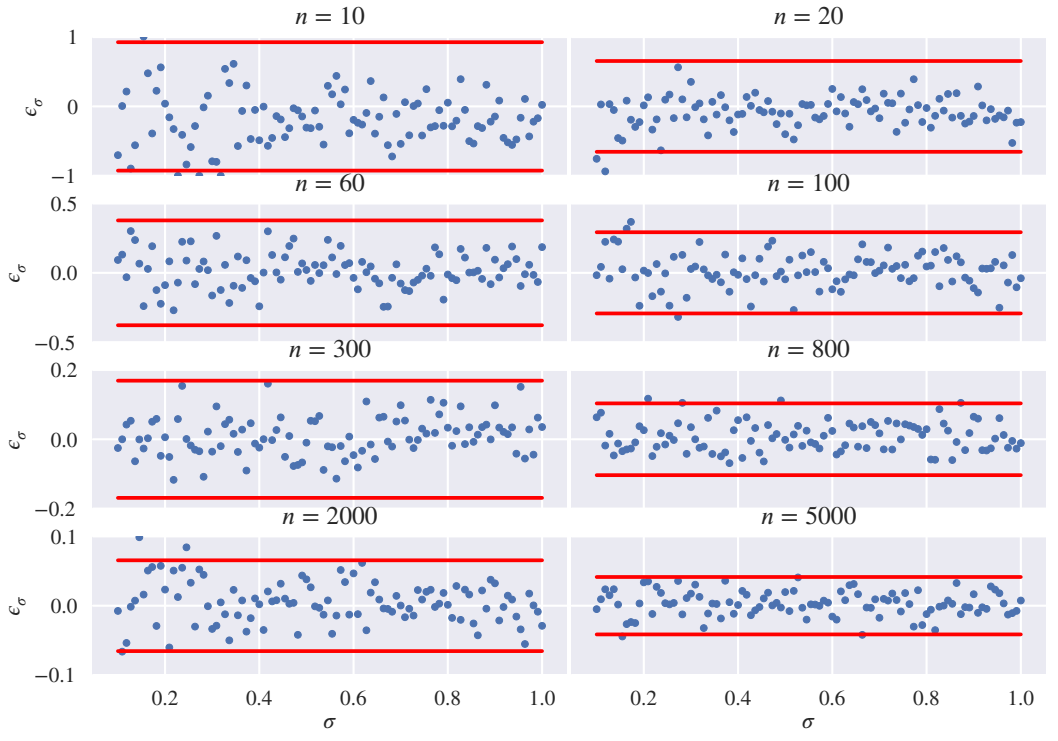


Figure 8: Results from fitting, depending on the sample size (n) and the shape parameter (σ): relative errors on the evaluation of σ (ϵ_σ). The envelopes (red curves) are approximations, with respect to Eq. (37).

of n only, that is:

$$-\bar{\epsilon}_\sigma \leq \epsilon_\sigma \leq \bar{\epsilon}_\sigma \tag{36}$$

with:

$$\bar{\epsilon}_\sigma = \frac{2.402}{\sqrt{n}} \tag{37}$$

Again, the latter envelopes are defined such that they bound 95% of experimental points.

Figure 9 illustrates the evolutions of $\bar{\epsilon}_E$ and $\bar{\epsilon}_\sigma$, depending on the sample size and the shape parameter for lognormal distribution (contour plots). If one wants to ensure that both E and σ are estimated within a given confidence interval, the maximum value between $\bar{\epsilon}_E$ and $\bar{\epsilon}_\sigma$ should be considered. Thus, Figure 9 also illustrates such value (see coloured regions). It appears that 5% un-

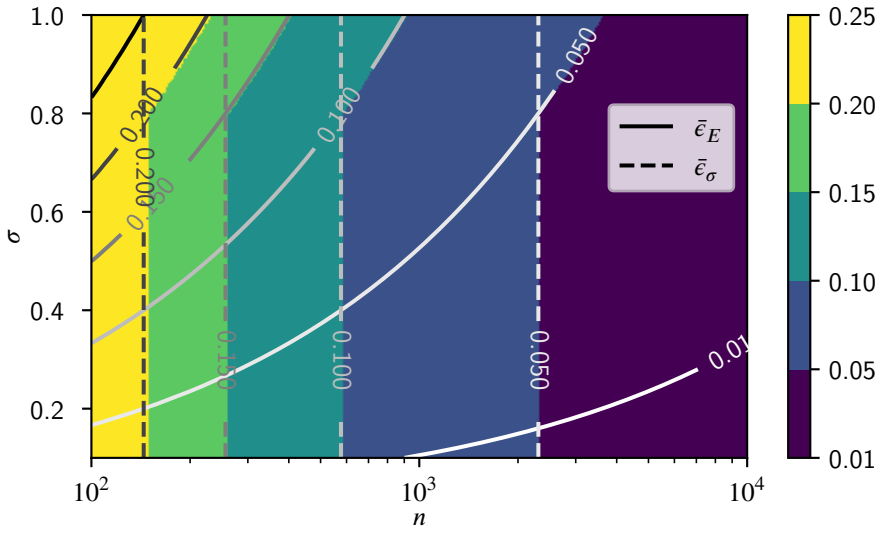


Figure 9: Values of relative errors on estimators, depending on the sample size (n) and shape parameter (σ) for lognormal distribution. Contour plots illustrate the values of $\bar{\epsilon}_E$ and $\bar{\epsilon}_\sigma$, whereas the background colour indicates the maximum value between $\bar{\epsilon}_E$ and $\bar{\epsilon}_\sigma$.

