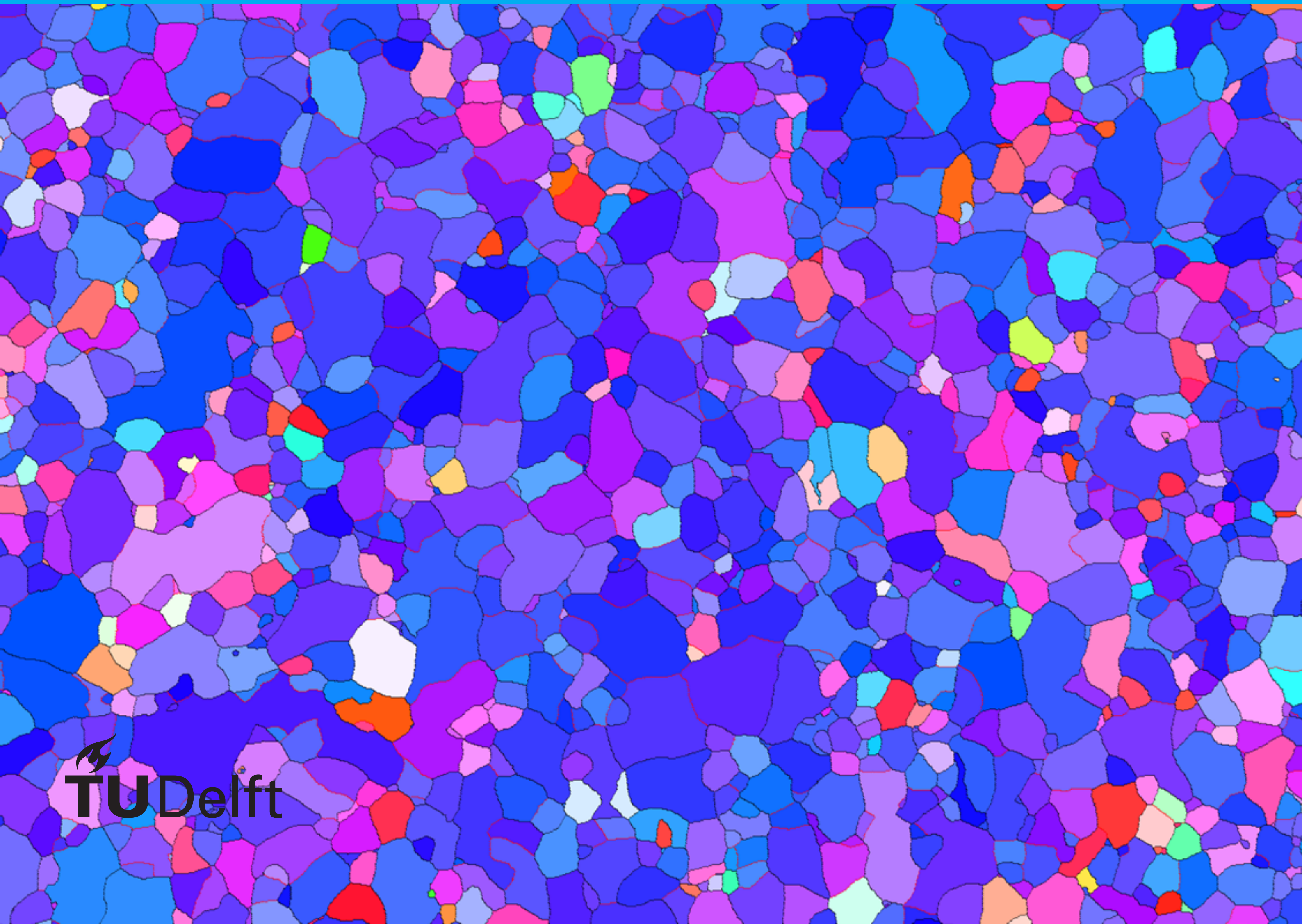
# Estimating the Particle Size Distribution Using the Size and Shape of Section Profiles

J. Faas



# Estimating the Particle Size Distribution Using the Size and Shape of Section Profiles

by

J. Faas

to obtain the degree of Bachelor of Science
at the Delft University of Technology,
to be defended publicly on Monday June 30, 2025 at 14:00.

| | | |
|---|---|---|
| Student number: | 5311004 | |
| Project duration: | February 24, 2025 – June 30, 2025 | |
| Thesis committee: | M. Vittorietti, | TU Delft, supervisor |
| | T. van der Jagt, | TU Delft, supervisor |
| | D. de Laat, | TU Delft |

An electronic version of this thesis is available at https://repository.tudelft.nl/.

Source of the cover image: Van der Jagt et al. (2025).

# Laymen's Summary

Imagine cutting a three-dimensional object in two parts. This reveals a two-dimensional image of its true three-dimensional internal structure. This structure contains important information about the properties of the object. For example, the internal structure of steel provides information about its hardness. However, the observed two-dimensional image does not correspond to the true three-dimensional internal structure. This image can, however, be used to estimate the true three-dimensional internal structure. Methods to obtain this estimate exist, although are never completely accurate. This thesis takes one such existing method, which uses size-related information found in the observed two-dimensional image, and aims to improve its accuracy by also including shape-related information found in the image. By establishing and implementing both of these estimation methods, their performances can be compared. This is done through simulations as well as application to a real two-dimensional image of the true three-dimensional internal structure of steel. This thesis concludes that the new method is, in theory, more accurate. However, this improvement only holds when the information found in the image is observed accurately, which is less likely in practice.

# Summary

The sizes of three-dimensional particles at a microscopic level reveal properties at a macroscopic level for many applications in materials science, but can be difficult to measure. This thesis builds on existing methods that estimate the size distribution of such particles of the same three-dimensional shape, using information obtained from their profiles in two-dimensional cross-sections. This is a well-known stereological problem. An existing method is explained, which yields a maximum likelihood estimator based on the two-dimensional sizes of observed profiles. This method is expanded, resulting in a new maximum likelihood estimator, which is based on the paired two-dimensional sizes and shapes of the observed profiles. The performance of both the existing and the new estimators is analysed in simulations, as well as in an application to real data obtained from a steel microstructure. By comparing the performances of both estimators, this thesis aims to answer whether or not the additional shape-related information improves the resulting estimator. Its conclusions are that the new estimator performs better on average than the existing estimator, but not in general. Moreover, the new estimator only performs better in applications when the observed information is accurate, which is not always the case.

# Contents

# 1 Introduction

One hundred years ago, there was a need among anatomists to investigate tissue of the spleen, an organ in the human body which plays an important role in the immune system. Responsible for this role are clusters of cells of different sizes spread around the organ. It was the size distribution of these so-called follicles that was of interest to anatomical research. However, these could not be directly observed. At the time, the only available observations came from post-mortem studies, where section cuts were taken of the organ, which contained profiles of intersected follicles. Thus, the problem at hand was to estimate the size distribution of follicles in an organ using their observed section profiles. This is a problem in the field of stereology, which focuses on estimating higher-dimensional information from samples of lower-dimensional observations.

Wicksell (1925) introduced a mathematical approach to solve this problem. Since a follicle is approximately of a spherical shape, its profile is approximately circular and thus the sizes of both shapes can be described by their radii. Assuming such a setting with spherical particles, he established an integral equation relating the distribution of the three-dimensional spherical radii to that of the observed two-dimensional circular radii, and used this relation to find an estimate for the distribution of three-dimensional spherical radii, i.e., an estimate for the particle size distribution.

Fast forward a hundred years, and modern-day scanning technology is capable of thoroughly investigating any organ - even inside a living human being - without making a single cut. This has replaced Wicksell's solution in its original application. However, the methods Wicksell used can be applied to various other problems, in particular when considering particles of other shapes than just spheres. A generalisation of the problem is considered in Van der Jagt et al. (2024), where the particles are general, similarly-shaped convex bodies.

This problem generalisation is formally described as follows: consider a three-dimensional opaque medium, which contains randomly positioned and orientated particles. The particles are convex bodies, being compact and convex sets with a non-empty interior, all of the same shape, but of varying sizes. A planar section of the space is taken, yielding an observation containing profiles of all intersected particles. The problem is now to estimate the three-dimensional particle size distribution using the two-dimensional observed section profiles.

Applications of this generalisation are for example found in materials science, where it may be used to estimate the volume distribution of grains in the microstructure of steel, based on two-dimensional observations of this microstructure at surface-level. This grain volume distribution is used to reveal the hardness of steel, without needing to stress-test and inevitably break the bar. However, in order for an estimate of the particle size distribution to be useful in, for example, determining the hardness of steel, it should of course be accurate.

In Van der Jagt et al. (2024) a solution of the generalised problem is presented. A result similar to that of Wicksell is obtained, which estimates the particle volume distribution by using the observed distribution of section profile areas. Accuracy of the resulting estimate varies depending on the particle and true size distribution to estimate, as well as on the sample size of observations. Since this procedure only uses the areas of section profiles to determine its estimate of the particle size distribution, one might wonder whether incorporating more information from section profiles would improve the accuracy of the estimate. In particular, considering that the area describes the size of a section profile, we would like to include information on the shape of a section profile in the estimation procedure.

Hence, in the setting described above, this thesis aims to answer the following question: *Can the joint distribution of area and number of vertices in an observed section profile be used to more accurately estimate the true size distribution of particles, when compared to using the distribution of only the section profile areas?*

To answer this question, a description of a general section profile and its parameters is first given in Chapter 2, along with the introduction of relevant terminology. This section profile information is used as observed samples in Chapter 3, which first describes the aforementioned existing estimation procedure from Van der Jagt et al. (2024), based on the section profile areas as observations. A new estimation procedure is defined next, which is based on paired observations of the area and the number of vertices in a section profile. Then, both the existing and new estimation procedure are implemented in a simulation study, as described in Chapter 4. The accuracy of each procedure is measured and comparisons between both procedures are made. Finally, both procedures are applied to a real data sample in Chapter 5, where the microstructure of steel is observed, and comparisons between performances are made.

# 2
# Describing the Section Profile of a Particle

When a three-dimensional object is cut in two, a two-dimensional cross-section of its interior is revealed. The size and shape of this cross-section depend on the size and shape of the original object, and vary depending on the location and orientation of the cut. In the context of the problem as described in Chapter 1, the observed profiles are such cross-sections of particles. This chapter first mathematically describes the section profile of a particle, along with the introduction of some relevant terminology. Then, several parameters are described that can be obtained from such a section profile. These parameters are used in Chapter 3 for observed section profiles to estimate the particle size distribution with.

## 2.1. General Section Profile

Consider a space $Q \subset \mathbb{R}^3$ containing a finite amount of particles $K_1, \dots, K_N$, as in the problem description in Chapter 1. Since all particles are bodies of the same shape, define $K$ to be the body of that shape which is of volume $1$. This body of reference $K$ will be referred to as the *reference particle*. A body of the same shape with any volume could be used, but taking the body of unit volume as definition of the reference particle, is a choice that will simplify calculations later. For the purposes of this thesis, possible shapes of $K$ will be restricted to polyhedra.

Note that, for any $i \in \{1, \dots, N\}$, particle $K_i$ can now be written as $K_i = \Lambda_i M_i K + x_i = \{\Lambda_i M_i k + x_i : k \in K\}$, where $\Lambda_i > 0$ is some scalar, $M_i$ is some rotation in three dimensions and $x_i$ is a displacement for its position in the space $Q$. Since $K$ has unit volume by above definition, $K_i$, which is scaled by $\Lambda_i$ in three dimensions with respect to $K$, must now be of volume $\Lambda_i^3$. $\Lambda_i$ will be referred to as the *size* of particle $K_i$. It is the distribution of these sizes that is of interest to this thesis.

Next, let $T$ be a random two-dimensional plane cutting through space $Q$. Following the definition in Van der Jagt et al. (2024), $T$ is defined as an Isotropic Uniformly Random (IUR) plane hitting $Q$. Intuitively, this definition ensures that any such IUR plane hitting $Q$ has an equal probability of occurring. The intersection $T \cap Q$ is called the *planar section* resulting from $T$ hitting $Q$.

For any particle $K$ in $Q$, the intersection $T \cap K$ is the *section profile* of that particle, given that this intersection is non-empty. Otherwise, the particle has no profile in $T$. The condition that $T \cap K \neq \emptyset$, i.e., the intersection is non-empty, will be referred to as $T$ hitting $K$. Since $T$ is randomly positioned and orientated, the probability that $T$ hits $K$ precisely along a vertex or an edge of $K$ is $0$. Therefore, conditionally on $T$ hitting $K$, the section profile $T \cap K$ of $K$ exists, has a positive area and is a two-dimensional subset of $Q$. This allows the section profile to be projected to $\mathbb{R}^2$, preserving distances between points. In order to better differentiate between the settings of three-dimensional particles and two-dimensional section profiles, the aforementioned projection to $\mathbb{R}^2$ is done by default throughout this thesis.

Moreover, any section profile $T \cap K$ is a convex two-dimensional polygon. This holds since the original three-dimensional particle $K$ is a polyhedron, i.e., a convex hull of finitely many three-dimensional vertices.

## 2.2. Parameters of a Section Profile

Now that the idea of a general section profile is established, several parameters are introduced to describe a profile with. One parameter is considered to describe the size of a section profile and another is considered to describe its shape. It is from here on assumed that the particle $K$ is hit by a plane $T$.

The first parameter considered is the *area* $A$ of a section profile. Simply defined as $A := \text{area}(T \cap K)$, the area describes the size of a section profile. Note that $A$ only attains positive values. In order to simplify calculations later in Chapter 3, it turns out that transforming the section profile area $A$ is helpful. Hence, $S := \sqrt{A}$ is defined as the *square root transformed area* of the section profile of $K$. This is its default size parameter. Note that $S$ also attains positive values only.

The number of vertices in a section profile is the second parameter considered, which describes the section profile shape. Let $\mathcal{V}$ be the set containing all vertices in the convex polygonal shape of a section profile $T \cap K$. Then, the *number of vertices* $V$ in the section profile of $K$ is simply defined as $V := |\mathcal{V}|$, the size of the set $\mathcal{V}$. Since a section profile is a polygon with positive area, as is stated in Paragraph 2.1, its number of vertices $V$ must be some integer in $\{3, 4, \dots, V_K^{max}\}$ for some maximum $V_K^{max}$, which depends on the shape of $K$.

# 3 Estimating the Particle Size Distribution

In order to be able to estimate the three-dimensional particle size distribution, this chapter describes two methods to find a maximum likelihood estimator for this distribution. Each method is based on different amounts of information that can be obtained from a section profile, as described in Chapter 2.

First, an overview of the existing method is given, using only the section profile area as parameter, one describing the two-dimensional size of the section profile. This procedure yields the first type of estimator for the three-dimensional particle size distribution. This method is adapted next, incorporating the number of vertices, an additional parameter describing section profile shape, alongside its area. This yields a second type of estimator for the particle size distribution, which is based on both a two-dimensional size and a shape parameter. The performance of estimators resulting from each of the two methods is then studied in Chapter 4.

## 3.1. Particle Size Distribution

Consider the setting of space $Q \subset \mathbb{R}^3$ containing particles $K_1, \dots, K_N$ and reference particle $K$ as in Paragraph 2.1. For $i \in \{1, \dots, N\}$, let $\Lambda_i > 0$ be the size of particle $K_i$. These three-dimensional particle sizes are distributed according to some unknown distribution, described by cumulative distribution function (CDF) $H$ and corresponding probability density function (PDF) $h$. These are functions of size as input, which will be denoted by $\lambda$. When cutting $Q$ with an IUR plane $T$ as described in Paragraph 2.1, however, larger particles are more likely to be hit by $T$ than smaller particles are. Therefore, sizes of particles hit by $T$ follow a different, so-called, *length-biased* distribution. The CDF and PDF of this three-dimensional length-biased particle size distribution are denoted by $H^b$ and $h^b$, respectively. See Arratia et al. (2019) for further reading about length- and size-biased distributions.

In Van der Jagt et al. (2024), the following relations between the original and length-biased size distribution functions $H$ and $H^b$ are stated:

$$H^b(\lambda) = \frac{\int_0^\lambda x \mathrm{d}H(x)}{\int_0^\infty x \mathrm{d}H(x)}, \qquad \text{and} \qquad H(\lambda) = \frac{\int_0^\lambda \frac{1}{x} \mathrm{d}H^b(x)}{\int_0^\infty \frac{1}{x} \mathrm{d}H^b(x)}, \qquad \lambda \geq 0. \tag{3.1}$$

Thus, any result in terms of $H^b$ can be related back to $H$ and vice versa, using Equation 3.1. Keeping the above in mind, the methods in Paragraphs 3.2 and 3.3 can focus on estimating $H^b$ instead of $H$.

## 3.2. Estimation Using Section Profile Area

Suppose $T$ is an IUR plane hitting reference particle $K$. The resulting random section profile of $K$ has an area, denoted by $Z$, which is distributed according to distribution and density functions $G_K^A$ and $g_K^A$, respectively. For known reference particles, $g_K^A$ can be approximated arbitrarily closely as described in Van der Jagt et al. (2023), by taking a very large sample of areas observed in simulated section profiles $T \cap K$. $g_K^A$ is then obtained based on this sample by computing the kernel density estimator. An approximation resulting from this method is shown in the left of Figure 3.1 for a cube particle $K$, for example.