certainty is hard to ensure, for it requires at least $n = 2310$. But the 10% uncertainty can be achieved with only $n = 580$ if $\sigma < 0.8$. Indeed, it is clear that the required number of values increases with σ if $\sigma > 0.8$. For the 15% uncertainty, only 260 values are required if $\sigma < 0.8$; 400 otherwise. Furthermore, according to the literature, it appears that σ is usually smaller than 0.8 in realistic applications (see e.g. Table 1). As a result, 580 values are enough to reach the 10% uncertainty in most cases. This number is far smaller than that usually reported for other stereology techniques like the Saltykov method or MDE, which both require at least 1000 values (Lopez-Sanchez and Llana-Fúnez, 2016; Depriester and Kubler, 2019b). The accuracy from method of moments appears to be same order of magnitude (Farr et al., 2017), but they cannot be extensively compared with the present results, for Farr et al. (2017) mainly focus on the moments of the distribution, not on its parameters. However, those authors have also reported an increasing error with increasing value for σ.

In practice, when using the MLE, one wants to estimate the bounds on the real value of σ, based on the estimated one (denoted $\hat{\sigma}$ above). Thus, the definition of ϵ_σ in Eq. (35), combined

with Eqs. (36) and (37) leads to:

$$\sigma_{\min} \leq \sigma \leq \sigma_{\max} \tag{38}$$

with

$$\sigma_{\min} = \frac{\hat{\sigma}}{1 + \frac{2.402}{\sqrt{n}}}\,, \tag{39a}$$

$$\sigma_{\max} = \frac{\hat{\sigma}}{1 - \frac{2.402}{\sqrt{n}}}\,. \tag{39b}$$

Eqs. (34) and (38) lead to:

$$E_{\min} \leq E \leq E_{\max} \tag{40}$$

with

$$E_{\min} = \frac{\hat{E}}{1 + \frac{3.003\sigma_{\max}}{\sqrt{n}}}\,, \tag{41a}$$

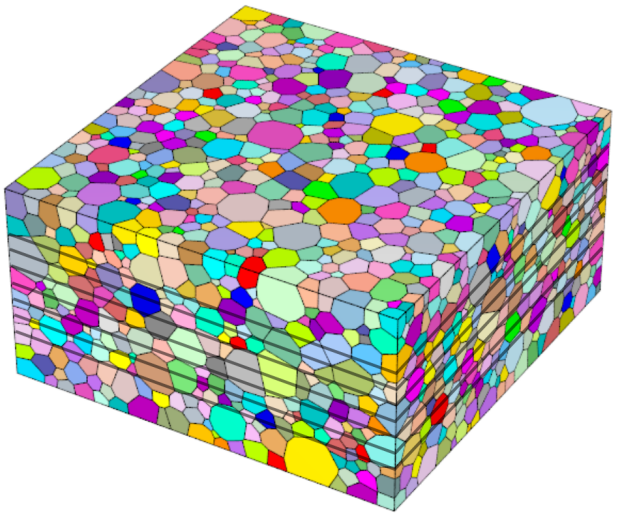
$$E_{\max} = \frac{\hat{E}}{1 - \frac{3.003\sigma_{\max}}{\sqrt{n}}}\,. \tag{41b}$$

Finally, the uncertainty on μ can be estimated as follows:

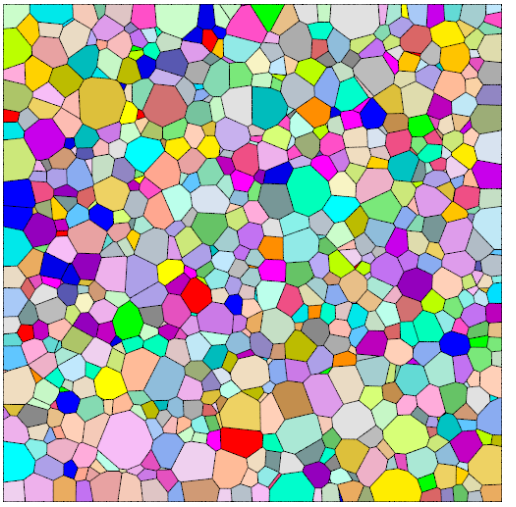
$$\log\left(E_{\min}\right) - \frac{\sigma_{\max}^2}{2} \leq \mu \leq \log\left(E_{\max}\right) - \frac{\sigma_{\min}^2}{2}\,. \tag{42}$$

2.6. Validation on synthetic polycrystals

This section investigates the effect of applying MLE on grains which are convex polyhedrons, instead of spheres. The idea is to apply the MLE on a microstructure where both 2D and 3D distributions are known a priori. To do so, synthetic polycrystalline aggregates were generated using the Neper software (Quey et al., 2011), as illustrated in Figure 10a. Those aggregates were generated so that the equivalent grain radii, as defined in Eq. (1), followed a lognormal distribution



(a) 3D tessellation. The black lines illustrate how the volume is sliced.



(b) 2D slice.

Figure 10: Example of synthetic material generated with Neper: a volume is first populated with polyhedrons with equivalent radii following lognormal distribution with $E = 1$ and $\sigma = 0.3$ (a); then this geometry is sliced (b). Each grain is randomly coloured.

with $E = 1$ and σ ranging from 0.1 to 0.9. The populated volume was of size $40 \times 40 \times 20$[2]. In each case, the polycrystal was sliced 7 times at different vertical positions, as schematically illustrated in Figure 10a. An example of such a slice is given in Figure 10b. Then, the equivalent size of apparent grains were computed with respect to Eq. (2). In order to avoid artefacts due to truncated grains at the domain boundaries, all apparent grains at the border of each slice were removed from the dataset. As a result, for a given value of σ, from 2579 to 5356 unique values of apparent radii were recorded.