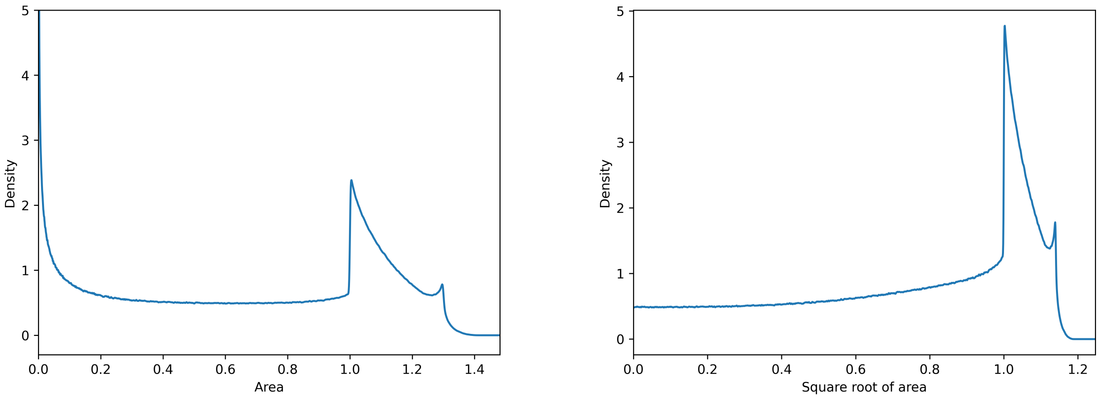


Figure 3.1: Approximations of $g_K^A$ (left) and $g_K^S$ (right) for the cube particle $K$. Made using a sample of $10^7$ simulated section profiles.

Next, let $T$ be an IUR plane hitting space $Q$. Under the assumption that particle $K_i$ is hit by $T$ for some $i \in \{1, \dots, N\}$, let $A_i$ be the resulting section profile area. The distribution of such a random section profile area $A_i$, observed in the planar section $T \cap Q$, is described by distribution function $F^A$ and density function $f^A$. In applications, $F^A$ is the empirical CDF of observed section profile areas in the planar section of $Q$. In simulations, such as in Chapter 4, observations sampled from $F^A$ are generated using the following result, collected from Van der Jagt et al. (2024):

**Lemma 3.1** *Consider a distribution function $H$ with length-biased version $H^b$. Suppose $Z \sim G_K^A$ and $\Lambda \sim H^b$ with $Z$ and $\Lambda^2$ independent. Set $A = Z\Lambda^2$. Then, $A \sim F^A$, and $F^A$, $G_K^A$ and $H^b$ are related via:*

$$F^A(a) = \int_0^\infty G_K^A\left(\frac{a}{\lambda^2}\right) \mathrm{d}H^b(\lambda). \tag{3.2}$$

As a consequence, independently drawing random sizes from a known length-biased particle size distribution $H^b$, and taking areas from simulated IUR planes hitting the reference particle $K$, yields a sample from $F^A$ following Lemma 3.1.

Having collected the result in Equation 3.2, which relates $F^A$, $G_K^A$ and $H^b$, applying a square root transformation to the areas in the equation results in a form that is easier to interpret. For the section profile of particle $K_i$, the square root transformed area $S = \sqrt{A}$ is distributed according to distribution function $F^S$, with corresponding density function $f^S$. Meanwhile, for the section profile of reference particle $K$, the square root transformed area $\sqrt{Z}$ has a distribution described by distribution and density functions $G_K^S$ and $g_K^S$, respectively. Note that $g_K^S$ can be approximated similarly to $g_K^A$ for known reference particles, as is done in the right of Figure 3.1 for a cube particle $K$, for example.

The resulting transformed version of Equation 3.2 is given by (Van der Jagt et al., 2024):

$$F^S(s) = \int_0^\infty G_K^S\left(\frac{s}{\lambda}\right) \mathrm{d}H^b(\lambda). \tag{3.3}$$

The form of Equation 3.3 is recognised as that of a distribution function corresponding to the product of two independent random variables, which is precisely what is stated in Lemma 3.1. Alternatively, the equivalent relation in terms of density functions is given by:

$$f^S(s) = \int_0^\infty g_K^S\left(\frac{s}{\lambda}\right) \frac{1}{\lambda} \mathrm{d}H^b(\lambda). \tag{3.4}$$

Equation 3.4 can be used to find a maximum likelihood (ML) estimator for $H^b$. However, a set of possible functions estimating $H^b$ to maximise over is still missing. Thus, following Van der Jagt et al. (2024), consider observed areas $a_1, \dots, a_n$ that are realisations of the i.i.d. sample $A_1, \dots, A_n \sim f^A$, and corresponding square root transformed areas $s_1, \dots, s_n$, being realisations of i.i.d. sample $S_1, \dots, S_n \sim f^S$.

Set $s_{(1)} < s_{(2)} < \cdots < s_{(n)}$ as the order statistics of $s_1, \dots, s_n$ and set $0 < s_{(0)} < s_{(1)}$. Now, let $\mathcal{F}^+$ be the set containing all distribution functions with domain $(0, \infty)$. Then, define:

$$\mathcal{F}_n^+ := \left\{F \in \mathcal{F}^+ : F(s_{(0)}) = 0 \text{ and } F \text{ is constant on } \left[s_{(i-1)}, s_{(i)}\right),\ i = 1, \dots, n\right\}. \tag{3.5}$$

In Equation 3.5, the set $\mathcal{F}_n^+$ is defined as the subset of $\mathcal{F}^+$ containing all distribution functions with jump locations precisely at the observations $s_i$ and constant elsewhere.

Using Equation 3.4, the log-likelihood function $L$ is defined for $H^b \in \mathcal{F}_n^+$ as:

$$\begin{aligned} L\left(H^b; s_1, \dots, s_n\right) : &= \log\left(\prod_{i=1}^{n} f^S(s_i)\right) = \sum_{i=1}^{n} \log\left(f^S(s_i)\right) \\ &= \sum_{i=1}^{n} \log\left(\int_0^\infty g_K^S\left(\frac{s_i}{\lambda}\right)\frac{1}{\lambda}\mathrm{d}H^b(\lambda)\right). \end{aligned} \tag{3.6}$$

Based on the observed sample of size $n$, finding a ML estimator $\hat{H}^b_{n,A} \in \mathcal{F}_n^+$ that maximises $L$ as in Equation 3.6 is done by taking the argument function $H^b$ that solves the following maximisation problem:

$$\begin{aligned} \hat{H}^b_{n,A} &:= \arg\max_{H^b \in \mathcal{F}_n^+} L\left(H^b; s_1, \dots, s_n\right) \\ &= \arg\max_{H^b \in \mathcal{F}_n^+} \sum_{i=1}^{n} \log\left(\sum_{j=1}^{n} g_K^S\left(\frac{s_i}{s_{(j)}}\right)\frac{1}{s_{(j)}}\left(H^b(s_{(j)}) - H^b(s_{(j-1)})\right)\right). \end{aligned} \tag{3.7}$$

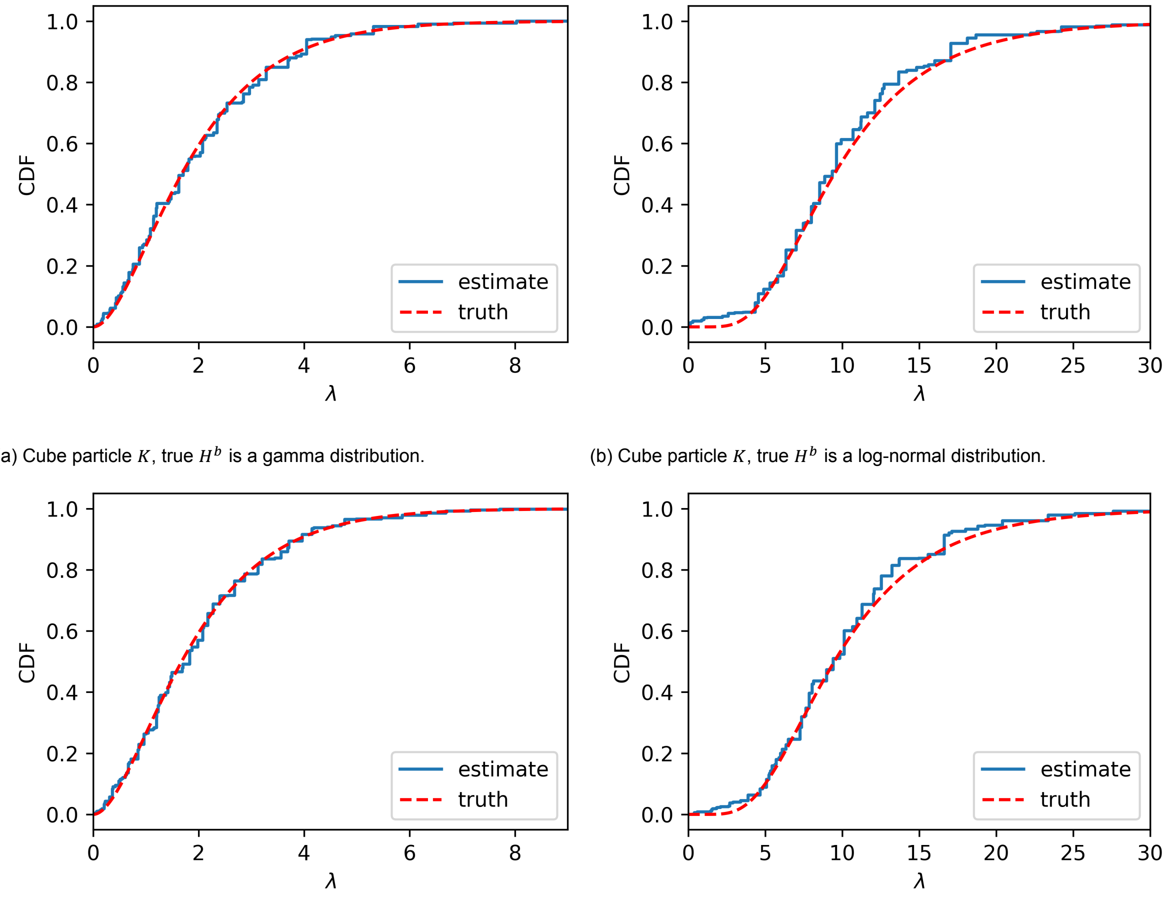
The final step in Equation 3.7 simplifies the expression of $L$ as in Equation 3.6 by discretising the integral with respect to the ordered observations $s_{(0)}, s_{(1)}, \dots, s_{(n)}$. Solving Equation 3.7 is done by solving an optimisation problem, the details of which are found in Van der Jagt et al. (2024) and are also described in Chapter 4 for performed simulations.

Using random samples of $n = 1000$ observed section profile areas, the ML estimates of $H^b$ resulting from Equation 3.7 are visualised in Figure 3.2 for several chosen reference particles $K$ and true size distributions $H$. Choices of $H$ are made such that the corresponding true $H^b$, following from Equation 3.1, is another well-known distribution. As is also stated in Chapter 4: if $H \sim$ Exp$(1)$, then $H^b \sim$ Gamma$(2, 1)$; if $H \sim$ Log-normal$(2, 0.5^2)$, then $H^b \sim$ Log-normal$(2{+}0.5^2, 0.5^2)$. Based on this individual random sample, ML estimator $\hat{H}^b_{n,A}$ appears to approximate the true $H^b$ well, a statement that is further verified in Van der Jagt et al. (2024).

## 3.3. Estimation Using Section Profile Area and Number of Vertices

When the number of vertices is observed in addition to the area of a section profile, this new parameter can be incorporated into the estimation procedure. The observations are now realisations of the paired random variables: $(A_i, V_i)$, where $A_i$ is the section profile area and $V_i$ is the number of vertices, for all observations $i \in \{1, \dots, n\}$. Note that $A_i$ and $V_i$ are not independent, but pairs $(A_i, V_i)$ and $(A_j, V_j)$ are, for $i \neq j$. Then, the $i$-th observation $(a_i, v_i)$ is a realisation of $(A_i, V_i)$. Note that, as is established in Paragraph 2.2, $V$ attains values in $\{3, 4, \dots, V_K^{max}\}$, for some $V_K^{max}$ that depends on $K$, so $V$ is a discrete random variable.

In Figure 3.3 empirical approximations of the density function $g_K^S$ for square root transformed areas are visualised, for several known shapes of $K$. Here, each colour represents the contribution in probability density originating from different observed numbers of vertices $v$. As is clearly shown in this figure, observations with different values of $v$ contribute to the probability density at different values of areas, which is an important observation that motivates the need to incorporate $V$ into the estimation procedure.



(a) Cube particle $K$, true $H^b$ is a gamma distribution.

(b) Cube particle $K$, true $H^b$ is a log-normal distribution.

(c) Dodecahedron particle $K$, true $H^b$ is a gamma distribution.

(d) Dodecahedron particle $K$, true $H^b$ is a log-normal distribution.

Figure 3.2: Maximum likelihood estimates $\hat{H}^b_{n,A}$ of the length-biased size distribution $H^b$. Made using $n = 1000$ sampled section profile areas, following the simulation procedure described in Chapter 4.

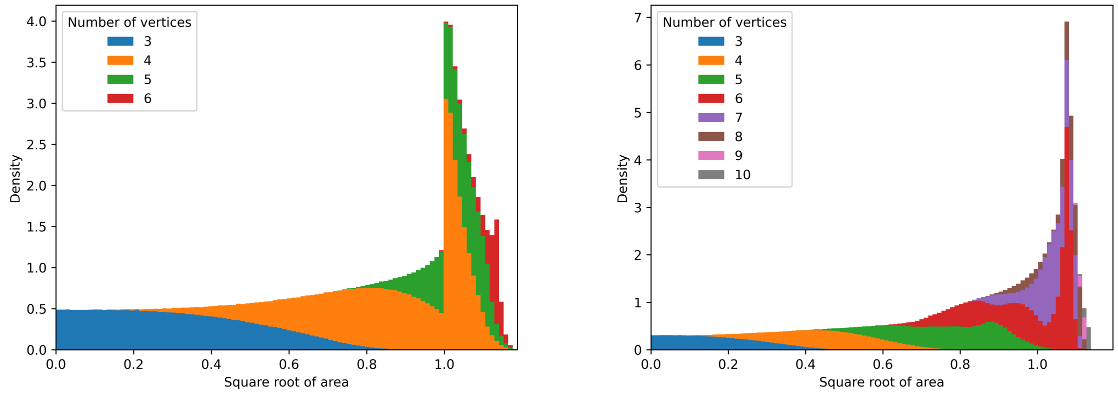


Figure 3.3: Approximations of the probability density function $g^S_K$ of square root transformed section profile areas, split into different colours representing the contributions of different observed numbers of vertices. As reference particle $K$, a cube (left) and a dodecahedron (right) of unit volume have been used. Made using $n = 10^7$ sampled section profiles of $K$.

Similarly to the procedure from Paragraph 3.2, the distribution of profiles in the planar section, and the distribution of the section profiles through the reference particle $K$, are relevant in finding the particle size distribution $H$. However, these functions should now be seen as joint distribution functions of both the square root transformed area $S$ as well as the number of vertices $V$ of the same section profile.

In order to incorporate the new variable $V$, the distribution functions $F^S$ and $G_K^S$ are adapted to the aforementioned joint distribution functions, which are denoted by $F^{S,V}$ and $G_K^{S,V}$, respectively. The distribution function $G_K^{S,V}$ and corresponding density function $g_K^{S,V}$ can be derived from $G_K^S$ and $g_K^S$, respectively, using the rules of conditional probability:

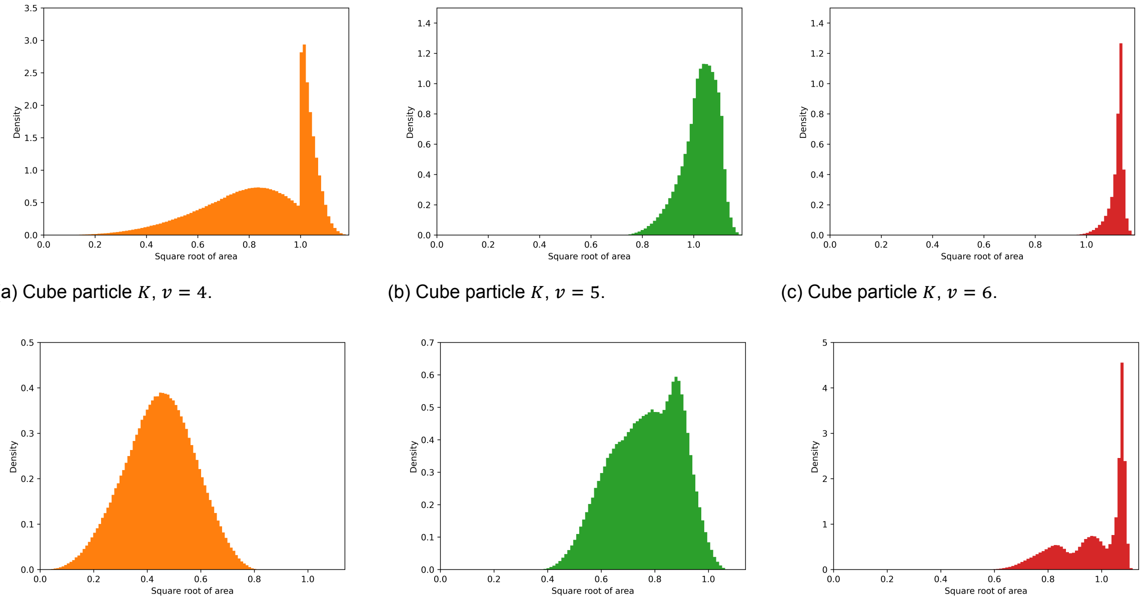
$$G_K^{S,V}(s,v) = G_K^S(s|V \leq v)\mathbb{P}(V \leq v) \quad \text{and} \quad g_K^{S,V}(s,v) = g_K^S(s|V = v)\mathbb{P}(V = v). \tag{3.8}$$

The form of $G_K^{S,V}$ in Equation 3.8 ensures that $\lim_{s\to\infty} G_K^{S,V}(s, V_K^{max}) = 1$ is satisfied, as well as the following relation with the density function $g_K^{S,V}$:

$$G_K^{S,V}(s,v) = \sum_{w=3}^{v} \int_{t=0}^{s} g_K^{S,V}(t,w)\mathrm{d}t.$$

Note that the density function $g_K^{S,V}$ of the square root transformed area and number of vertices in a section profile can be approximated for known reference particles $K$, similarly to the approximation of $g_K^S$, by simulating many IUR planes hitting a particle of shape $K$. This same method may also be used to approximate the quantity $\mathbb{P}(V \leq v)$ for any $v$, which is the probability that a section profile of a particle $K$ has a number of vertices less than or equal to $v$.