For a given sample size n and a given value for σ, n values were randomly picked from the corresponding dataset; then MLE was performed on this sub-sample. This operation was repeated 50 times in order to estimate the variance in each case. The results are illustrated in Figure 11. It is clear that the mean value for the expectation (solid blue lines) is always underestimated whereas that for σ (dashed blue lines) is always overestimated. Details about the evolutions of those mean values, depending on σ, are given in Figure 12. From those scatter plots, polynomial regressions

[2]For $\sigma = 0.7$ and 0.9, the domain was of size $60 \times 60 \times 30$ in order to ensure that the number of grains was large enough (> 3000).

can be made (see continuous lines in Figure 12), giving:

$$\frac{\hat{E}}{E} \approx \ 0.402\sigma^2 - 0.589\sigma + 1.019\,, \tag{43a}$$

$$\hat{\sigma} \approx -0.430\sigma^2 + 1.361\sigma - 0.014\,. \tag{43b}$$

Those equations can be used to correct the bias introduced by assuming grains as spheres instead of polyhedrons in a polycrystalline aggregate.

Figure 11 also shows the standard deviations (std), made for each sample size n, as error bars. Each bar represents $\pm$2std (so that it approximately covers 95% of the whole results). It appears that the standard deviations decrease when n increases, and tend toward 0 when $n \to \infty$. As a comparison, the envelopes given by Eqs. (34) and (36) are also illustrated (shaded regions); it is worth remembering that those envelopes were defined so that they also bound 95% of the populations (see section 2.5). Hence, it appears that, for $n > 100$, those envelopes are in good agreement with the actual spreads given by MLE on polycrystalline aggregates; thus the confidence intervals given in Eqs. (34) and (36) appear to be still valid, even if the grains are not perfect spheres. Conversely, Figure 11 illustrates large discrepancies between the envelopes and the actual spreads when $n \leq 100$.

3. Applications

MLE was performed on two datasets introduced in previous works:

1. Grains of uranium dioxide (UO_2), manufactured by sintering, and imaged by EBSD by Soulacroix (2014);
2. Quartz grains in mylonite (Lopez-Sanchez and Llana-Fúnez, 2016), imaged by OM then manually segmented[3].

Because of the finite resolution in imaging techniques, grains with size smaller than 1 pixel cannot be taken into account, resulting in truncated empirical distribution. One usually neglects grains

[3]The dataset corresponds to map E in (Lopez-Sanchez and Llana-Fúnez, 2016); see Suppl. materials in ref.

Table 2
Results from MLE on UO_2 and quartz, assuming lognormal distributions. Brackets give the confidence interval (for 95% confidence level) for the estimators.

		UO_2	Quartz in mylonite
Sample	n	4264	3582
	$r_{\min}$	1.128 µm	0.615 µm
Estimator	$\hat{\sigma}$	0.400 [0.386, 0.415]	0.506 [0.486, 0.527]
	$\hat{E}$	5.038 [4.943, 5.136]	17.77 [17.31, 18.25]
	$\hat{\mu}$	1.537 [1.512, 1.562]	2.750 [2.713, 2.786]
KS test	D_n	0.0344	0.0236
	$D_{n,0.1}$	0.0187	0.0204
	$D_{n,0.2}$	0.0164	0.0179

smaller than 6 to 10 pixels since they can induce artefact effects in the apparent distribution, as pointed out by Lopez-Sanchez (2020). Still, a threshold value of 1 pixel has been used here in order to ensure consistency with the procedure used in (Depriester and Kubler, 2019b). Let $r_{\min}$ be the left-truncating value. The truncated PDF and CDF are respectively:

$$\tilde{f}^*(r) = \frac{\tilde{f}(r)}{1 - \tilde{F}\left(r_{\min}\right)}, \tag{44a}$$

$$\tilde{F}^*(r) = \frac{\tilde{F}(r) - \tilde{F}\left(r_{\min}\right)}{1 - \tilde{F}\left(r_{\min}\right)}. \tag{44b}$$

The values for $r_{\min}$ and the sample sizes (n) are given in Table 2.

At first, a lognormal has been assumed as the underlying distribution for both datasets. The results from MLE and the corresponding confidence intervals are given in Table 2. Thanks to the large number of grains, the confidence intervals are very small here.

Based on the estimators given in Table 2, the KS tests have been performed against the experimental data, as presented in this table. Critical values for KS statistics at significance level α ($D_{n,\alpha}$), estimated from Eq. (31), are given as well. Thus, the null hypothesis should be rejected at significance level $\alpha = 0.1$ for both datasets. Nevertheless, considering the low value for D_n for quartz grains, assuming a lognormal distribution in this case appears to be a reasonable approximation.

Table 3
Results from MLEs performed on the example dataset, and corresponding KS statistics. The values in parentheses give the results from MDE (Depriester and Kubler, 2019b), for comparison.