As Figure 3.3 already hinted at, the distribution of section profile areas can vary greatly among different numbers of vertices. For several known shapes of $K$ and values of $v$, Figure 3.4 shows approximations of different joint density function $g_K^{S,V}$, resulting from a simulated sample of $10^7$ IUR planes through $K$. This figure highlights the variation in values of the density function $g_K^{S,V}$ that exists between different values of $v$ and between shapes $K$.



(a) Cube particle $K$, $v = 4$.

(b) Cube particle $K$, $v = 5$.

(c) Cube particle $K$, $v = 6$.

(d) Cube particle $K$, $v = 4$.

(e) Dodecahedron particle $K$, $v = 5$.

(f) Dodecahedron particle $K$, $v = 6$.

Figure 3.4: Approximations of the joint probability density function $g_K^{S,V}$ for various numbers of vertices $v$. Made using $n = 10^7$ samples, and either a cube (a, b, c) or a dodecahedron (d, e, f) as reference particle $K$.

Note that the distribution of $V$ obtained by hitting any particle $\Lambda K$ with an IUR plane depends only on the shape of $K$, and is independent of the size $\Lambda$. Thus, the joint distribution function $F^{S,V}$ of the observed area and number of vertices in a section profile, may be related to $G_K^{S,V}$ and $H^b$ similarly to

Equation 3.3 in the following way:

$$F^{S,V}(s,v) = \int_0^\infty G_K^{S,V}\left(\frac{s}{\lambda}, v\right) \mathrm{d}H^b(\lambda) = \mathbb{P}(V \leq v) \int_0^\infty G_K^S\left(\frac{s}{\lambda}\middle| V \leq v\right) \mathrm{d}H^b(\lambda). \tag{3.9}$$

The density function $f^{S,V}$ corresponding to $F^{S,V}$ follows again from the rules of conditional probability:

$$\begin{aligned} f^{S,V}(s,v) &= f^S(s|V=v)\mathbb{P}(V=v) \\ &= \mathbb{P}(V=v)\int_0^\infty g_K^S\left(\frac{s}{\lambda}\middle| V = v\right)\frac{1}{\lambda}\mathrm{d}H^b(\lambda) = \int_0^\infty g_K^{S,V}\left(\frac{s}{\lambda}, v\right)\frac{1}{\lambda}\mathrm{d}H^b(\lambda). \end{aligned} \tag{3.10}$$

### 3.3.1. Maximum Likelihood Estimator

Having established distributions $F^{S,V}$ and $G_K^{S,V}$, as well as their corresponding densities $f^{S,V}$ and $g_K^{S,V}$, a maximum likelihood estimator $\hat{H}_{n,A,V}^b$ for the particle size distribution $H^b$ can be obtained similarly to the process in Paragraph 3.2. Using these joint distribution and density functions, it is based on both the area and the number of vertices in a section profile.

Consider observed realisations $(s_i, v_i)$ of the $n$ pairs of random variables $(S_i, V_i) \sim f^{S,V}$, independent from other pairs, for $i = 1, \dots, n$. Set $s_{(1)} < s_{(2)} < \cdots < s_{(n)}$ as the order statistics of $s_1, \dots, s_n$ and set $0 < s_{(0)} < s_{(1)}$. Based on these order statistics, define $\mathcal{F}_n^+$ as in Equation 3.5, where $\mathcal{F}^+$ is again the set containing all distribution functions with domain $(0, \infty)$. Then, the log-likelihood function $L$ takes the following form:

$$\begin{aligned} L\left(H^b; (s_1, v_1), \dots, (s_n, v_n)\right) &= \log\left(\prod_{i=1}^n f^{S,V}(s_i, v_i)\right) = \sum_{i=1}^n \log\left(f^{S,V}(s_i, v_i)\right) \\ &= \sum_{i=1}^n \log\left(\int_0^\infty g_K^{S,V}\left(\frac{s_i}{\lambda}, v_i\right)\frac{1}{\lambda}\mathrm{d}H^b(\lambda)\right). \end{aligned} \tag{3.11}$$

Based on the observed sample, the distribution function $\hat{H}_{n,A,V}^b$ that maximises $L$ as in Equation 3.11 is a ML estimator for $H^b$ based on both the area and the number of vertices in a section profile. $\hat{H}_{n,A,V}^b$ is found by solving the following maximisation problem:

$$\begin{aligned} \hat{H}_{n,A,V}^b &:= \arg\max_{H^b \in \mathcal{F}_n^+} L\left(H^b; (s_1, v_1), \dots, (s_n, v_n)\right) \\ &= \arg\max_{H^b \in \mathcal{F}_n^+} \sum_{i=1}^n \log\left(\sum_{j=1}^n g_K^{S,V}\left(\frac{s_i}{s_{(j)}}, v_i\right)\frac{1}{s_{(j)}}\left(H^b(s_{(j)}) - H^b(s_{(j-1)})\right)\right). \end{aligned} \tag{3.12}$$

Similarly to Equation 3.7, the last step in Equation 3.12 simplifies the integral from $L$ as in Equation 3.11 by discretisation with respect to the ordered $s_{(0)}, s_{(1)}, \dots, s_{(n)}$. The solution to Equation 3.12 is found by solving an optimisation problem, as is described in Chapter 4.

# 4

# Simulating Methods of Estimation

In order to illustrate the results of both the method from Chapter 3, one based on the section profile area only, and the other based on both the section profile area and the number of vertices, and to be able to compare the performance of both methods, a simulation study is performed. This section first describes the procedure followed in the simulations. Several simulation results are given next, which are then used to draw conclusions comparing the estimates resulting from each method.

## 4.1. Simulation Procedure

The following procedure to perform simulations by, is similar to what is done in Van der Jagt et al. (2024). Several distributions $H$ are chosen such that their corresponding length-biased distributions $H^b$, according to the relation in Equation 3.1, are well-known, making comparisons of estimates to the true distributions straightforward. The following distributions are chosen:

- If $H$ is the standard *exponential* distribution, then the corresponding $H^b$ is a known *gamma* distribution:

  $$H(\lambda) = 1 - e^{-\lambda} \sim \text{Exp}(1), \quad \text{and} \quad H^b(\lambda) = 1 - (\lambda + 1)e^{-\lambda} \sim \text{Gamma}(2, 1), \quad \text{for } \lambda \geq 0.$$

- If $H$ is a *log-normal* distribution with parameters $\mu$ and $\sigma > 0$, then the corresponding $H^b$ is also a *log-normal* distribution, with parameters $\mu + \sigma^2$ and $\sigma$:

  $$\begin{aligned} H(\lambda) &= \Phi\left(\frac{\log(\lambda) - \mu}{\sigma}\right) \sim \text{Lognormal}(\mu, \sigma), \text{ and} \\ H^b(\lambda) &= \Phi\left(\frac{\log(\lambda) - \mu - \sigma^2}{\sigma}\right) \sim \text{Lognormal}(\mu + \sigma^2, \sigma), \qquad \text{for } \lambda > 0. \end{aligned}$$

  Here, $\Phi$ is the standard normal distribution function. In simulations, the parameters $\mu = 2$ and $\sigma = \frac{1}{2}$ are set.

Several convex regular polyhedra are chosen as shapes of the reference particle $K$: a tetrahedron, a cube and a dodecahedron. Using such a $K$ of unit volume, distribution functions $g_K^S$ and $g_K^{S,V}$ are approximated by simulating $10^7$ IUR planes hitting $K$ and storing the paired observations of square root transformed areas and numbers of vertices from resulting section profiles. Then, simulating an observed section profile from distribution $F^{S,V}$ can be done using Lemma 3.1. A size $\Lambda \sim H^b$ and a section profile through reference particle $K$, so $(\sqrt{Z}, V) \sim G_K^{S,V}$, where $Z$ is the area and $V$ is the number of vertices of a section profile, are sampled independently. By Lemma 3.1, it now follows that $A := Z\Lambda^2$ is distributed as an area sampled from $F^{A,V}$, without square root transformation. After applying the transformation $S := \sqrt{A} = \sqrt{Z}\Lambda$ to the area, it now follows that $(S, V)$ is a paired sample from $F^{S,V}$. This holds since $(\sqrt{Z}, V)$ was an initial paired sample, and applying a scaling by $\Lambda$ and a square root transformation to the area of a section profile do not change its number of vertices.

Thus, independent samples $(\sqrt{Z_1}, V_1), \dots, (\sqrt{Z_n}, V_n) \sim G_K^{S,V}$ and $\Lambda_1, \dots, \Lambda_n \sim H^b$ are generated to make a sample $(S_1, V_1), \dots, (S_n, V_n) \sim F^{S,V}$ of $n$ observed section profiles, with $S_i = \sqrt{Z_i}\Lambda_i$ for $i \in \{1, \dots, n\}$. This sampling process is repeated $100$ times for several values of $n$ and for each distribution $H^b$ and reference particle $K$ mentioned before.

### 4.1.1. Algorithm

Given a simulated sample of $n$ observed section profiles, a ML estimate $\hat{H}^b_{n,A,V}$ of length-biased distribution $H^b$ is computed. This is done following the method from Paragraph 3.3, solving Equation 3.12, using the sampled pairs $(S_1, V_1), \dots, (S_n, V_n)$ of square root transformed areas and numbers of vertices. Similarly, a ML estimate $\hat{H}^b_{n,A}$ is computed following the method from Paragraph 3.2, solving Equation 3.7, using only the square root transformed areas $S_1, \dots, S_n$ from the same sample.

In each case, the maximisation problem can be adapted to the following form. The likelihood equations to be maximised in both Equation 3.7 and Equation 3.12 are of the following form:

$$l(\beta) = \sum_{i=1}^{n} \log\left( \sum_{j=1}^{n} \alpha_{i,j} \left(\beta_j - \beta_{j-1}\right) \right), \ \beta \in \mathbb{R}^{n+1}, \tag{4.1}$$

with $\beta_0 = 0$, $\beta_j = H^b\left(s_{(j)}\right)$ for $j \in \{1, \dots, n\}$, as well as $\alpha_{i,j} = g_K^S\left(\frac{s_i}{s_{(j)}}\right)\frac{1}{s_{(j)}}$ when estimating $\hat{H}^b_{n,A}$ (as in Equation 3.7) and $\alpha_{i,j} = g_K^{S,V}\left(\frac{s_i}{s_{(j)}}, v_i\right)\frac{1}{s_{(j)}}$ when estimating $\hat{H}^b_{n,A,V}$ (as in Equation 3.12). The set $\mathcal{C}^+$ of possible vectors $\beta$, equivalent to the set $\mathcal{F}_n^+$ (as in Equation 3.5) of possible distribution functions $H^b$ to maximise over, is given by:

$$\mathcal{C}^+ := \left\{\beta \in \mathbb{R}^{n+1} : 0 = \beta_0 \leq \beta_1 \leq \dots \leq \beta_n \leq 1\right\}.$$

Note that the set $\mathcal{C}^+$ is closed and bounded, and thus compact, so $l$ has a maximum on $\mathcal{C}$. The maximisation problem is now given by $\hat{\beta} := \arg\max_{\beta \in \mathcal{C}^+} l(\beta)$. Moreover, since the function $l$ is concave, this maximisation problem is computationally tractable.

The coefficients $\alpha_{i,j}$ in Equation 4.1 are computed for all $i, j \in \{1, \dots, n\}$, using a Monte Carlo approximation of the corresponding density function, either $g_K^S$ or $g_K^{S,V}$, as described in Van der Jagt et al. (2023). This approximation is made using kernel density estimation with a boundary correction method described in Schuster (1985). The kernel density estimation uses a band-width determined by the Sheather-Jones method (Sheather and Jones, 1991), and $1000$ bins. Figure 4.1 shows examples of such resulting approximations of $g_K^{S,V}$ when $K$ is a cube or a dodecahedron. The corresponding examples of such approximations of $g_K^S$ are shown in Figure 3.1.

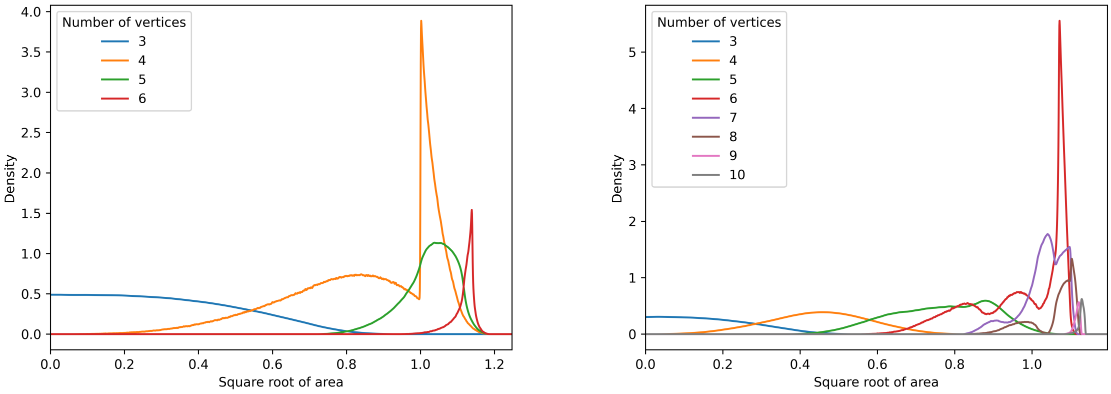


Figure 4.1: Monte Carlo approximations of $\tilde{g}_K$ for the cube (left) and dodecahedron (right) particles $K$.

This form of the maximisation problem is then solved by the *hybrid ICM-EM algorithm*, as proposed in Wellner and Zhan (1997). This hybrid algorithm is a combination of the *Expectation Maximisation*

(EM) and *Iterative Convex Minorant* (ICM) algorithms. The idea is to perform one iteration of ICM, followed by one iteration of EM, as one iteration of the hybrid algorithm. Each of these individual algorithms may also be used to obtain similar results, however, simulations in Wellner and Zhan (1997) and in Van der Jagt et al. (2024) showed that the hybrid algorithm is faster. It appears that EM converges quickly when far from the optimum and slower when closer to the optimum. Conversely, ICM appears to converge slowly when far from the optimum and quicker when closer to the optimum. The hybrid algorithm, instead, seems to inherit the strength of both algorithms and converges relatively fast at any point. Therefore, the maximisation problem is solved by hybrid ICM-EM in all simulations. It is implemented according to the following description:

1. **Initial estimate of $\beta$.**
   $\beta^{(0)}$, the initial estimate of $\beta$, is set to equal $\beta^{(0)} := \left(\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n} = 1\right) \in \mathcal{C}^+$. Iteration 1 of the algorithm begins next, at step 2.

2. **One iteration of the ICM algorithm.**
   The version of the ICM algorithm used here originates from Jongbloed (1998). This algorithm solves the minimisation problem $\hat{\beta} := \arg\min_{\beta \in \mathcal{C}^+} \phi(\beta)$, where the function $\phi$ is given by:

$$\phi(\beta) = -\beta_n - \frac{1}{n} l(\beta) = -\beta_n - \frac{1}{n} \sum_{i=1}^{n} \log\left(\sum_{j=1}^{n} \alpha_{i,j} \left(\beta_j - \beta_{j-1}\right)\right), \tag{4.2}$$

   with $l$ as in Equation 4.1. The term $-\beta_n$ is added to ensure that $\beta_n = 1$. For any $\beta \in \mathcal{C}^+$ with $\beta_n < 1$, adhering the condition $\beta_n = 1$ instead will increase the value of $l(\beta)$, since all $\alpha_{i,j} \geq 0$ and is therefore more optimal.

   In iteration $k$ of the hybrid algorithm, given previous estimate $\beta^{(k-1)}$, the ICM step uses the following expression $\phi_{(k)}$ to locally approximate $\phi$:

$$\begin{aligned} \phi_{(k)}(\beta) = &\left(\beta - \beta^{(k-1)} + W\left(\beta^{(k-1)}\right)^{-1} \nabla\phi\left(\beta^{(k-1)}\right)\right)^{\intercal} W\left(\beta^{(k-1)}\right) \\ &\cdot \left(\beta - \beta^{(k-1)} + W\left(\beta^{(k-1)}\right)^{-1} \nabla\phi\left(\beta^{(k-1)}\right)\right), \end{aligned} \tag{4.3}$$

   where $\nabla\phi$ is the gradient of $\phi$, and $W\left(\beta^{(k-1)}\right)$ is a diagonal matrix with the same diagonal entries as the Hessian matrix of $\phi$:

$$\left(W\left(\beta^{(k-1)}\right)\right)_{j,j} = \frac{\partial^2}{\partial \beta_j^2} \phi\left(\beta^{(k-1)}\right), \text{ for } j \in \{1, \ldots, n\}.$$

   The function $\phi_{(k)}$ as in Equation 4.3 is then minimised over $\mathcal{C}^+$, the details of which are found in Jongbloed (1998), yielding candidate estimate $\beta := \arg\min_{\beta \in \mathcal{C}^+} \phi_{(k)}$ for $\beta^{(k)}$. If $\phi(\beta)$, the value of $\phi$ evaluated at candidate $\beta$, is sufficiently decreasing when compared to $\phi\left(\beta^{(k-1)}\right)$, with $\phi$ as in Equation 4.2, then the $k$-th estimate $\beta^{(k)}$ is set to be $\beta^{(k)} := \beta$. Here, satisfying the following condition is considered to be a sufficient decrease:

$$\phi(\beta) < \phi\left(\beta^{(k-1)}\right) + (1 - \epsilon)\nabla\phi\left(\beta^{(k-1)}\right)^{\intercal}\left(\beta - \beta^{(k-1)}\right), \tag{4.4}$$

   for some $\epsilon \in \left(0, \frac{1}{2}\right)$. In simulations, $\epsilon = 0.25$ is set. If the condition from Equation 4.4 is not met, then $\beta^{(k)}$ is obtained according to the line-search process described in Jongbloed (1998), as a convex combination of $\beta$ and $\beta^{(k-1)}$.

3. **One iteration of the EM algorithm.**
   The description of the EM algorithm from Wellner and Zhan (1997) is followed. For notational convenience, a probability vector $p$ is introduced, corresponding to an estimate $\beta$, with coefficients:

$$p_j := \beta_j - \beta_{j-1} \geq 0, \tag{4.5}$$

for $j \in \{1, \dots, n\}$. Furthermore, $p$ satisfies $\sum_{j=1}^{n} p_j = 1$. Thus, the set of probability vectors $\mathcal{P}_n$ that the EM algorithm maximises over is given by:

$$\mathcal{P}_n = \left\{ (p_1, \dots, p_n) \in \mathbb{R}^n : \sum_{i=1}^{n} p_i = 1 \text{ and } p_i \geq 0, \text{ for } i \in \{1, \dots, n\} \right\}.$$

In iteration $k$, probability vector $p^{(k-1)} \in \mathcal{P}_n$ corresponding to the given estimate $\tilde{\beta}^{(k)}$ resulting from step 2 is constructed from $\tilde{\beta}^{(k)}$ following Equation 4.5. As described in Van der Jagt et al. (2024), the update rule is now given by:

$$p_j^{(k)} = \frac{1}{n} \sum_{i=1}^{n} \frac{\alpha_{i,j}}{\sum_{q=1}^{n} \alpha_{i,q}\, p_q^{(k-1)}} p_j^{(k-1)}, \tag{4.6}$$

for $j \in \{1, \dots, n\}$. Then, estimate $p^{(k)}$ resulting from Equation 4.6 is converted back to the definitive estimate $\beta^{(k)}$ of iteration $k$ as follows:

$$\beta_j^{(k)} = \sum_{i=1}^{j} p_i^{(k)}, \text{ for } j \in \{1, \dots, n\} \qquad \text{and} \qquad \beta_0^{(k)} = 0.$$

4. **Checking the stopping criterion.**
After step 3 in iteration $k$, the following stopping criterion is evaluated:

$$\max_{j \in \{1, \dots, n\}} \left| \beta_j^{(k)} - \beta_j^{(k-1)} \right| < \varepsilon. \tag{4.7}$$

In simulations, $\varepsilon = 10^{-4}$ is taken. The algorithm is terminated if the condition in Equation 4.7 is satisfied for $10$ successive iterations. $\hat{\beta}$ is then given by the final estimate $\beta^{(k)}$. Otherwise, the algorithm performs iteration $k+1$ from step 2.

### 4.1.2. De-biasing and Regularisation

When de-biasing an estimate of the length-biased particle size distribution to an estimate of the true size distribution, it was proposed in Van der Jagt et al. (2024) to apply a regularisation first. This is motivated by the observation that, in order to obtain estimate $\hat{H}_{n,A}$, directly plugging a close approximation $\hat{H}_{n,A}^{b}$ of biased size distribution $H^b$ into Equation 3.1 fails to approximate the true particle size distribution $H$ in simulations. This deviation is a result of the behaviour of $\hat{H}_{n,A}^{b}$ near zero. Therefore, both $\hat{H}_{n,A}$ and $\hat{H}_{n,A,V}$ are regularised here in a similar way.

For truncation parameter $t_n > 0$, consider estimate $\hat{H}_n^b$, either $\hat{H}_{n,A}^b$ or $\hat{H}_{n,A,V}^b$, truncated at $t_n$ as follows:

$$\hat{H}_n^b(\lambda; t_n) := \begin{cases} \frac{\hat{H}_n^b(\lambda) - \hat{H}_n^b(t_n)}{1 - \hat{H}_n^b(t_n)} & \text{if } \lambda \geq t_n, \\ 0 & \text{otherwise.} \end{cases} \tag{4.8}$$

Now, plugging truncated estimate $\hat{H}_n^b$ from Equation 4.8 into Equation 3.1 yields:

$$\hat{H}_n(\lambda; t_n) := \frac{\int_0^\lambda \frac{1}{x} \mathrm{d}\hat{H}_n^b(x; t_n)}{\int_0^\infty \frac{1}{x} \mathrm{d}\hat{H}_n^b(x, t_n)} = \begin{cases} \frac{\int_{t_n}^{\lambda} \frac{1}{x} \mathrm{d}\hat{H}_n^b(x)}{\int_{t_n}^{\infty} \frac{1}{x} \mathrm{d}\hat{H}_n^b(x)} & \text{if } \lambda \geq t_n, \\ 0 & \text{if } 0 \leq \lambda < t_n. \end{cases} \tag{4.9}$$

The resulting estimate $\hat{H}_n$ corresponds to $\hat{H}_{n,A}$ or $\hat{H}_{n,A,V}$, respectively, truncated at $t_n$. These estimate the true particle size distribution $H$. In simulations, resulting estimates of $H$ will correspond to this truncated estimate of the true size distribution $H$, given by Equation 4.9, instead of the direct plug-in estimate.

As is established in Van der Jagt et al. (2024), $\hat{H}_{n,A}$ as in Equation 4.9 is indeed a close approximation of $H$ given a close approximation $\hat{H}_{n,A}^b$, and for an appropriate choice of truncation parameter $t_n$. When a sequence of estimates $\hat{H}^b_{n,A}$ converging to $H^b$ is given, convergence of resulting truncated estimates $\hat{H}_{n,A}$ to the true $H$ is even proven. It is remarked that $t_n = \sqrt{\|\hat{H}^b_{n,A} - H^b\|_\infty}$ theoretically may be taken as an appropriate choice. In practice, however, $H^b$ is unknown and therefore this quantity is not known either. The following procedure to obtain a sufficiently accurate $t_n$ is proposed by Van der Jagt et al. (2024) instead. This is used for the computation of estimates $\hat{H}_{n,A}$, based on square root transformed section profile areas. For $s \in \mathbb{R}$, consider:

$$\begin{aligned} \hat{F}^S_n(s;t) &:= \int_0^\infty G^S_K\left(\frac{s}{\lambda}\right) \mathrm{d}\hat{H}^b_{n,A}(\lambda;t), \\ \overline{F}^S_n(s) &:= \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{s_i \le s\}. \end{aligned} \tag{4.10}$$

Note that, for some truncation $t > 0$, the function $\hat{F}^S_n$ as in Equation 4.10 is the distribution of observed square root transformed areas following from Equation 3.3, if the truncated $\hat{H}^b_{n,A}$ is the true length-biased size distribution. Meanwhile, $\overline{F}^S_n$ is the empirical distribution of observed square root transformed areas in the sample. The proposed choice $\hat{t}_{n,A}$ for $t_n$ is now motivated by minimising the distance between these two functions in the $L^1$-norm, and is given by:

$$\hat{t}_{n,A} := \arg\min_{t \in \{s_1,\dots,s_n\}} \int_0^\infty \left| \hat{F}^S_n(s;t) - \overline{F}^S_n(s) \right| \mathrm{d}s, \tag{4.11}$$

with $\hat{F}^S_n$ and $\overline{F}^S_n$ as in Equation 4.10. Note that minimising the integral in Equation 4.11 over the sampled $\{s_1, \dots, s_n\}$ is done in order to reduce computations. Simulations performed in Van der Jagt et al. (2024) have shown that this choice of $\hat{t}_{n,A}$ results in an accurate estimate $\hat{H}_{n,A}$.

When computing estimates $\hat{H}_{n,A,V}$ based on both the square root transformed area and number of vertices of a section profile, a convergence result similar to that of in the previous setting is not proven. It is still attempted to adapt above procedure to this setting first, using the following functions for $s \in \mathbb{R}$ and $v \in \mathbb{N}$:

$$\begin{aligned} \hat{F}^{S,V}_n(s,v;t) &:= \int_0^\infty G^{S,V}_K\left(\frac{s}{\lambda}, v\right) \mathrm{d}\hat{H}^b_{n,A,V}(\lambda;t), \\ \overline{F}^{S,V}_n(s,v) &:= \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{s_i \le s\}\, \mathbb{1}\{v_i \le v\}. \end{aligned} \tag{4.12}$$

Note that now $\hat{F}^{S,V}_n$ as in Equation 4.12 is the joint distribution of observed square root transformed area and number of vertices in a section profile, if the truncated $\hat{H}^b_{n,A,V}$ is the true length-biased size distribution. Similarly, $\overline{F}^{S,V}_n$ is now the empirical joint distribution of square root transformed areas and numbers of vertices observed in the sample. Then, the corresponding choice $\hat{t}_{n,A,V}$ for $t_n$ in this case is given by:

$$\hat{t}_{n,A,V} := \arg\min_{t \in \{s_1,\dots,s_n\}} \sum_{v=3}^{V^{max}_K} \int_0^\infty \left| \hat{F}^{S,V}_n(s,v;t) - \overline{F}^{S,V}_n(s,v) \right| \mathrm{d}s, \tag{4.13}$$

with $\hat{F}^{S,V}_n$ and $\overline{F}^{S,V}_n$ as in Equation 4.12.

However, taking $\hat{t}_{n,A,V}$ as in Equation 4.13 appears to structurally choose lower values than desired in simulations, when compared to the theoretical $t_n = \sqrt{\|\hat{H}^b_{n,A} - H^b\|_\infty}$, which is known in simulations. Truncating at a lower value than desired may result in an unwanted, relatively large, jump in the estimate at low values. Consider computing $\hat{H}_{n,A,V}$ with such a $\hat{t}_{n,A,V}$ following from Equation 4.9. Say $\hat{t}_{n,A,V}$ is equal to a point $s_{(i)}$ for some $i \in \{1, \dots, n\}$. $\hat{H}_{n,A,V}$ is computed at a point $s_{(j)}$, for some $j \in \{i, \dots, n\}$ such that $s_{(j)} \ge s_{(i)} = \hat{t}_{n,A,V}$, by:

$$\hat{H}_{n,A,V}(s_{(j)}) = \frac{\sum_{k=i}^{j} \frac{1}{s_{(k)}} \left(\hat{H}^b_{n,A,V}(s_{(k)}) - \hat{H}^b_{n,A,V}(s_{(k-1)})\right)}{\sum_{k=i}^{n} \frac{1}{s_{(k)}} \left(\hat{H}^b_{n,A,V}(s_{(k)}) - \hat{H}^b_{n,A,V}(s_{(k-1)})\right)}. \tag{4.14}$$

If the estimate $\hat{H}^b_{n,A,V}$ has non-zero probability mass in the point $s_{(k)}$ for some small $k \geq i$, which is given by $\hat{H}^b_{n,A,V}(s_{(k)}) - \hat{H}^b_{n,A,V}(s_{(k-1)})$, then the numerators in the sum terms of Equation 4.14, equal to this probability mass of $\hat{H}^b_{n,A,V}$ at $s_{(k)}$, will be greater than zero. A problem arises at values of the denominator $s_{(k)}$ that are close to zero, if it is of an order close to, or even smaller than that of the numerator. In such cases, this term will inflate the total of both sums, and thereby raising the function values of estimate $\hat{H}_{n,A,V}$, which results from Equation 4.14, to be relatively close to $1$ starting from values close to zero.

On the left side of Figure 4.2 a simulation is shown where this phenomenon is present, but clearly not desired. Several estimates are inflated to function values greater than $0.6$, starting from very small values of $\lambda$, while the true distribution $H$ still has function values close to zero at this $\lambda$.

In simulations, this phenomenon occurred mostly in estimates $\hat{H}_{n,A,V}$ when using the cube or tetrahedron $K$, for different values of $n$ and different true distributions $H$. The only times it occurred in estimates $\hat{H}_{n,A}$ was when using the tetrahedron $K$ and mostly for the exponentially distributed true $H$, but to less extreme extents than estimates $\hat{H}_{n,A,V}$ in the same simulations. Moreover, the phenomenon appears to diminish for higher values of $n$ in estimates $\hat{H}_{n,A}$ using the tetrahedron $K$ and exponentially distributed $H$, whereas it appears to remain equally present in estimates $\hat{H}_{n,A,V}$ of the same simulations for any value of $n$.

Simulations have thus shown that the phenomenon described above is mostly a problem for estimates $\hat{H}_{n,A,V}$, based on both the area and number of vertices of a section profile, and not for estimates $\hat{H}_{n,A}$, based on the section profile area only. A possible explanation for this is the lack of a convergence result of estimates $\hat{H}_{n,A,V}$, as mentioned before, even when these are de-biased from estimates $\hat{H}^b_{n,A,V}$ that converge to $H^b$.

Another explanation is that the algorithm assigns slightly higher, or more concentrated probability mass to values of $s_{(k)}$ which are close to zero, when computing $\hat{H}^b_{n,A,V}$, than it does when computing $\hat{H}^b_{n,A}$. This results in higher values of $\hat{H}^b_{n,A,V}(s_{(k)}) - \hat{H}^b_{n,A,V}(s_{(k-1)})$, the numerators of the sums in Equation 4.14, which in turn inflates the resulting de-biased estimates $\hat{H}_{n,A,V}$.

In order to resolve this issue, a possible adjustment could be to more evenly spread the probability mass in estimates $\hat{H}^b_{n,A,V}$, or at least to use such an adjusted version of $\hat{H}^b_{n,A,V}$ in the de-biasing step of Equation 4.9. This is left as a suggestion for future research. For now, a small adjustment is made to the choice of $\hat{t}_{n,A,V}$ as in Equation 4.13, which attempts to catch and avoid the most problematically large values in the sum terms of Equation 4.14. The idea is to bound these sum terms by an upper bound which depends on the existing spread of probability mass in estimates $\hat{H}^b_{n,A,V}$. This is done by checking the following condition for $k \in \{1, \ldots, n\}$:

$$\frac{\hat{H}^b_{n,A,V}(s_{(k)}) - \hat{H}^b_{n,A,V}(s_{(k-1)})}{s_{(k)}} > M \max_{l \in \{1,\ldots,n\}} \left(\hat{H}^b_{n,A,V}(s_{(l)}) - \hat{H}^b_{n,A,V}(s_{(l-1)})\right). \tag{4.15}$$

This condition has been tested in several simulations and the factor $M$ on the right-hand side is chosen to be $M = 3$. When testing the condition in simulations, this value of $M$ ensured that the condition was only satisfied in select estimates $\hat{H}^b_{n,A,V}$, at values $s_{(k)}$ close to zero with relatively high probability mass $\hat{H}^b_{n,A,V}(s_{(k)}) - \hat{H}^b_{n,A,V}(s_{(k-1)})$, without being satisfied in estimates $\hat{H}^b_{n,A,V}$ where the de-biasing process did not need to be adjusted any further. Hence, the condition in Equation 4.15 is a way to detect exactly the values of $s_{(k)}$ with the potential to cause inflated estimates $\hat{H}_{n,A,V}$. To prevent their respective terms from being in the sum of Equation 4.14 altogether, the truncation parameter $\hat{t}_{n,A,V}$ is simply chosen to be greater than any $s_{(k)}$ that satisfy the condition.

Let $\tilde{s}_{(1)}, \ldots, \tilde{s}_{(m)}$ be the ordered statistics of $s_1, \ldots, s_n$, limited to values strictly greater than any that satisfy Equation 4.15. Note that $m \leq n$, with equality possible only if none of the values satisfy the above condition. Then, the adjusted choice $\hat{t}_{n,A,V}$ for $t_n$ is given by:

$$\hat{t}_{n,A,V} := \arg\min_{t \in \{\tilde{s}_{(1)},\ldots,\tilde{s}_{(m)}\}} \sum_{v=1}^{\infty} \int_0^{\infty} \left| \hat{F}_n^{S,V}(s,v;t) - \overline{F}_n^{S,V}(s,v) \right| \mathrm{d}s, \tag{4.16}$$

with $\hat{F}_n^{S,V}$ and $\overline{F}_n^{S,V}$ as in Equation 4.12. The right side of Figure 4.2 shows how this choice of the truncation parameter $\hat{t}_{n,A,V}$ as in Equation 4.16 improves several estimates in simulations, when compared the same estimates in the left of the figure, which use the previous choice of $\hat{t}_{n,A,V}$ as in Equation 4.13.

Similar, partially improved results have been observed in simulations using other particles $K$, sample sizes $n$ and true distributions $H$.

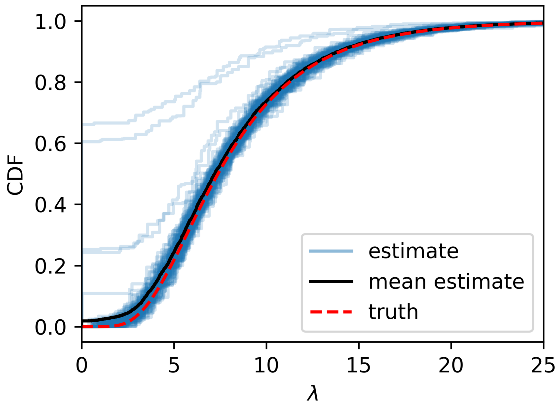


Figure 4.2: Per panel, $100$ truncated estimates $\hat{H}_{n,A,V}$ (blue) and their point-wise mean (black), each resulting from the same $100$ length-biased estimates $\hat{H}^b_{n,A,V}$. Based on a sample of size $n = 1000$, using the cube particle $K$ and true $H$ (red) is a log-normal distribution. Left: inflated estimates $\hat{H}_{n,A,V}$ with truncation parameter given by Equation 4.13, truncated too early. Right: estimates with proposed truncation parameter given by Equation 4.16.