Underlying distribution	Parameters			D_n
Positive normal	Mode	(μ):	3.547 (3.876)	0.0158
	Shape	(σ):	3.006 (2.816)	
Weibull	Scale	(k):	2.017 (2.106)	0.0176
	Shape	(λ):	5.103 (5.203)	
Rayleigh	Mode	(σ):	3.587 (3.612)	0.0176
Gamma	Scale	(θ):	1.001 (1.026)	0.0225
	Shape	(k):	4.822 (4.724)	
Lognormal	Scale	(μ):	1.537 (1.519)	0.0344
	Shape	(σ):	0.400 (0.428)	

As a comparison, Lopez-Sanchez and Llana-Fúnez (2016) found $\sigma = 0.57$, $E = 20.9$ and $\mu = 2.87$ on quartz grains using the two-step method with 12 classes. Thus, it seems that both methods lead to similar estimators, although MLE results in a slightly sharper distribution (lower value for σ) and slightly lower value for the expectation. Applying the correction rules for polyhedrons, as defined in Eqs. (43), we get $\sigma = 0.341$, $E = 5.823$ and $\mu = 1.704$ for UO_2, and $\sigma = 0.444$, $E = 21.24$ and $\mu = 2.957$ for quartz grains.

In (Depriester and Kubler, 2019b), other distributions were investigated on UO_2, namely: positive normal, Weibull, Rayleigh and Gamma (see Table 4 in appendix for details). Thus, MLE was performed assuming the aforementioned distributions. In each case, the KS goodness-of-fit test was computed against the empirical data. The results, in terms of estimators and KS statistics, are given in Table 3. According to the values of D_n, it appears that a positive normal distribution is the best candidate for the underlying distribution, whereas lognormal is the worst. This is consistent with the results from MDE (Depriester and Kubler, 2019b). Considering the values of D_n, the null hypothesis related to lognormal and Gamma should be rejected at significance level $\alpha = 0.1$, whereas those related to positive normal, Weibull and Rayleigh should not. At significance level $\alpha = 0.2$, all the null hypotheses should be rejected, except that related to the positive

normal distribution (see the critical values in Table 2).

As a comparison, the results obtained by MDE on the same distributions are given in parentheses in Table 3. Thus, it is clear that both MLE and MDE lead to very similar estimators. As a conclusion, MLE and MDE are consistent with each other.

In order to assess the estimators given in Table 3, one can plot the transformed PDFs on-top of the empirical histogram, as illustrated in Figure 13. It appears that the positive normal distribution fits well with these empirical data, specially at $r > 5\,\mu\text{m}$, but it slightly underestimates the modal value (location of the maximum frequency). In contrast, the lognormal distribution fails to fit the empirical data, specially for $r < 5\,\mu\text{m}$.

For comparative purposes, the corresponding PDFs are plotted in Figure 14. It is clear that they are quite different from each other, specially for lower radii (less than 6 µm), except the Weibull and Rayleigh distributions, which appear superimposed in Figure 14. This is because the Weibull's scale parameter (denoted k in Table 3) is very close to 2 (the Rayleigh distribution is a special case of the Weibull distribution with $k = 2$). Conversely, all the PDFs are barely distinguishable on the right tails (radii larger than 7 µm). It is worth mentioning that the same applies in the results presented in Depriester and Kubler (2019b) (e.g. see Fig. 5 in reference). Those tails correspond to 15% of the population. Thus, it seems that, regardless of the underlying parametric distribution being used, both MDE and MLE actually mainly use the larger values to unfold the distribution, rather than on the whole population. In other words, more "weight" is given to larger radii. This is because changing the proportion of large grains would change the whole transformed (apparent) distribution, whereas changing the proportion of small grains would only affect the left tail. As a conclusion, it is not possible to determine which method is superior based on these analyses. Nevertheless, it appears that both MDE and MLE are equally good techniques, except MLE seems to require fewer data for a given accuracy. In addition, it seems that the unfolded distributions from both the methods are more accurate with respect to larger radii than to the smaller ones.

In Figure 13, the positive normal distribution evidences somehow large frequencies at lower radii, especially at 0 value ($f_{\mathcal{N}^+}(0) = 0.075$). This result does not seem realistic considering the

material. As a reminder, the latter (UO_2) was made by sintering, thus the grain size distribution is mainly inherited from that of the powder used for compaction. Hence, one should expect a null frequency for $R = 0$. As a result, the Weibull distribution seems to be the best candidate as a compromise between realisticity (i.e. it ensures that $f_W(0) = 0$) and goodness-of-fit ($D_n = 0.0176$ for the Weibull distribution).

4. Conclusion

The equations presented in a former work (Depriester and Kubler, 2019b) have been used to evaluate the efficiency of MLE for solving the corpuscle problem. The special case of a lognormal distribution for the underlying distribution has been studied in detail in this work and interval errors (for 95% confidence level) have been evaluated. Thus, this paper provides a way to estimate the interval errors when unfolding a distribution, assuming a 3D lognormal distribution. The ability of the proposed method to unfold the grain size distribution, when grains are polyhedrons instead of perfect spheres is also investigated. The following conclusions have been reached:

- if the underlying distribution is lognormal, 10% uncertainty on the estimators can be achieved with only 580 unique values as empirical data; this number is almost half that usually required for the Saltykov method or MDE;
- both MLE and MDE applied on real-world data (grain size distributions in UO_2 and mylonite) lead to very similar results for estimators; hence, considering that larger samples are required for MDE, it appears that MLE is more efficient than MDE for solving the corpuscle problem;
- if the grains are polyhedrons with sizes following a lognormal distribution, a set of correction laws for the distribution parameters estimated by MLE has been established;

If the underlying distribution is not known a priori, MLE can be used to test which distribution, among several ones, leads to the highest goodness-of-fit. Since this paper mainly focuses on the 3D lognormal distribution only, the efficiency of MLE on other distributions (e.g. positive normal, Weibull, Rayleigh etc.) may be studied in a future work.