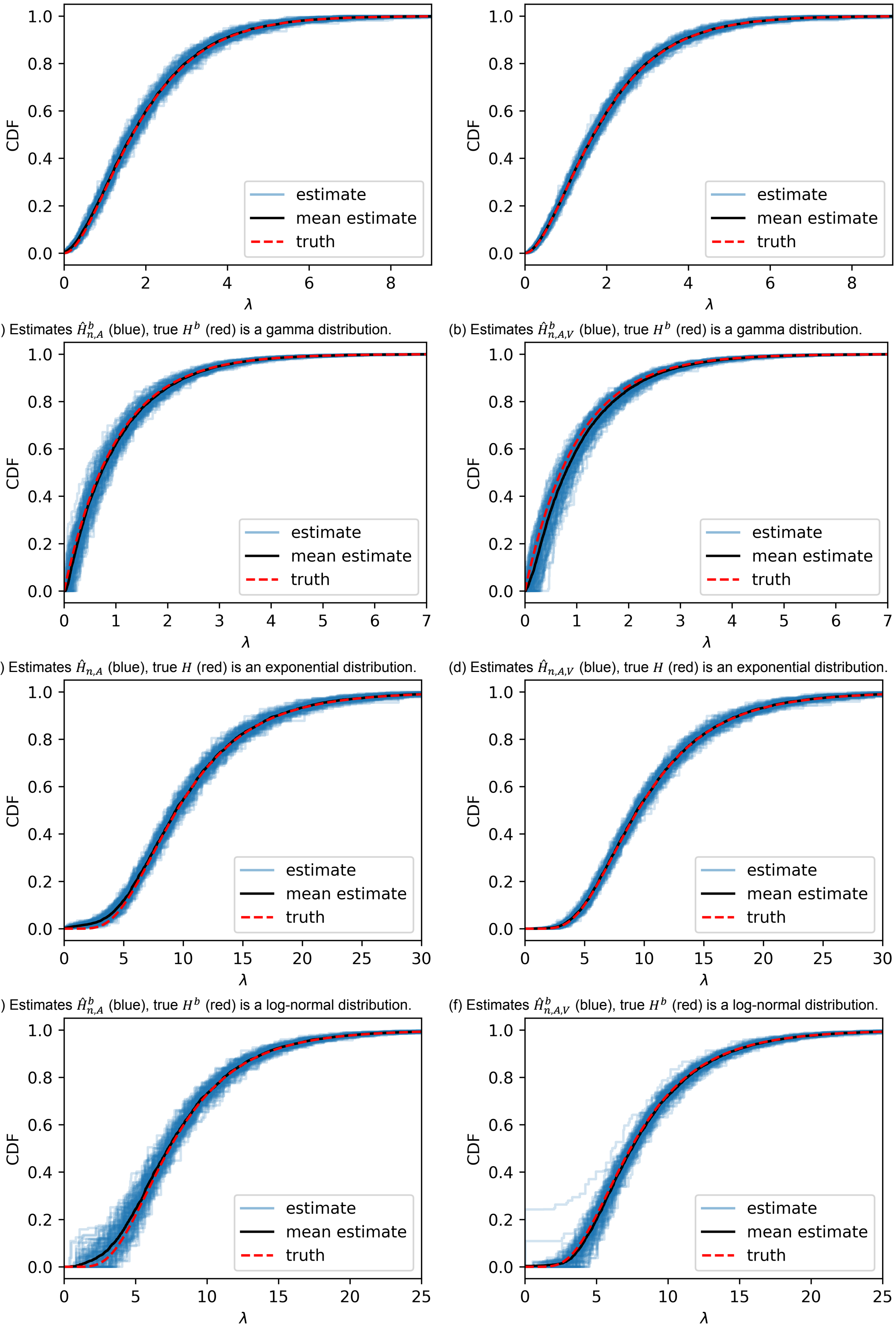
It should be noted that this adjustment in choosing the truncation parameter is not a general solution to the problem of poorly performing estimates of the true particle size distribution $H$, and only serves as a way to show results that estimate $H$ well most of the time. Moreover, the choice of factor $M$, although motivated by results in simulations, is subjective and questionable. A more robust technique to de-bias estimates of $H^b$ to estimates of $H$ is desired, but is beyond the scope of this thesis.

## 4.2. Simulation Results

Following the procedure from Paragraph 4.1, repeated $100$ times for each chosen combination of $n$, $H$ and $K$, simulations first result in $100$ ML estimates of the length-biased particle size distribution $H^b$ for each method, determined by the hybrid ICM-EM algorithm. The corresponding ML estimate of the true size particle distribution $H$ is also computed for each method, following Equation 4.9, using the respective estimate of $H^b$ truncated at $t_n$ either as in Equation 4.11 or as in Equation 4.16. All simulations are performed using the code found at https://github.com/JeroenFaas/adapted-pysizeunfolder. Note that much of this code is adapted or directly reused from https://github.com/thomasvdj/pysizeunfolder, which is used in Van der Jagt et al. (2024) to simulate the estimation method based only on section profile areas.

Figure 4.3 visualises such simulation results for both methods using the cube particle $K$, different true size distributions $H$ and sample size $n = 1000$. Each blue line corresponds to one of the $100$ estimates of either $H$ or $H^b$, the black line is their point-wise mean and the red line is the corresponding true distribution $H$ or $H^b$. The figure shows that the different length-biased size distributions $H^b$ are estimated well using either method, with estimates $\hat{H}^b_{n,A,V}$ performing slightly better than estimates $\hat{H}^b_{n,A}$. Meanwhile, corresponding estimates of the true size distributions $H$ appear to approximate the true distributions less closely. Estimates $\hat{H}_{n,A}$ seem to perform slightly better than estimates $\hat{H}_{n,A,V}$ when the true $H$ is exponentially distributed, but the opposite appears to hold when $H$ is log-normally distributed, with the exception of one outlying estimate $\hat{H}_{n,A,V}$. Note that, in each panel, the mean of estimates follows the true distribution closely.

Similarly, Figures A.1 and A.2 in the Appendix show simulation results for both methods using the dodecahedron and tetrahedron particle $K$, respectively, different true size distributions $H$ and sample size $n = 1000$. The estimates for both $H^b$ and $H$ behave similarly to those for the cube particle, as described above. Using the dodecahedron particle, however, resulting deviations from the true distributions appear overall less extreme, with respect to those using the cube, while such deviations appear more extreme and more frequently using the tetrahedron particle.



(a) Estimates $\hat{H}^b_{n,A}$ (blue), true $H^b$ (red) is a gamma distribution.

(b) Estimates $\hat{H}^b_{n,A,V}$ (blue), true $H^b$ (red) is a gamma distribution.

(c) Estimates $\hat{H}_{n,A}$ (blue), true $H$ (red) is an exponential distribution.

(d) Estimates $\hat{H}_{n,A,V}$ (blue), true $H$ (red) is an exponential distribution.

(e) Estimates $\hat{H}^b_{n,A}$ (blue), true $H^b$ (red) is a log-normal distribution.

(f) Estimates $\hat{H}^b_{n,A,V}$ (blue), true $H^b$ (red) is a log-normal distribution.

(g) Estimates $\hat{H}_{n,A}$ (blue), true $H$ (red) is a log-normal distribution.

(h) Estimates $\hat{H}_{n,A,V}$ (blue), true $H$ (red) is a log-normal distribution.

Figure 4.3: $100$ ML estimates of distribution $H^b$ and corresponding estimates of distribution $H$ based on a sample size $n = 1000$ and the cube particle $K$. a-d: estimates computed by both methods and compared side by side for each row, from the same sample with an exponentially distributed true $H$. e-h: similar, from the same sample with a log-normally distributed true $H$.

In particular, simulations using the tetrahedron $K$ and log-normally distributed $H$ still tend to result in several inflated estimates $\hat{H}_{n,A,V}$. In this case, however, there are enough good estimates for the mean of estimates $\hat{H}_{n,A,V}$ to remain close to the true distribution $H$, more so than the mean of estimates $\hat{H}_{n,A}$.

Overall, estimates of $H^b$ seem to perform well, computed by either method. Estimates $\hat{H}^b_{n,A,V}$ do appear slightly better than $\hat{H}^b_{n,A}$ in all simulations. Corresponding de-biased estimates of $H$ have differing performances, since either method seems to have simulations where its resulting estimates outperform estimates of the other method.

## 4.3. Comparing Methods of Estimation

### 4.3.1. Overall Comparison of Estimates

To quantify the performance of each individual estimate, the *error* of the ML estimate of $H^b$ is chosen to be defined as the supremum of the point-wise distance between the true $H^b$ and the estimate $\hat{H}^b_n$. The error of the ML estimate of $H$ is defined in a similar way. This definition of the estimate error corresponds to $\|\hat{H}^b_n - H^b\|_\infty$ and similarly $\|\hat{H}_n - H\|_\infty$, the infinity (or supremum) norm of the difference between the estimated and true distribution functions.

Given the errors of all $100$ estimates of $H^b$ and $H$ computed by the same method for any chosen combination of $n$, $H$ and $K$, the mean error of all estimates is computed, as well as the $2.5\%$- and $97.5\%$-quantiles of the errors. For all performed simulations, resulting errors for each method and combination of $n$ and true size distribution $H$ are shown in Table 4.1 for the dodecahedron, in Table 4.2 for the cube and in Table 4.3 for the tetrahedron as particle $K$.

The tables are split in two parts, the top part contains estimate errors of the length-biased size distribution $H^b$ and the bottom part contains the corresponding estimate errors of the size distribution $H$. The combinations of $n$ and $H$ are mentioned in the left column, and the row of each combination contains the mean errors and quantiles of the errors resulting from the corresponding simulations of each method. In the middle column, results are from estimates $\hat{H}^b_{n,A}$ and $\hat{H}_{n,A}$, which are based only on the section profile areas, while results in the right column are from estimates $\hat{H}^b_{n,A,V}$ and $\hat{H}_{n,A,V}$, based on the pairs of section profile area and the number of vertices. Note that, for each row, the different estimates resulting from the two methods are based on the same $100$ samples of observed section profiles. This allows for performance comparisons of the different estimation methods, by comparing between the errors in the middle and right columns.

| Estimates of $H^b$ | | $\|\hat{H}^b_{n,A} - H^b\|_\infty$ | | $\|\hat{H}^b_{n,A,V} - H^b\|_\infty$ | |
|---|---|---|---|---|---|
| $n$ | $H$ | mean error | $(2.5\%, 97.5\%)$ | mean error | $(2.5\%, 97.5\%)$ |
| $1000$ | *Exponential* | $0.0593$ | $(0.0393, 0.0924)$ | $0.0466$ | $(0.0309, 0.0671)$ |
| $1000$ | *Log-normal* | $0.0683$ | $(0.0453, 0.0940)$ | $0.0540$ | $(0.0355, 0.0752)$ |
| $2000$ | *Exponential* | $0.0448$ | $(0.0324, 0.0621)$ | $0.0352$ | $(0.0243, 0.0501)$ |
| $2000$ | *Log-normal* | $0.0530$ | $(0.0379, 0.0739)$ | $0.0402$ | $(0.0279, 0.0604)$ |
| $5000$ | *Exponential* | $0.0314$ | $(0.0246, 0.0422)$ | $0.0240$ | $(0.0180, 0.0339)$ |
| $5000$ | *Log-normal* | $0.0384$ | $(0.0290, 0.0499)$ | $0.0283$ | $(0.0198, 0.0371)$ |
| $10000$ | *Exponential* | $0.0247$ | $(0.0189, 0.0346)$ | $0.0187$ | $(0.0133, 0.0261)$ |
| $10000$ | *Log-normal* | $0.0299$ | $(0.0235, 0.0410)$ | $0.0214$ | $(0.0162, 0.0286)$ |
| Estimates of $H$ | | $\|\hat{H}_{n,A} - H\|_\infty$ | | $\|\hat{H}_{n,A,V} - H\|_\infty$ | |
| $n$ | $H$ | mean error | $(2.5\%, 97.5\%)$ | mean error | $(2.5\%, 97.5\%)$ |
| $1000$ | *Exponential* | $0.1196$ | $(0.0685, 0.2076)$ | $0.1286$ | $(0.0439, 0.2509)$ |
| $1000$ | *Log-normal* | $0.0908$ | $(0.0541, 0.1581)$ | $0.0706$ | $(0.0428, 0.1396)$ |
| $2000$ | *Exponential* | $0.0962$ | $(0.0566, 0.1634)$ | $0.0871$ | $(0.0346, 0.1911)$ |
| $2000$ | *Log-normal* | $0.0762$ | $(0.0441, 0.1272)$ | $0.0518$ | $(0.0294, 0.0918)$ |
| $5000$ | *Exponential* | $0.0712$ | $(0.0408, 0.1136)$ | $0.0726$ | $(0.0231, 0.1362)$ |
| $5000$ | *Log-normal* | $0.0516$ | $(0.0339, 0.0821)$ | $0.0390$ | $(0.0237, 0.0586)$ |
| $10000$ | *Exponential* | $0.0578$ | $(0.0315, 0.0976)$ | $0.0530$ | $(0.0188, 0.0882)$ |
| $10000$ | *Log-normal* | $0.0404$ | $(0.0285, 0.0602)$ | $0.0259$ | $(0.0174, 0.0354)$ |

Table 4.1: Estimate errors resulting from simulations using the dodecahedron particle $K$.

| Estimates of $H^b$ | | $\lVert\hat{H}^b_{n,A} - H^b\rVert_\infty$ | | $\lVert\hat{H}^b_{n,A,V} - H^b\rVert_\infty$ | |
|---|---|---|---|---|---|
| $n$ | $H$ | mean error | $(2.5\%, 97.5\%)$ | mean error | $(2.5\%, 97.5\%)$ |
| $1000$ | *Exponential* | $0.0683$ | $(0.0500, 0.1027)$ | $0.0549$ | $(0.0373, 0.0774)$ |
| $1000$ | *Log-normal* | $0.0763$ | $(0.0511, 0.1024)$ | $0.0606$ | $(0.0409, 0.0868)$ |
| $2000$ | *Exponential* | $0.0516$ | $(0.0384, 0.0762)$ | $0.0412$ | $(0.0299, 0.0578)$ |
| $2000$ | *Log-normal* | $0.0625$ | $(0.0469, 0.0842)$ | $0.0481$ | $(0.0346, 0.0641)$ |
| $5000$ | *Exponential* | $0.0369$ | $(0.0285, 0.0472)$ | $0.0290$ | $(0.0215, 0.0405)$ |
| $5000$ | *Log-normal* | $0.0455$ | $(0.0352, 0.0584)$ | $0.0333$ | $(0.0236, 0.0440)$ |
| $10000$ | *Exponential* | $0.0310$ | $(0.0238, 0.0392)$ | $0.0228$ | $(0.0167, 0.0309)$ |
| $10000$ | *Log-normal* | $0.0364$ | $(0.0270, 0.0491)$ | $0.0265$ | $(0.0196, 0.0367)$ |

| Estimates of $H$ | | $\lVert\hat{H}_{n,A} - H\rVert_\infty$ | | $\lVert\hat{H}_{n,A,V} - H\rVert_\infty$ | |
|---|---|---|---|---|---|
| $n$ | $H$ | mean error | $(2.5\%, 97.5\%)$ | mean error | $(2.5\%, 97.5\%)$ |
| $1000$ | *Exponential* | $0.1244$ | $(0.0660, 0.2085)$ | $0.1349$ | $(0.0560, 0.2632)$ |
| $1000$ | *Log-normal* | $0.1088$ | $(0.0651, 0.1832)$ | $0.0870$ | $(0.0453, 0.1690)$ |
| $2000$ | *Exponential* | $0.1048$ | $(0.0551, 0.1780)$ | $0.1070$ | $(0.0414, 0.2210)$ |
| $2000$ | *Log-normal* | $0.0896$ | $(0.0538, 0.1493)$ | $0.0696$ | $(0.0417, 0.1382)$ |
| $5000$ | *Exponential* | $0.0782$ | $(0.0423, 0.1245)$ | $0.0829$ | $(0.0301, 0.1446)$ |
| $5000$ | *Log-normal* | $0.0605$ | $(0.0411, 0.0877)$ | $0.0471$ | $(0.0270, 0.0700)$ |
| $10000$ | *Exponential* | $0.0663$ | $(0.0377, 0.1058)$ | $0.0527$ | $(0.0218, 0.1196)$ |
| $10000$ | *Log-normal* | $0.0478$ | $(0.0315, 0.0707)$ | $0.0342$ | $(0.0219, 0.0424)$ |

Table 4.2: Estimate errors resulting from simulations using the cube particle $K$.

When comparing the estimate errors resulting from both methods, it is clear that estimates $\hat{H}^b_{n,A,V}$ of the length-biased size distribution $H^b$ perform better than estimates $\hat{H}^b_{n,A}$ in the same simulation. This is the case for all simulations performed, and can be seen by reduced mean errors, as well as reduced $2.5\%$- and $97.5\%$-quantiles in all rows of the top parts of Tables 4.1, 4.2 and 4.3. The improvement varies among combinations of $n$, $H$ and $K$, but a reduction is typically between $0.01$ and $0.02$ in terms of their mean error for simulations with $n = 1000$, and the magnitude of this reduction decreases as $n$ increases. Similar reductions in terms of both quantiles can be observed, although these reductions vary more than those in terms of the mean error.

Similarly to what is noted in Paragraph 4.2, estimates $\hat{H}_{n,A,V}$ of the true size distribution $H$ do not always improve over estimates $\hat{H}_{n,A}$ in the same simulation. In terms of mean errors, estimates $\hat{H}_{n,A}$ appear to perform better than estimates $\hat{H}_{n,A,V}$ in simulations using the dodecahedron or cube $K$, and the exponentially distributed true $H$. For larger sample sizes $n$, estimates $\hat{H}_{n,A,V}$ improve in performance and by $n = 10000$ perform better than estimates $\hat{H}_{n,A}$. Estimates $\hat{H}_{n,A,V}$ perform better in all other cases, except for simulations using the tetrahedron $K$, the log-normally distributed $H$ and $n \geq 2000$. These particular simulations all appear to have several outlying estimates $\hat{H}_{n,A,V}$, as can be seen by the $97.5\%$-quantiles of their errors.

The $2.5\%$-quantile is lower overall for estimates $\hat{H}_{n,A,V}$ in all simulations. However, in terms of the $97.5\%$-quantile, each method has different simulations where its resulting estimates perform better than estimates of the other method. In simulations using the dodecahedron or cube $K$, this quantile lies higher for estimates $\hat{H}_{n,A,V}$ when the true $H$ is exponentially distributed, but lower for these estimates when the true $H$ is log-normally distributed, compared to estimates $\hat{H}_{n,A}$. Conversely, when using the tetrahedron $K$ instead, estimates $\hat{H}_{n,A}$ have a lower $97.5\%$-quantile with the exponential $H$ than estimates $\hat{H}_{n,A,V}$, which in turn have a lower $97.5\%$-quantile with the log-normal $H$.

| Estimates of $H^b$ | | $\|\hat{H}^b_{n,A} - H^b\|_\infty$ | | $\|\hat{H}^b_{n,A,V} - H^b\|_\infty$ | |
|---|---|---|---|---|---|
| $n$ | $H$ | mean error | $(2.5\%, 97.5\%)$ | mean error | $(2.5\%, 97.5\%)$ |
| 1000 | *Exponential* | 0.0942 | (0.0680, 0.1336) | 0.0808 | (0.0611, 0.1097) |
| 1000 | *Log-normal* | 0.1138 | (0.0840, 0.1610) | 0.0984 | (0.0744, 0.1374) |
| 2000 | *Exponential* | 0.0759 | (0.0553, 0.1016) | 0.0653 | (0.0472, 0.0896) |
| 2000 | *Log-normal* | 0.0956 | (0.0709, 0.1303) | 0.0811 | (0.0607, 0.1079) |
| 5000 | *Exponential* | 0.0614 | (0.0483, 0.0782) | 0.0511 | (0.0395, 0.0650) |
| 5000 | *Log-normal* | 0.0760 | (0.0583, 0.0998) | 0.0653 | (0.0499, 0.0821) |
| 10000 | *Exponential* | 0.0518 | (0.0402, 0.0678) | 0.0431 | (0.0333, 0.0588) |
| 10000 | *Log-normal* | 0.0619 | (0.0474, 0.0806) | 0.0515 | (0.0418, 0.0643) |

| Estimates of $H$ | | $\|\hat{H}_{n,A} - H\|_\infty$ | | $\|\hat{H}_{n,A,V} - H\|_\infty$ | |
|---|---|---|---|---|---|
| $n$ | $H$ | mean error | $(2.5\%, 97.5\%)$ | mean error | $(2.5\%, 97.5\%)$ |
| 1000 | *Exponential* | 0.1981 | (0.0849, 0.4200) | 0.1504 | (0.0623, 0.2897) |
| 1000 | *Log-normal* | 0.1734 | (0.1008, 0.3325) | 0.1500 | (0.0763, 0.4628) |
| 2000 | *Exponential* | 0.1622 | (0.0770, 0.3217) | 0.1279 | (0.0569, 0.2898) |
| 2000 | *Log-normal* | 0.1420 | (0.0860, 0.2832) | 0.1463 | (0.0735, 0.5737) |
| 5000 | *Exponential* | 0.1335 | (0.0621, 0.2825) | 0.0978 | (0.0454, 0.2261) |
| 5000 | *Log-normal* | 0.1069 | (0.0725, 0.1705) | 0.1098 | (0.0574, 0.4025) |
| 10000 | *Exponential* | 0.1110 | (0.0543, 0.2458) | 0.0725 | (0.0361, 0.1645) |
| 10000 | *Log-normal* | 0.0854 | (0.0562, 0.1417) | 0.0987 | (0.0464, 0.5685) |

Table 4.3: Estimate errors resulting from simulations using the tetrahedron particle $K$.

### 4.3.2. Pair-wise Comparison of Estimates

Alternatively to the overall estimate error analysis above, it is useful to directly compare the performance of the two estimation methods in simulations using the same sample. Therefore, the *pair-wise error* is considered for the pair of ML estimates $\hat{H}^b_{n,A}$ and $\hat{H}^b_{n,A,V}$ resulting from the same sample. This pair-wise error is defined as the difference between their individual estimate errors and is quantified by the following difference:

$$\|\hat{H}^b_{n,A,V} - H^b\|_\infty - \|\hat{H}^b_{n,A} - H^b\|_\infty . \tag{4.17}$$

Note that the result of Equation 4.17 is negative if $\hat{H}^b_{n,A,V}$ has a smaller estimate error than $\hat{H}^b_{n,A}$, and positive otherwise. A similar expression $\|\hat{H}_{n,A,V} - H\|_\infty - \|\hat{H}_{n,A} - H\|_\infty$, the *pair-wise error* for the pair of estimates $\hat{H}_{n,A}$ and $\hat{H}_{n,A,V}$ of $H$, is also considered.

The pair-wise errors of all $100$ pairs of estimates of $H^b$ are computed, as well as those of the corresponding pairs of estimates of $H$, with every pair consisting of a ML estimate resulting from each method, using the same sample. This is done for any chosen combination of $n$, $H$ and $K$ and the resulting pair-wise errors are collected. For each combination, the mean, $2.5\%$- and $97.5\%$-quantiles of the $100$ pair-wise errors are listed in Table 4.4 for the dodecahedron, in Table 4.5 for the cube and in Table 4.6 for the tetrahedron as particle $K$.

The lay-out of these tables is comparable to the ones in the previous section. However, data about the pair-wise errors for pairs of estimates of the length-biased particle size distribution $H^b$ is found in the middle column, while the right column contains the same data for corresponding pairs of estimates of the true particle size distribution $H$. Combinations of $n$ and $H$ used in each simulation are listed in the left column, and the row of each combination contains its corresponding mean, $2.5\%$- and $97.5\%$-quantile data per column.

| Pair-wise errors | | $\|\hat{H}^b_{n,A,V} - H^b\|_\infty - \|\hat{H}^b_{n,A} - H^b\|_\infty$ | | $\|\hat{H}_{n,A,V} - H\|_\infty - \|\hat{H}_{n,A} - H\|_\infty$ | |
|---|---|---|---|---|---|
| $n$ | $H$ | mean error | $(2.5\%, 97.5\%)$ | mean error | $(2.5\%, 97.5\%)$ |
| 1000 | *Exponential* | −0.0127 | (−0.0318, 0.0100) | 0.0090 | (−0.1284, 0.1329) |
| 1000 | *Log-normal* | −0.0143 | (−0.0418, 0.0184) | −0.0202 | (−0.0873, 0.0348) |
| 2000 | *Exponential* | −0.0096 | (−0.0228, 0.0064) | −0.0091 | (−0.0853, 0.0889) |
| 2000 | *Log-normal* | −0.0128 | (−0.0319, 0.0068) | −0.0245 | (−0.0823, 0.0189) |
| 5000 | *Exponential* | −0.0074 | (−0.0173, 0.0016) | 0.0014 | (−0.0612, 0.0699) |
| 5000 | *Log-normal* | −0.0101 | (−0.0249, 0.0022) | −0.0126 | (−0.0479, 0.0135) |
| 10000 | *Exponential* | −0.0061 | (−0.0142, 0.0024) | −0.0047 | (−0.0404, 0.0399) |
| 10000 | *Log-normal* | −0.0085 | (−0.0198, 0.0025) | −0.0146 | (−0.0323, 0.0004) |

Table 4.4: Pair-wise errors of estimates resulting from simulations using the dodecahedron particle $K$.

| Pair-wise errors | | $\|\hat{H}^b_{n,A,V} - H^b\|_\infty - \|\hat{H}^b_{n,A} - H^b\|_\infty$ | | $\|\hat{H}_{n,A,V} - H\|_\infty - \|\hat{H}_{n,A} - H\|_\infty$ | |
|---|---|---|---|---|---|
| $n$ | $H$ | mean error | $(2.5\%, 97.5\%)$ | mean error | $(2.5\%, 97.5\%)$ |
| 1000 | *Exponential* | −0.0134 | (−0.0370, 0.0045) | 0.0105 | (−0.1002, 0.1527) |
| 1000 | *Log-normal* | −0.0157 | (−0.0458, 0.0128) | −0.0218 | (−0.1042, 0.0684) |
| 2000 | *Exponential* | −0.0104 | (−0.0301, 0.0079) | 0.0022 | (−0.0963, 0.1300) |
| 2000 | *Log-normal* | −0.0144 | (−0.0340, 0.0093) | −0.0200 | (−0.0849, 0.0459) |
| 5000 | *Exponential* | −0.0079 | (−0.0181, 0.0030) | 0.0047 | (−0.0819, 0.0830) |
| 5000 | *Log-normal* | −0.0123 | (−0.0270, −0.0008) | −0.0134 | (−0.0473, 0.0136) |
| 10000 | *Exponential* | −0.0081 | (−0.0165, −0.0002) | −0.0136 | (−0.0731, 0.0483) |
| 10000 | *Log-normal* | −0.0099 | (−0.0211, 0.0000) | −0.0136 | (−0.0375, 0.0018) |

Table 4.5: Pair-wise errors of estimates resulting from simulations using the cube particle $K$.

| Pair-wise errors | | $\|\hat{H}^b_{n,A,V} - H^b\|_\infty - \|\hat{H}^b_{n,A} - H^b\|_\infty$ | | $\|\hat{H}_{n,A,V} - H\|_\infty - \|\hat{H}_{n,A} - H\|_\infty$ | |
|---|---|---|---|---|---|
| $n$ | $H$ | mean error | $(2.5\%, 97.5\%)$ | mean error | $(2.5\%, 97.5\%)$ |
| 1000 | *Exponential* | −0.0134 | (−0.0416, 0.0109) | −0.0477 | (−0.2921, 0.1029) |
| 1000 | *Log-normal* | −0.0154 | (−0.0571, 0.0246) | −0.0234 | (−0.2183, 0.3005) |
| 2000 | *Exponential* | −0.0106 | (−0.0330, 0.0087) | −0.0343 | (−0.2108, 0.1315) |
| 2000 | *Log-normal* | −0.0145 | (−0.0466, 0.0106) | 0.0043 | (−0.1276, 0.4354) |
| 5000 | *Exponential* | −0.0102 | (−0.0268, 0.0041) | −0.0358 | (−0.2240, 0.0893) |
| 5000 | *Log-normal* | −0.0107 | (−0.0344, 0.0083) | 0.0029 | (−0.0796, 0.2964) |
| 10000 | *Exponential* | −0.0087 | (−0.0227, 0.0055) | −0.0385 | (−0.1479, 0.0516) |
| 10000 | *Log-normal* | −0.0104 | (−0.0294, 0.0057) | 0.0133 | (−0.0743, 0.4274) |

Table 4.6: Pair-wise errors of estimates resulting from simulations using the tetrahedron particle $K$.

Based on what is shown in the middle columns of Tables 4.4, 4.5 and 4.6, it is clear that estimates $\hat{H}^b_{n,A,V}$ of $H^b$ generally perform better than corresponding estimates $\hat{H}^b_{n,A}$. The mean of pair-wise errors is negative with a magnitude of around $0.01$ in all combinations of $n$, $H$ and $K$. There is variation in this magnitude, and in general it decreases as $n$ increases. The $2.5\%$-quantile of each combination is negative, however, the $97.5\%$-quantile is always positive, with the exception of two combinations using the cube particle $K$ and either $n = 5000$ and $H$ is log-normally distributed, or $n = 10000$ and $H$ is exponentially distributed. These exceptions are the only cases where nearly all estimates $\hat{H}^b_{n,A,V}$ perform better than corresponding estimates $\hat{H}^b_{n,A}$, since even the $97.5\%$-quantile is negative. In all other combinations, the positive $97.5\%$-quantile implies that there are several estimates $\hat{H}^b_{n,A,V}$ that perform worse than the corresponding estimates $\hat{H}^b_{n,A}$ based on the same sample. In such cases, the pair-wise error is mostly small, and decreases in magnitude for larger sample sizes $n$.

Finally, when looking at estimates of $H$ again, it is clear that estimates $\hat{H}_{n,A,V}$ do not all improve over estimates $\hat{H}_{n,A}$, as can be seen in the right columns of Tables 4.4, 4.5 and 4.6. In terms of the pair-wise errors, most of the combinations of $n$, $H$ and $K$ result in negative means. However, a substantial amount of simulations does not yield a negative pair-wise error, and in some combinations the mean is even positive as well. A positive mean of pair-wise errors corresponds to the same combinations where the mean in individual errors is higher for estimates $\hat{H}_{n,A,V}$ than it is for estimates $\hat{H}_{n,A}$, as mentioned in the previous paragraph. This concerns most combinations using the cube or dodecahedron $K$, and the exponentially distributed $H$, or using the tetrahedron $K$ and the log-normally distributed $H$. In terms of quantiles of pair-wise errors, all combinations have a negative $2.5\%$-quantile and a positive $97.5\%$-quantile, even in cases where the $97.5\%$-quantile in the middle column is negative. This once again underlines varying performances of each method when estimating $H$, even within the same combination of $n$, $H$ and $K$.

## 4.4. Conclusions

The established methods from Chapter 3 of estimating the length-biased particle size distribution $H^b$ and the true particle size distribution $H$ have been tested in different simulations based on several chosen combinations of sample size $n$, true distribution $H$ and shape of reference particle $K$, according to the procedure described in Paragraph 4.1. Based on the simulation results in Paragraphs 4.2 and 4.3, estimates $\hat{H}^b_{n,A,V}$ of $H^b$, based on the area as well as the number of vertices in section profiles, generally improve over estimates $\hat{H}^b_{n,A}$, based on section profile areas only. The improvement is a notable, structural reduction in the average of estimate errors across all combinations, which appears to slightly decrease in magnitude as the sample size increases. These results imply that including the number of vertices in the procedure of estimating the length-biased size distribution, improves the resulting estimate on average.

When looking at the pair-wise error between estimates of $H^b$ resulting from the different methods, based on the same sample, it is revealed that the reduction in estimate errors is not present in general. For almost all combinations of $n$, $H$ and $K$, there are samples that yield a better performing estimate $\hat{H}^b_{n,A}$ than an estimate $\hat{H}^b_{n,A,V}$.

Contrarily to this result in the estimation of length-biased distribution $H^b$, results from estimating the true distribution $H$ imply that there is an average improvement when including the number of vertices in the estimation procedure only in some cases. In other cases it leads to worse estimates on average. However, the de-biasing procedure appears to be blamed for this, since estimates $\hat{H}^b_{n,A,V}$ do improve over $\hat{H}^b_{n,A}$ overall, but the estimates $\hat{H}_{n,A,V}$ that follow from those same $\hat{H}^b_{n,A,V}$ do not seem to improve over estimates $\hat{H}_{n,A}$ that follow from the respective estimates $\hat{H}^b_{n,A}$. Questions are raised regarding the procedure to de-bias and regularise estimates of $H^b$ to estimates of $H$, as described in Paragraph 4.1. In particular, the described phenomenon, which results in outlying estimates $\hat{H}_{n,A,V}$ of $H$, persists in simulations, despite the implementation of the suggested adjustment. A general solution to improve the de-biasing procedure is missing.

# 5

# Application to Steel Microstructure

Two procedures for estimating the particle size distribution are established in Chapter 3, and implemented in Chapter 4. In this chapter, both methods are applied to real data. The dataset is described first. A method is established next, which is needed to be able to apply the procedures to the dataset. Then, the dataset is used in both procedures and their performances are compared.

## 5.1. Observing the Steel Microstructure

This dataset is based on real observations of the microstructure of steel grains. It can be found at `https://github.com/JeroenFaas/adapted-pysizeunfolder`, in the files '`sample_data_2d.txt`' and '`sample_data_3d.txt`' of the '`examples`' folder. The two files consist of two kinds of observations. The first is a two-dimensional planar section of the steel, revealing the steel microstructure over an area of $500\ \mu m \times 500\ \mu m$. Section profiles of the intersected grains are observed. This observation is shown on the left in Figure 5.1.

The observation is processed to a form that can be better analysed. As shown on the right in Figure 5.1, an approximate polygonal recreation of this same planar section is made, without losing relevant information. The colours of the profiles in the figure are not relevant for the purposes of this analysis.