Table 4
Parametric density functions used in this paper: related parameters and expression of the PDFs.

Name	Parameter(s)	PDF
Positive normal	$\boldsymbol{\theta} = (\mu, \sigma)$	$f_{\mathcal{N}^+}(R \mid \mu, \sigma) = \frac{1}{\sigma(1-\Phi_0)}\phi\left(\frac{R-\mu}{\sigma}\right)$ with: $\phi(x) = \frac{1}{\sqrt{2\pi}}\exp\left(\frac{-x^2}{2}\right)$ and $\Phi_0 = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{-\mu}{\sigma\sqrt{2}}\right)\right]$
Gamma	$\boldsymbol{\theta} = (\theta, k)$	$f_\gamma(R \mid k, \theta) = \frac{1}{\Gamma(k)\theta^k} R^{k-1}\exp\left(-\frac{R}{\theta}\right)$ with: $\Gamma(k) = \int_0^\infty x^{k-1}\exp(-x)\,\mathrm{d}x$
Weibull	$\boldsymbol{\theta} = (k, \lambda)$	$f_{\mathrm{W}}(R \mid \lambda, k) = \frac{k}{\lambda}\left(\frac{R}{\lambda}\right)^{k-1}\exp\left(-\left(\frac{R}{\lambda}\right)^k\right)$
Rayleigh	$\boldsymbol{\theta} = \sigma$	$f_{\mathrm{R}}(R \mid \mu) = \frac{R}{\sigma^2}\exp\left(\frac{-R^2}{2\sigma^2}\right)$

A. Parametric probability functions used in this paper

All the parametric density functions used in this paper are summarized in Table 4.

CRediT authorship contribution statement

Dorian Depriester: Conceptualization of this study, Software, Validation, Formal analysis, Investigation, Writing - Original Draft. **Régis Kubler:** Writing - Review & Editing, Supervision.

References

Al-Harthi, A., Al-Amri, R., Shehata, W., 1999. The porosity and engineering properties of vesicular basalt in Saudi Arabia. Engineering Geology 54, 313–320. URL: https://www.sciencedirect.com/science/article/pii/S0013795299000502, doi:https://doi.org/10.1016/S0013-7952(99)00050-2.

Altman, D.G., Bland, J.M., 2005. Standard deviations and standard errors. Bmj 331, 903.

Bellido, G.G., Scanlon, M.G., Page, J.H., Hallgrimsson, B., 2006. The bubble size distribution in wheat flour dough. Food Research International 39, 1058 – 1066. URL: http://www.sciencedirect.com/science/article/pii/S0963996906001141, doi:https://doi.org/10.1016/j.foodres.2006.07.020. physical Properties VI.

Berger, A., Herwegh, M., Schwarz, J.O., Putlitz, B., 2011. Quantitative analysis of crystal/grain sizes and their distributions in 2d and 3d. Journal of Structural Geology 33, 1751 – 1763. URL: http://www.sciencedirect.com/science/article/pii/S0191814111001179, doi:https://doi.org/10.1016/j.jsg.2011.07.002.

Bose, N., Dutta, D., Mukherjee, S., 2018. Role of grain-size in phyllonitisation: insights from mineralogy, microstructures, strain analyses and numerical modeling. Journal of Structural Geology 112, 39–52.

412 Cashman, K.V., Marsh, B.D., 1988. Crystal size distribution (csd) in rocks and the kinetics and dynamics of crystallization ii: Makaopuhi lava lake. Contributions to Mineralogy and Petrology 99, 292–305.

414 Chan, K.C.G., Qin, J., 2016. Nonparametric maximum likelihood estimation for the multisample Wicksell corpuscle problem. Biometrika 103, 273–286. URL: https://doi.org/10.1093/biomet/asw011, doi:10.1093/biomet/asw011,
416 arXiv:https://academic.oup.com/biomet/article-pdf/103/2/273/23590465/asw011.pdf.

Chiu, S.N., Stoyan, D., Kendall, W.S., Mecke, J., 2013. Stochastic geometry and its applications. John Wiley & Sons.

418 Conrad, H., Swintowski, M., Mannan, S., 1985. Effect of cold work on recrystallization behavior and grain size distribution in titanium. Metallurgical transactions A 16, 703–708.

420 Czaplińska, D., Piazolo, S., Zibra, I., 2015. The influence of phase and grain size distribution on the dynamics of strain localization in polymineralic rocks. Journal of Structural Geology 72, 15–32.

422 Czaplińska, D., Piazolo, S., Zibra, I., 2015. The influence of phase and grain size distribution on the dynamics of strain localization in polymineralic rocks. Journal of Structural Geology 72, 15–32. URL: https://www.sciencedirect.com/science/article/pii/
424 S0191814115000024, doi:https://doi.org/10.1016/j.jsg.2015.01.001.

Depriester, D., Kubler, R., 2019a. Radical Voronoï tessellation from random pack of polydisperse spheres: Prediction of the cells' size distri-
426 bution. Computer-Aided Design 107, 37 – 49. URL: http://www.sciencedirect.com/science/article/pii/S0010448518300083, doi:https://doi.org/10.1016/j.cad.2018.09.001.