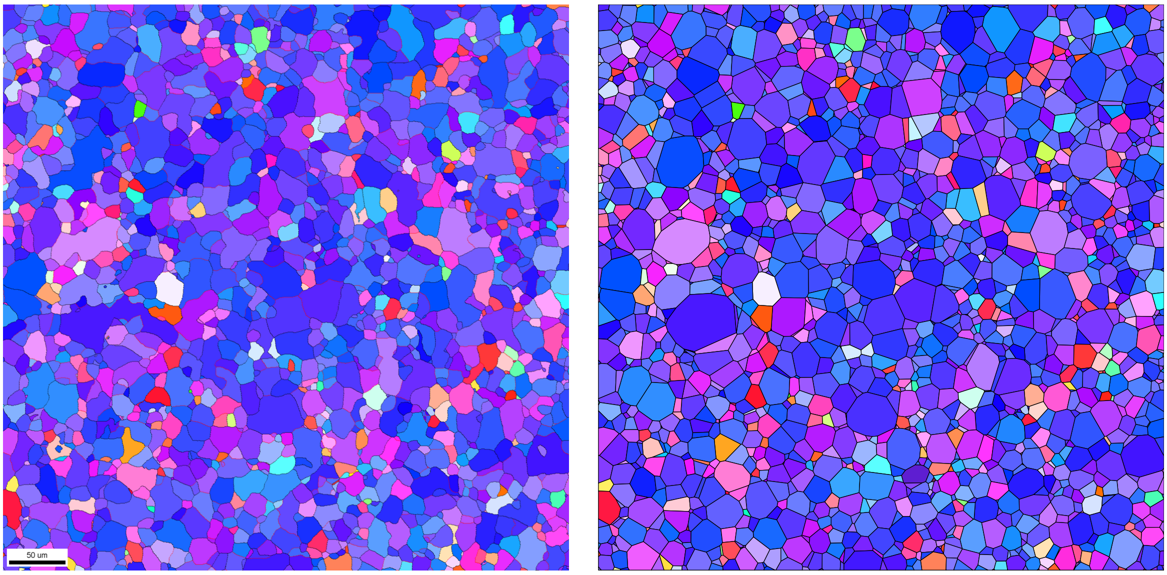
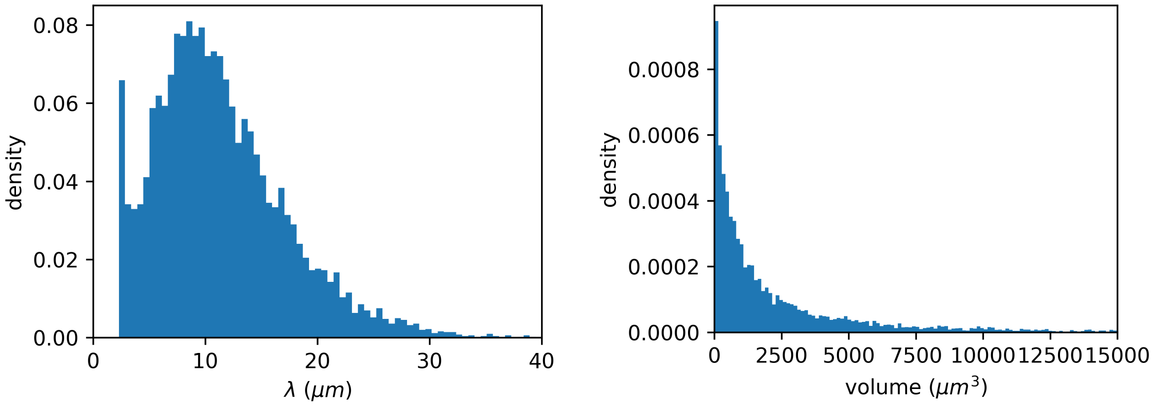


Figure 5.1: Left: Image of the true observed section profiles (source: Van der Jagt et al., 2025). Right: Polygonal recreation of the true observations (provided by T. van der Jagt).

Each polygonal recreation has the same area as the original section profile, and its center of mass is located at the same position as that of the original profile. Each recreation of a section profile has a number of vertices, which is equal to the number of neighbours in the recreation. Note that this is not necessarily the same as the number of neighbours of the original section profile. The data of this polygonal recreation is stored in the first file. Some grain profiles along the boundary of the observed area are only partially visible, making their observed data incomplete. Therefore, data from these profiles is discarded in the rest of this chapter.

Aside from this observation, the same steel microstructure has also been scanned in three dimensions in order to obtain the true volume distribution of the grains. The grain volumes are found in the second file, and are visualised in Figure 5.2(b), measured in cubic micrometers. In the context of the methods developed in Chapters 3 and 4, the grain volumes are also transformed to grain sizes. Grain sizes are defined as a scaling by $\lambda > 0$ in three dimensions with respect to the reference particle $K$, which has a volume of $1\ \mu m^3$. Thus, sizes are obtained by applying a cubic root transformation to the grain volumes. The resulting grain sizes are visualised in Figure 5.2(a), measured in micrometers.



(a) True sizes of the steel dataset, measured in micrometers.

(b) True volumes of the steel dataset, measured in cubic micrometers.

Figure 5.2: Histograms of the known true sizes (a) and volumes (b) of grains in the steel dataset.

## 5.2. Fitting Observations to Known Shapes

In order to be able to apply the methods from Chapters 3 and 4, a shape for reference particle $K$ should be known. However, in this and other applications, that is not a given. It is therefore proposed to take several known candidate reference particles $K$ and choose the best fit for the observed section profiles in the estimation procedures. Candidates are chosen of the tetrahedron, cube and dodecahedron shape. Choosing the best fit shape is done using the distributions of shape-dependent parameter $V$, the number of vertices in a section profile. We would like to fit the observed distribution of $V$ to that in section profiles of each candidate for $K$.

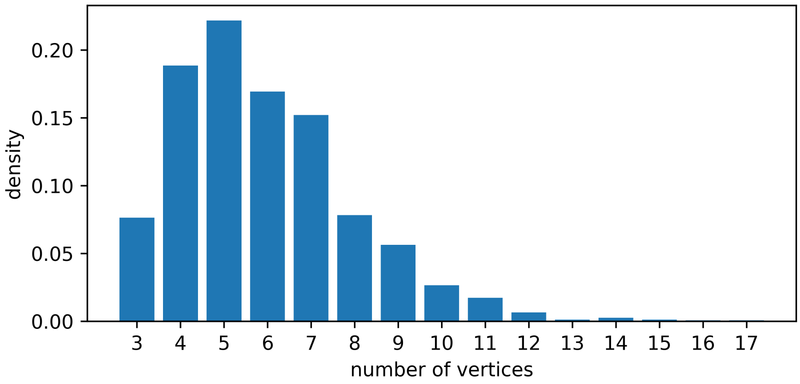


Figure 5.3: Numbers of vertices in the polygonal recreation of the observed grain section profiles.

For the recreation of the grain section profiles, Figure 5.3 shows the observed numbers of vertices. It is clear from this figure that a problem arises when trying to fit the data of observed $V$ to any of the candidate reference particles. Remark that the recreations of section profiles have up to $17$ vertices, while none of the candidates allow for such profiles to occur. One way to resolve this problem is to choose other candidate shapes that do allow up to $17$ vertices in a section profile. Another is to interpret the observed $V$ in a way that works for the chosen candidates, which is done instead. This still raises the questions what to do with the profiles containing more than vertices than the maximum $V_K^{max}$ each candidate allows, and whether profiles with less than or equal to $V_K^{max}$ vertices should even be interpreted as such.

Two solutions are proposed, which interpret the observed $V$ in different ways. These different interpretations of the observed data are justified by the fact that the observed numbers of vertices are based on a polygonal recreation of the true section profiles. These polygonal approximations of section profiles have numbers of vertices that do not necessarily describe the shape of the true section profiles, and are therefore prone to having noise in their values of $V$. Thus, for each candidate $K$:

- the first interpretation is to treat any observed section profile $(a, v)$ with $v$ greater than $V_K^{max}$ as a section profile with $v = V_K^{max}$, and does not change observations otherwise. The resulting distribution of numbers of vertices are shown in Figure 5.4, for all candidate reference particles $K$;

- the second interpretation is to translate the distribution of observed $V$, denoted by $F^V$, to fit the distribution of $V$ in section profiles of $K$, denoted by $G_K^V$. For each $v \in \{3, \dots, 17\}$, the idea is to find the $w_v \in \{3, \dots, V_K^{max}\}$ that minimises the absolute distance between the observed distribution of $V$ at $v$ and that in section profiles of $K$. Formally, for each $v \in \{3, \dots, 17\}$, this $w_v$ is defined as:

$$w_v := \arg\min_{w \in \{3,\dots,V_K^{max}\}} \left| F^V(v) - G_K^V(w) \right|.$$

  Then, each observed section profile with $v$ vertices is treated as having $w_v$ vertices instead. This process is done to the observations for all candidate reference particles $K$, and the resulting distribution of translated numbers of vertices is shown in Figure 5.5.

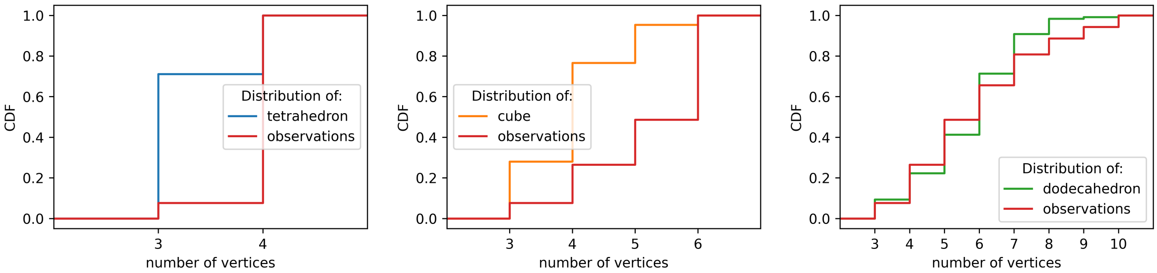


Figure 5.4: Distributions of observed data of number of vertices in a recreated section profile compared to that of each of the candidate reference particles $K$, adjusted according to the first interpretation, which interprets observed $v \geq V_K^{max}$ as $v = V_K^{max}$. Made using the tetrahedron (left), cube (middle) and dodecahedron (right) candidates $K$.

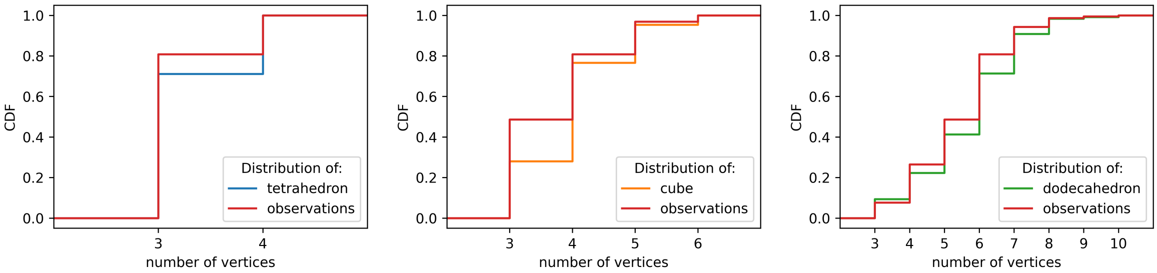


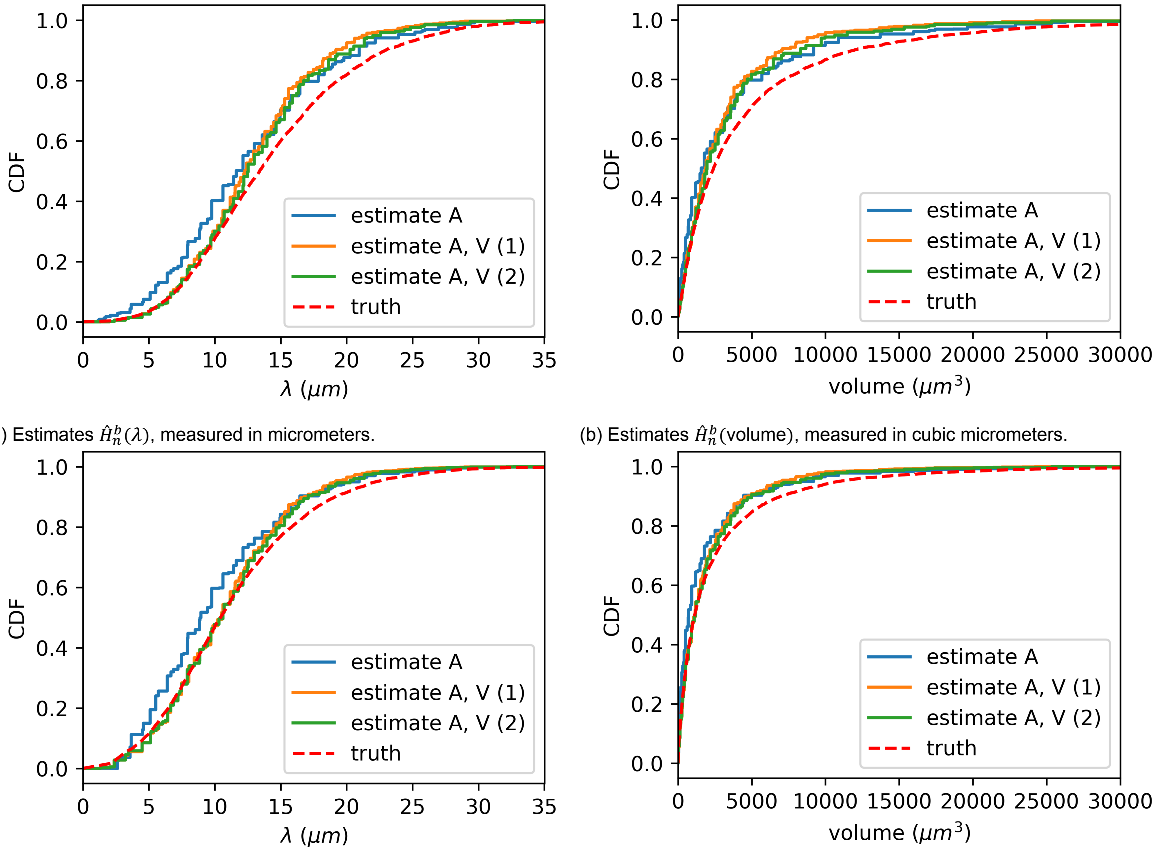
Figure 5.5: Distributions of observed data of number of vertices in a recreated section profile compared to that of each of the candidate reference particles $K$, adjusted according to the second interpretation, which translates the distribution of observed $v$ to that of the candidate $K$. Made using the tetrahedron (left), cube (middle) and dodecahedron (right) candidates $K$.

For each interpretation of the observations and each candidate reference particle $K$, the maximum absolute deviation between the distribution of $V$ corresponding to the candidate and the distribution of $V$ of interpreted observations is computed. Using the first interpretation of the observed data, this deviation for the tetrahedron $K$ is $0.6349$, the greatest out of all candidates. The cube follows next, with a deviation of $0.5016$. It is the smallest for the dodecahedron candidate $K$, with a maximum absolute deviation of $0.1002$. This means that the dodecahedron is the best candidate for the observed dataset. This result is also clearly shown in Figure 5.4, where the distribution corresponding to the dodecahedron candidate follows the observed distribution most closely.

Similarly, the maximum absolute deviations are also computed between the distribution of $V$ for each candidate reference particle $K$ and the observed distribution of $V$, following the second interpretation. This time, the deviation is greatest for the cube candidate $K$, at $0.2066$. The other two candidates have similar deviations, with the tetrahedron $K$ having $0.0968$ and the dodecahedron $K$ having a maximum absolute deviation of $0.0948$. Although close this time, the dodecahedron turns out to be the best candidate for the observed data again. This result is difficult to see in Figure 5.5, due to the small difference in deviations between the tetrahedron and dodecahedron.

## 5.3. Estimating the Grain Volume Distribution

Following each method as described in Chapter 3 and implemented in Chapter 4, ML estimates $\hat{H}^b_{n,A}$ and $\hat{H}^b_{n,A,V}$ are computed based on the observed section profiles. The observed data is adjusted according to each interpretation from Paragraph 5.2. Note that the estimation procedure to obtain $\hat{H}^b_{n,A}$ and $\hat{H}_{n,A}$ does not depend on the number of vertices in a section profile, and will therefore not change between the different interpretations of $V$ in the observations.



(a) Estimates $\hat{H}^b_n(\lambda)$, measured in micrometers.

(b) Estimates $\hat{H}^b_n$(volume), measured in cubic micrometers.

(c) Estimates $\hat{H}_n(\lambda)$, measured in micrometers.

(d) Estimates $\hat{H}^b_n$(volume), measured in cubic micrometers.

Figure 5.6: Estimates of the true $H$ and length-biased $H^b$ grain size and volume distributions (red), computed by all methods using the dodecahedron reference particle $K$. Observations are adjusted according to the first interpretation in estimates (1), shown in orange, and according to the second interpretation in estimates (2), shown in green.

At first, the estimates are computed using the dodecahedron as reference particle $K$, since it has been established in Paragraph 5.2 to be the most fitting candidate in either interpretation of the observations. Figure 5.6 shows the resulting estimates of each estimation method using the dodecahedron $K$.

The estimated distribution functions in Figure 5.6 show that all estimates approximate the true size and volume distributions well. Estimates $\hat{H}^b_{n,A}$ and $\hat{H}_{n,A}$, based on the method using only the square root transformed areas of section profiles, shown in blue, tend to consistently overestimate the true distributions. Meanwhile, estimates $\hat{H}^b_{n,A,V}$ and $\hat{H}_{n,A,V}$, based on the method that uses pairs of the square root transformed area and the number of vertices in a section profile, appear much more accurate for low values of both size and volume. This holds in either interpretation of the observations, with the estimate resulting from the first interpretation shown in orange and the estimate resulting from the second interpretation shown in green. For higher values of size and volume, however, the estimates $\hat{H}^b_{n,A,V}$ and $\hat{H}_{n,A,V}$ also overestimate by approximate equal amounts as the estimates $\hat{H}^b_{n,A}$ and $\hat{H}_{n,A}$ when using the second interpretation of the observations. This can be observed in the orange functions in Figure 5.6. Moreover, when the observations are adjusted according to the first interpretation, estimates $\hat{H}^b_{n,A,V}$ and $\hat{H}_{n,A,V}$ perform worse than estimates $\hat{H}^b_{n,A}$ and $\hat{H}_{n,A}$ for higher values of size and volume, as can be seen in the green functions in Figure 5.6.

Similar figures can be found in Appendix A for the other candidate reference particles. Figure A.3 shows the resulting estimates when using the cube $K$, following the first and second interpretation of the observations, respectively. Figure A.4 shows the resulting estimates when using the tetrahedron $K$, following the first and second interpretation of the observations, respectively.