428 Depriester, D., Kubler, R., 2019b. Resolution of the Wicksell's equation by Minimum Distance Estimation. Image Analysis & Stereology 38, 213–226. URL: https://www.ias-iss.org/ojs/IAS/article/view/2133, doi:10.5566/ias.2133.

430 Devroye, L., 2006. Chapter 4 nonuniform random variate generation, in: Henderson, S.G., Nelson, B.L. (Eds.), Simulation. Elsevier. volume 13 of *Handbooks in Operations Research and Management Science*, pp. 83 – 121. URL: http://www.sciencedirect.com/science/article/
432 pii/S0927050706130042, doi:https://doi.org/10.1016/S0927-0507(06)13004-2.

Farr, R.S., Honour, V.C., Holness, M.B., 2017. Mean grain diameters from thin sections: matching the average to the problem. Mineralogical
434 Magazine 81, 515–530. doi:10.1180/minmag.2016.080.107.

Gibbons, J.D., Chakraborti, S., 2014. Nonparametric Statistical Inference: Revised and Expanded. CRC press.

436 Heilbronner, R., Bruhn, D., 1998. The influence of three-dimensional grain size distributions on the rheology of polyphase rocks. Journal of Structural Geology 20, 695 – 705. URL: http://www.sciencedirect.com/science/article/pii/S0191814198000108, doi:https:
438 //doi.org/10.1016/S0191-8141(98)00010-8.

Holly, J., Hampton, D., Thomas, M.D., 1993. Modelling relationships between permeability and cement paste pore microstructures. Cement and
440 Concrete Research 23, 1317 – 1330. URL: http://www.sciencedirect.com/science/article/pii/000888469390069L, doi:https:
//doi.org/10.1016/0008-8846(93)90069-L.

442 Keiding, N., Jensen, S.T., 1972. Maximum likelihood estimation of the size distribution of liver cell nuclei from the observed distribution in a plane section. Biometrics , 813–829.

444 King, M., Zimmerman, R., Corwin, R., 1988. Seismic and electrical properties of unconsolidated permafrost. Geophysical Prospecting 36, 349–364.

Kong, M., Bhattacharya, R.N., James, C., Basu, A., 2005. A statistical approach to estimate the 3d size distribution of spheres from 2d size
446 distributions. Geological Society of America Bulletin 117, 244–249. doi:10.1130/B25000.1.

Kosugi, K., Hopmans, J., 1998. Scaling water retention curves for soils with lognormal pore-size distribution. Soil Science Society of America
448 Journal 62, 1496–1505.

Kretz, R., 1993. A garnet population in Yellowknife schist, Canada. Journal of Metamorphic Geology 11, 101–120. URL:

https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1525-1314.1993.tb00134.x, doi:https://doi.org/10.1111/j.1525-1314.1993.tb00134.x, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1525-1314.1993.tb00134.x.

Lopez-Sanchez, M., Llana-Funez, S., 2015. An evaluation of different measures of dynamically recrystallized grain size for paleopiezometry or paleowattometry studies. Solid Earth 6, 475–495. doi:10.5194/se-6-475-2015.

Lopez-Sanchez, M.A., 2020. Which average, how many grains, and how to estimate robust confidence intervals in unimodal grain size populations. Journal of Structural Geology 135, 104042. URL: https://www.sciencedirect.com/science/article/pii/S0191814119305358, doi:https://doi.org/10.1016/j.jsg.2020.104042.

Lopez-Sanchez, M.A., Llana-Fúnez, S., 2016. An extension of the Saltykov method to quantify 3d grain size distributions in mylonites. Journal of Structural Geology 93, 149 – 161. URL: http://www.sciencedirect.com/science/article/pii/S0191814116301778, doi:https://doi.org/10.1016/j.jsg.2016.10.008.

Nelder, J.A., Mead, R., 1965. A Simplex Method for Function Minimization. The Computer Journal 7, 308–313. URL: https://doi.org/10.1093/comjnl/7.4.308, doi:10.1093/comjnl/7.4.308, arXiv:https://academic.oup.com/comjnl/article-pdf/7/4/308/1013182/7-4-308.pdf.

Quey, R., Dawson, P., Barbe, F., 2011. Large-scale 3d random polycrystals for the finite element method: Generation, meshing and remeshing. Computer Methods in Applied Mechanics and Engineering 200, 1729–1745. URL: https://www.sciencedirect.com/science/article/pii/S004578251100003X, doi:https://doi.org/10.1016/j.cma.2011.01.002.

Rhines, F., Patterson, B., 1982. Effect of the degree of prior cold work on the grain volume distribution and the rate of grain growth of recrystallized aluminum. Metall Trans A 13, 985–993.

Sahagian, D., Proussevitch, A., 1998. 3d particle size distributions from 2d observations: Stereology for natural applications. Journal of Volcanology and Geothermal Research 84, 173–196. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0032455529&doi=10.1016%2fS0377-0273%2898%2900043-2&partnerID=40&md5=2a8a468a982b822dfa7d53f09acef167, doi:10.1016/S0377-0273(98)00043-2. cited By 233.

Saltykov, S., 1967. The determination of the size distribution of particles in an opaque material from a measurement of the size distribution of their sections, in: Elias, H. (Ed.), Stereology, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 163–173.

Schmalholz, S.M., Duretz, T., 2017. Impact of grain size evolution on necking in calcite layers deforming by combined diffusion and dislocation creep. Journal of Structural Geology 103, 37–56.

Smirnov, N., 1948. Table for estimating the goodness of fit of empirical distributions. The Annals of Mathematical Statistics 19, 279–281. URL: http://www.jstor.org/stable/2236278.

Soulacroix, J., 2014. Approche micromécanique du comportement du combustible dioxyde d'uranium. Ph.D. thesis. Paris, ENSAM. URL: http://www.theses.fr/2014ENAM0032.

Tucker, J.C., Chan, L.H., Rohrer, G.S., Groeber, M.A., Rollett, A.D., 2012. Comparison of grain size distributions in a ni-based superalloy in three and two dimensions using the saltykov method. Scripta Materialia 66, 554 – 557. URL: http://www.sciencedirect.com/science/article/pii/S1359646212000073, doi:https://doi.org/10.1016/j.scriptamat.2012.01.001.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors, 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods 17, 261–272. doi:10.1038/s41592-019-0686-2.

488 Wicksell, S.D., 1925. The corpuscle problem: A mathematical study of a biometric problem. Biometrika 17, 84–99. URL: http://www.jstor.org/stable/2332027.

490 Yamamoto, K., Wakita, J.i., 2016. Analysis of a stochastic model for bacterial growth and the lognormality of the cell-size distribution. Journal of the Physical Society of Japan 85, 074004. doi:10.7566/JPSJ.85.074004.

492 Zöllner, D., Streitenberger, P., 2006. Three-dimensional normal grain growth: Monte carlo potts model simulation and analytical mean field theory. Scripta materialia 54, 1697–1702.

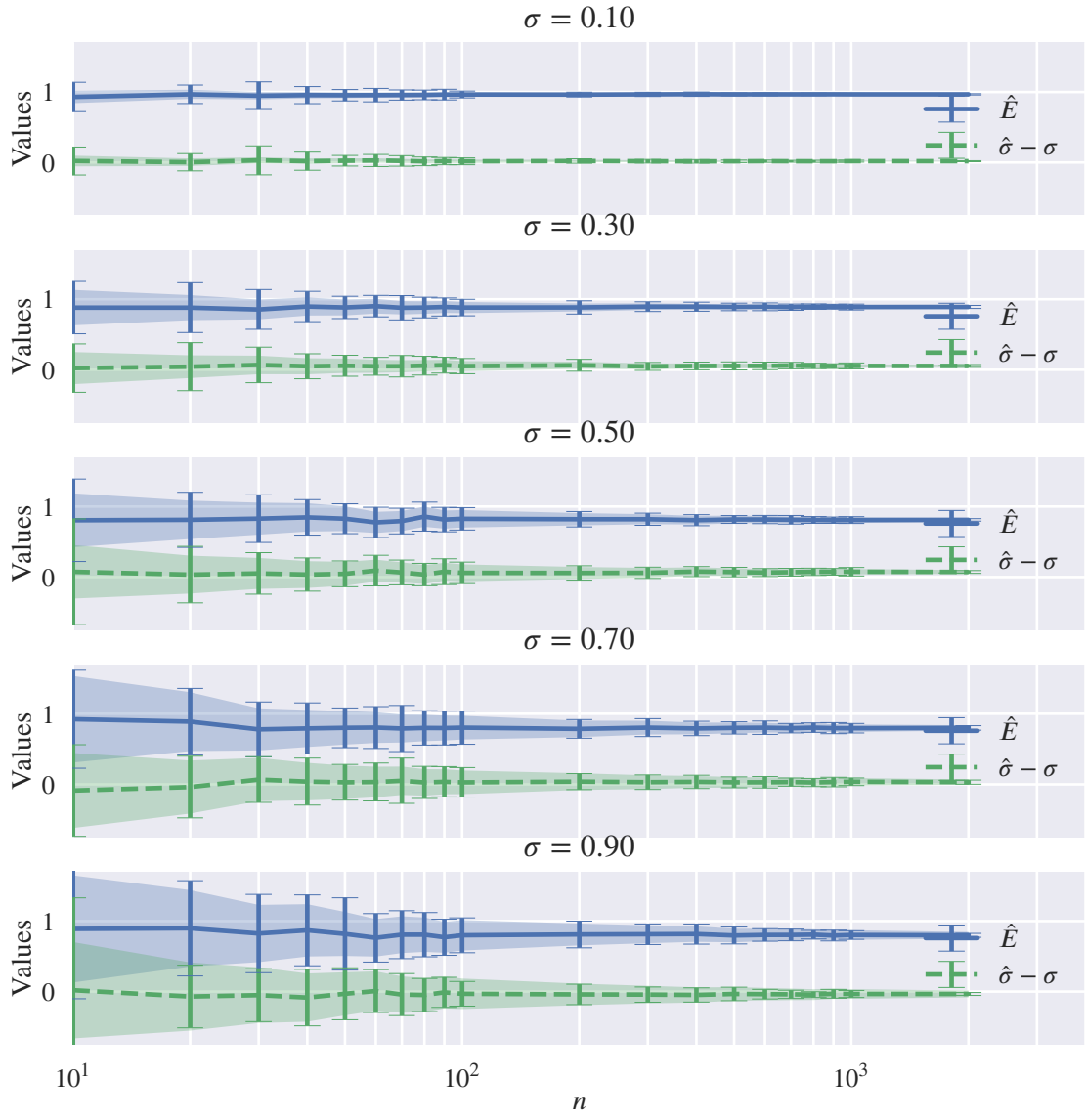


Figure 11: Results from MLE applied on synthetic polycrystals: evolution of mean values for the estimated parameters (broken lines) as functions of the sample size (n) and shape parameter for the lognormal distribution of equivalent radii (σ); the error bars represent ± 2std. The shaded regions illustrate the intervals given by inequations (34) and (36).

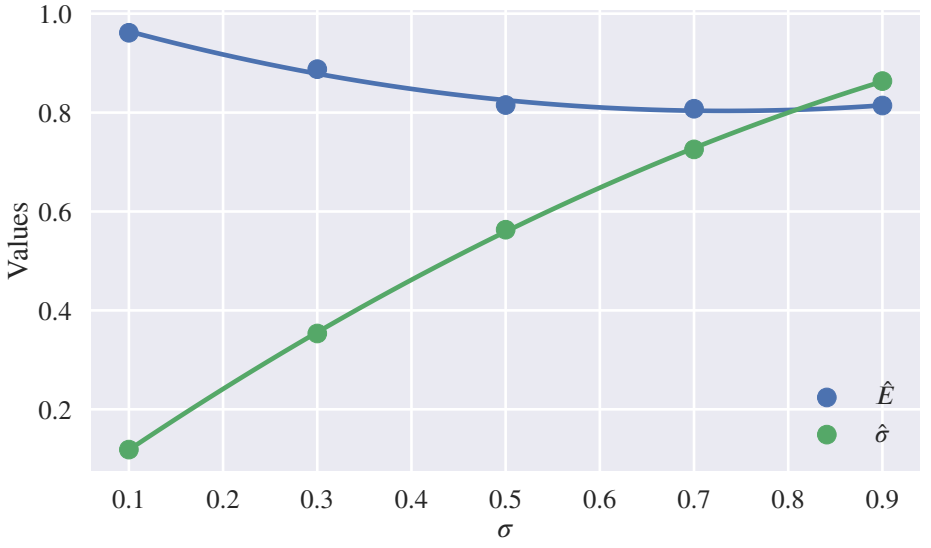


Figure 12: Evolution of mean values values for the estimators given by MLE on synthetic polycrystal (scatter plots). The continuous lines illustrate the polynomial regressions given in Eqs. (43).

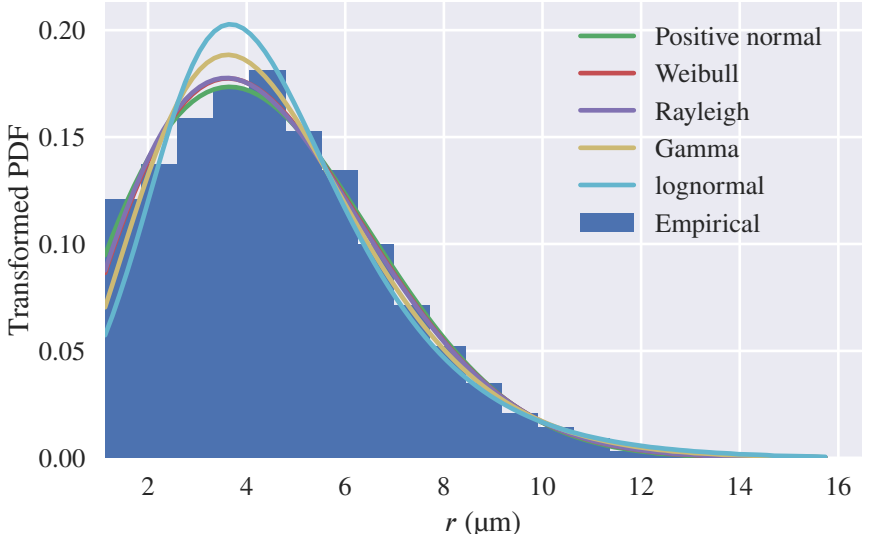


Figure 13: Results from MLE on UO_2: transformed PDFs ($\tilde{f}$), as compared to the empirical distribution (histogram).

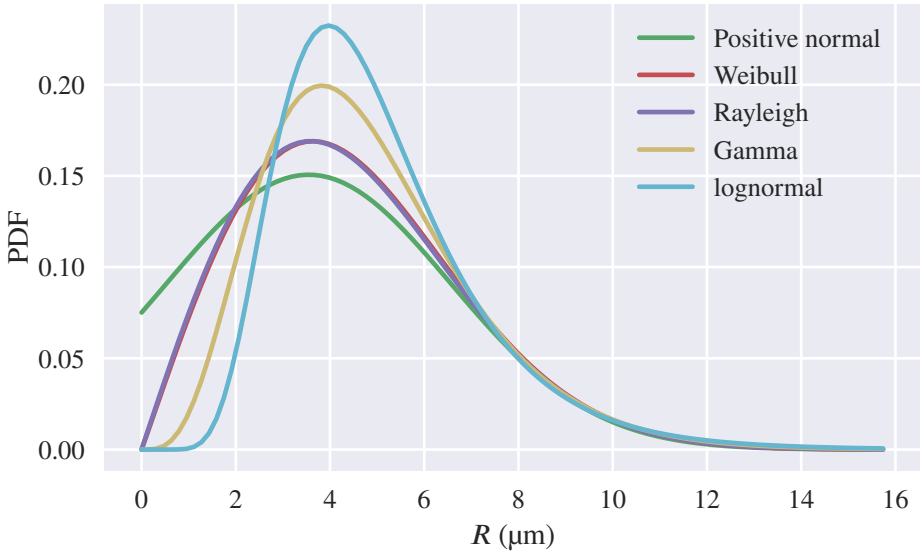


Figure 14: Results from MLE on UO_2: unfolded PDFs (f).