Similarly to Paragraph 4.3, the error of an estimate is considered to be the infinity norm of the difference between the estimated and corresponding true distribution functions. This estimate error is computed for the estimates of both the length-biased distribution $H^b$ and the true distribution $H$, following from each method and each interpretation of the observations. The errors of all estimates are listed in Table 5.1. Note that taking these distributions with respect to sizes or volumes makes no difference in estimate errors, since translating a data point between size and volume does not change its corresponding distribution function value.

| $K$ | Estimates using method & interpretation | | $\lVert\hat{H}^b_n - H^b\rVert_\infty$ | $\lVert\hat{H}_n - H\rVert_\infty$ |
|---|---|---|---|---|
| *Tetrahedron* | method $A$ ($\hat{H}^b_{n,A}$ / $\hat{H}_{n,A}$) | - | 0.1891 | 0.3502 |
| | method $A, V$ ($\hat{H}^b_{n,A,V}$ / $\hat{H}_{n,A,V}$) | 1 | 0.2494 | 0.2303 |
| | method $A, V$ ($\hat{H}^b_{n,A,V}$ / $\hat{H}_{n,A,V}$) | 2 | 0.4121 | 0.6106 |
| *Cube* | method $A$ ($\hat{H}^b_{n,A}$ / $\hat{H}_{n,A}$) | - | 0.0899 | 0.1355 |
| | method $A, V$ ($\hat{H}^b_{n,A,V}$ / $\hat{H}_{n,A,V}$) | 1 | 0.2792 | 0.2336 |
| | method $A, V$ ($\hat{H}^b_{n,A,V}$ / $\hat{H}_{n,A,V}$) | 2 | 0.2912 | 0.4183 |
| *Dodecahedron* | method $A$ ($\hat{H}^b_{n,A}$ / $\hat{H}_{n,A}$) | - | 0.1414 | 0.1449 |
| | method $A, V$ ($\hat{H}^b_{n,A,V}$ / $\hat{H}_{n,A,V}$) | 1 | 0.1432 | 0.0820 |
| | method $A, V$ ($\hat{H}^b_{n,A,V}$ / $\hat{H}_{n,A,V}$) | 2 | 0.1147 | 0.0636 |

Table 5.1: Estimate errors resulting from simulations using the observations in the dataset. Made using different reference particles $K$, listed in the first column, methods of estimation, listed in the second column, and interpretations of the observations, listed in the third column.

Table 5.1 clearly shows that using the dodecahedron reference particle $K$ leads to overall low estimate errors when estimating both $H^b$ and $H$, regardless of estimation method or interpretation of the observations. The lowest of these estimate errors may be found when using estimates $\hat{H}^b_{n,A,V}$ and $\hat{H}_{n,A,V}$, and when using the second interpretation of observations. However, the lowest error when estimating $H^b$ is actually achieved by estimate $\hat{H}^b_{n,A}$ using the cube $K$. This is a result which has also been observed in Van der Jagt et al. (2025), which analyses the same dataset. Estimates based on both the area and the number of vertices in a section profile do not perform well using the cube $K$, in either interpretation of the observations, because the deviation between distributions of $V$ for the dataset and the cube particle is too large. This result emphasises the sensitivity of the latter estimation method to deviations from the distribution of $V$.

Moreover, resulting errors in the table reveal the dependence of the estimation method based on both the section profile area and number of vertices on the right reference particle. This is especially relevant since the used reference particle $K$ determines what $g_K^{S,V}$, the density function of section profiles through $K$, looks like. This function varies greatly among shapes of $K$. This is a dependency that the method based only on the section profile areas does not have, and is what allows it to perform relatively well for any chosen reference particle $K$.

The results listed in the table also make it clear that the different interpretations of the observed numbers of vertices yield very different results. The first interpretation yields estimates that do better than those resulting from the second interpretation when using the tetrahedron or cube $K$, which performs very poorly, especially when using the tetrahedron $K$. On the other hand, when using the dodecahedron $K$, the second interpretation performs best of all methods.

# 6
# Conclusion & Discussion

In this thesis an existing estimation method for the approximation of the three-dimensional particle size distribution $H$ has been explained, which is based on observations of two-dimensional section profile areas. Estimates of the particle size distribution solve the problem described in Chapter 1, and may be used in materials science, for example, to determine the hardness of steel based on the grain size distribution in its microstructure.

This method is then adapted, incorporating the number of vertices as parameter describing the shape of an observed section profile, alongside its area, which describes its size. The intention of including this shape parameter in the estimation method is to see whether or not it would improve resulting estimates.

Chapter 3 describes how each of these procedures yields a maximum likelihood estimator for the length-biased particle size distribution $H^b$. This can then, in theory, be de-biased to obtain an estimate of the true size distribution $H$. However, when implementing the de-bias procedure from Chapter 4, several problems have arisen. The existing estimation method also encounters some problems when de-biasing, but has found an acceptable ‘rule of thumb’ to eliminate these problems and yield accurate estimates of $H$ as well. A similar, accurate de-biasing procedure for the method including the number of vertices of a section profile has not yet been found and is thus left as a suggestion for future research. Therefore, conclusions about the performance of each estimation method are best drawn based on the estimates of $H^b$, so that the different de-bias procedures do not affect outcomes.

Simulations in Chapter 4 have revealed that the estimate of $H^b$, which is based on a sample of both the area and number of vertices in an observed section profile, performs better on average than the estimate of $H^b$ based only on the section profile areas of the same sample. The reduction in estimate errors, as defined in Chapter 4, varies depending on used reference particles and true size distributions, but it is between $0.01$ and $0.02$ on average. This reduction decreases slightly in magnitude when the sample size increases. When examining these errors for each pair of estimates based on the same sample of observations, it is revealed that this reduction in estimate error is not structural.

On the other hand, mixed conclusions may be drawn regarding the performance of estimates of $H$, based on the different estimation methods. Different methods of estimating $H$ perform better in different situations. This has been shown in Chapter 4, using simulations based on various sample sizes, true size distributions and reference particles. Here, the different de-bias procedures should also be taken into account, as they do affect results.

When applying both estimation methods in practice, an additional step may be required. Both methods require a reference particle to be known, which is not always the case in applications. Therefore, Chapter 5 describes how to determine what reference particle best fits the observations out of several known candidates. This decision is based on the distribution of numbers of vertices in the observed data. Another problem is revealed, however, when a number of vertices in the observations cannot possibly be obtained from some or any of the candidate reference particles. To resolve this issue in processing the observations, and to be able to apply the method which incorporates the number of vertices in a section profile to it, two suggestions have been done in Chapter 5. Both provide alternative interpretations of the observations, and are motivated by the presence of noise in the process of obtaining the observed numbers of vertices in section profiles. This noise in the observed number of vertices of a section profile raises the question whether this is a useful shape parameter in practical situations, and depending on the way it is observed, could be a reason not to use the method incorporating its observed data altogether.

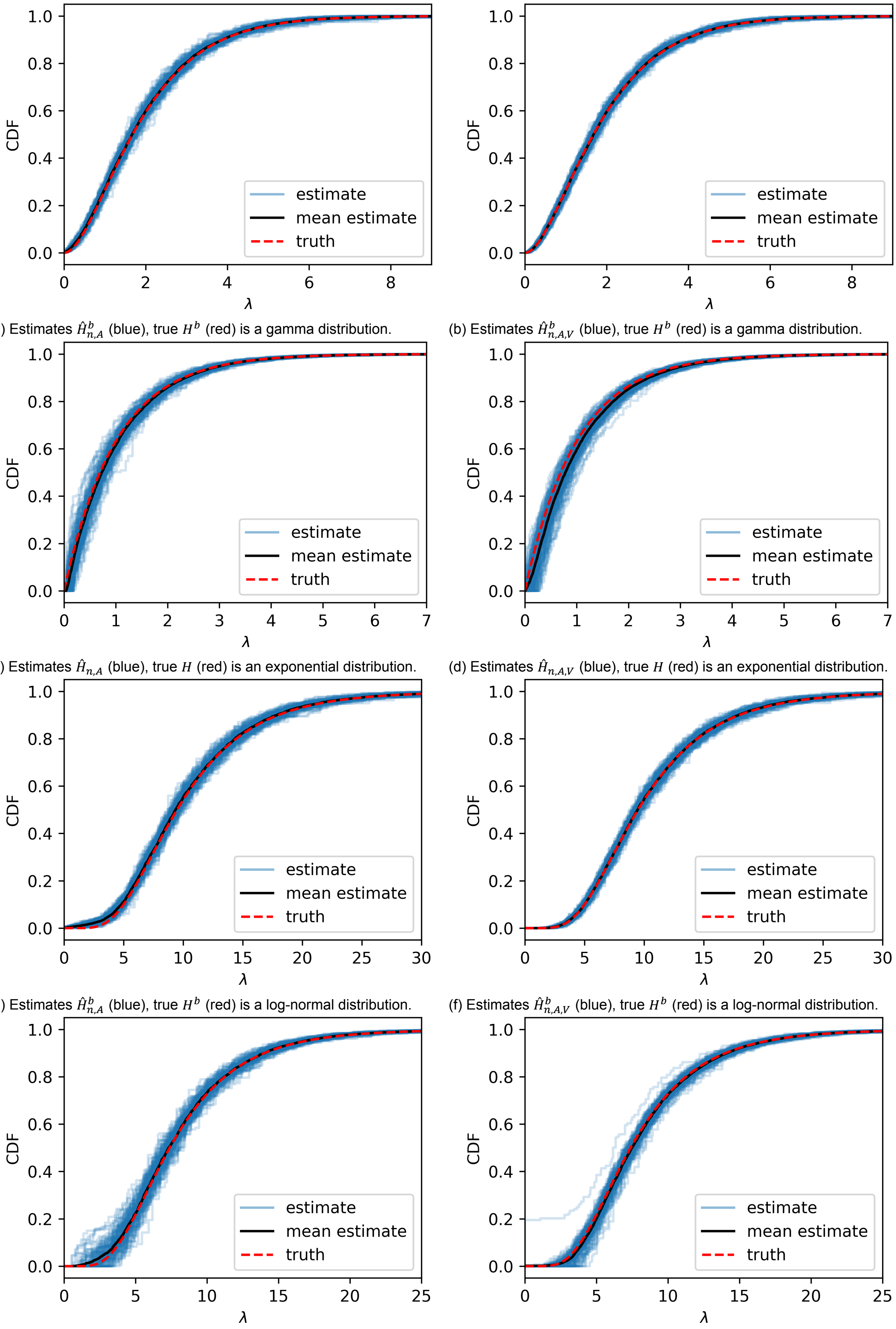
The final results are obtained when applying both estimation methods to a real sample of observed section profiles of grains in a steel microstructure. The true grain volume distribution $H$ is known, and is used to compare the performance of each estimate. The observed numbers of vertices are interpreted according to both suggestions and result into two variants of the estimation method using both the area and the number of vertices in a section profile. The different interpretations do not alter the other estimation method. Thus, estimates resulting from the three methods are compared to each other using the candidate reference particle that best fits the observed distribution of numbers of vertices. The best estimate of the length-biased grain volume distribution $H^b$ is concluded to follow from the method based on both the area and number of vertices in a section profile. However, this is only the best estimate when adjusting the observed numbers of vertices in the second interpretation given in Chapter 5. Moreover, a better estimate of $H^b$ is obtained when using the method based only on the section profile areas, with the cube as a reference particle.

This result highlights the dependence of the method incorporating the number of vertices on an accurate reference particle and the sensitivity of this method to noise in the observed numbers of vertices. The method based only on the section profile areas still requires a reference particle to work, but the accuracy of its estimates is less dependent on this choice and is even independent of the different ways to interpret observed numbers of vertices.

With all of the above taken into account, it is noted that the incorporation of the number of vertices in the estimation procedure might not be useful when this data is not precisely observable. Other parameters to describe the shape of a section profile could be considered in future research instead of the number of vertices, preferably parameters which are less sensitive to the observation process.

In conclusion, the incorporation of observed numbers of vertices improves the estimation procedure of the particle size distribution in theory. On average, the resulting estimator deviates less than the result of the estimator based only on the observed section profile areas, but in general this is not a given. Moreover, the former estimator requires more computations to approximate, resulting in a trade-off between accuracy and computational cost. Therefore, a different method could be preferred depending on the application. In such an application, the estimate based on both the area and number of vertices in a section profile could face more difficulties, depending on the way the numbers of vertices are observed. Since the method is highly sensitive and dependent on this observed data, its application might lead to reduced accuracy when compared to the theoretical gains, and the method based only on the section profile areas, which is less dependent on this data, might be preferred.

# A
# Figures



(a) Estimates $\hat{H}^b_{n,A}$ (blue), true $H^b$ (red) is a gamma distribution.

(b) Estimates $\hat{H}^b_{n,A,V}$ (blue), true $H^b$ (red) is a gamma distribution.

(c) Estimates $\hat{H}_{n,A}$ (blue), true $H$ (red) is an exponential distribution.

(d) Estimates $\hat{H}_{n,A,V}$ (blue), true $H$ (red) is an exponential distribution.

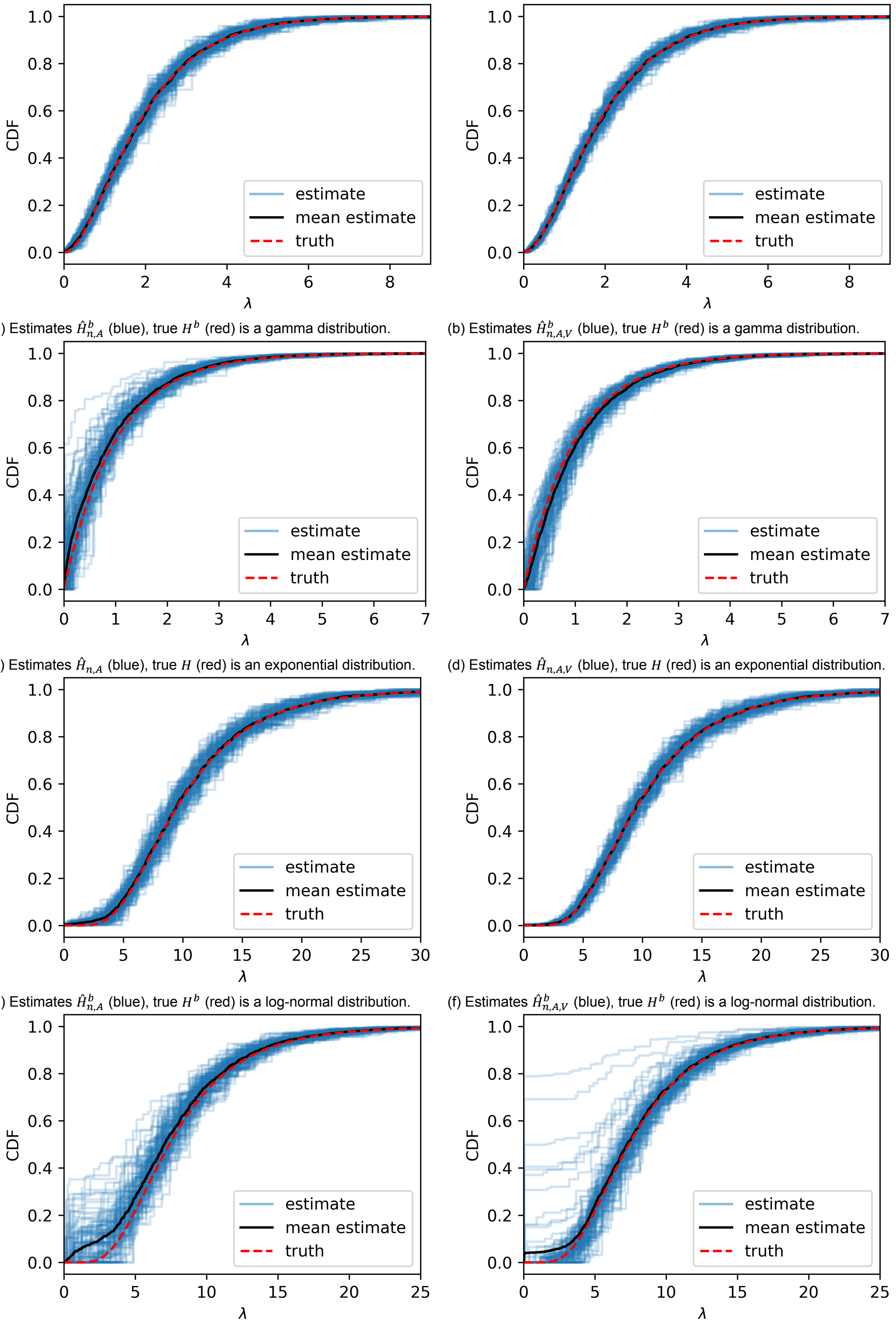
(e) Estimates $\hat{H}^b_{n,A}$ (blue), true $H^b$ (red) is a log-normal distribution.

(f) Estimates $\hat{H}^b_{n,A,V}$ (blue), true $H^b$ (red) is a log-normal distribution.

(g) Estimates $\hat{H}_{n,A}$ (blue), true $H$ (red) is a log-normal distribution.

(h) Estimates $\hat{H}_{n,A,V}$ (blue), true $H$ (red) is a log-normal distribution.

Figure A.1: Maximum likelihood estimates of distribution $H^b$ and corresponding distribution $H$ for the dodecahedron particle $K$.



(a) Estimates $\hat{H}^b_{n,A}$ (blue), true $H^b$ (red) is a gamma distribution.

(b) Estimates $\hat{H}^b_{n,A,V}$ (blue), true $H^b$ (red) is a gamma distribution.

(c) Estimates $\hat{H}_{n,A}$ (blue), true $H$ (red) is an exponential distribution.

(d) Estimates $\hat{H}_{n,A,V}$ (blue), true $H$ (red) is an exponential distribution.

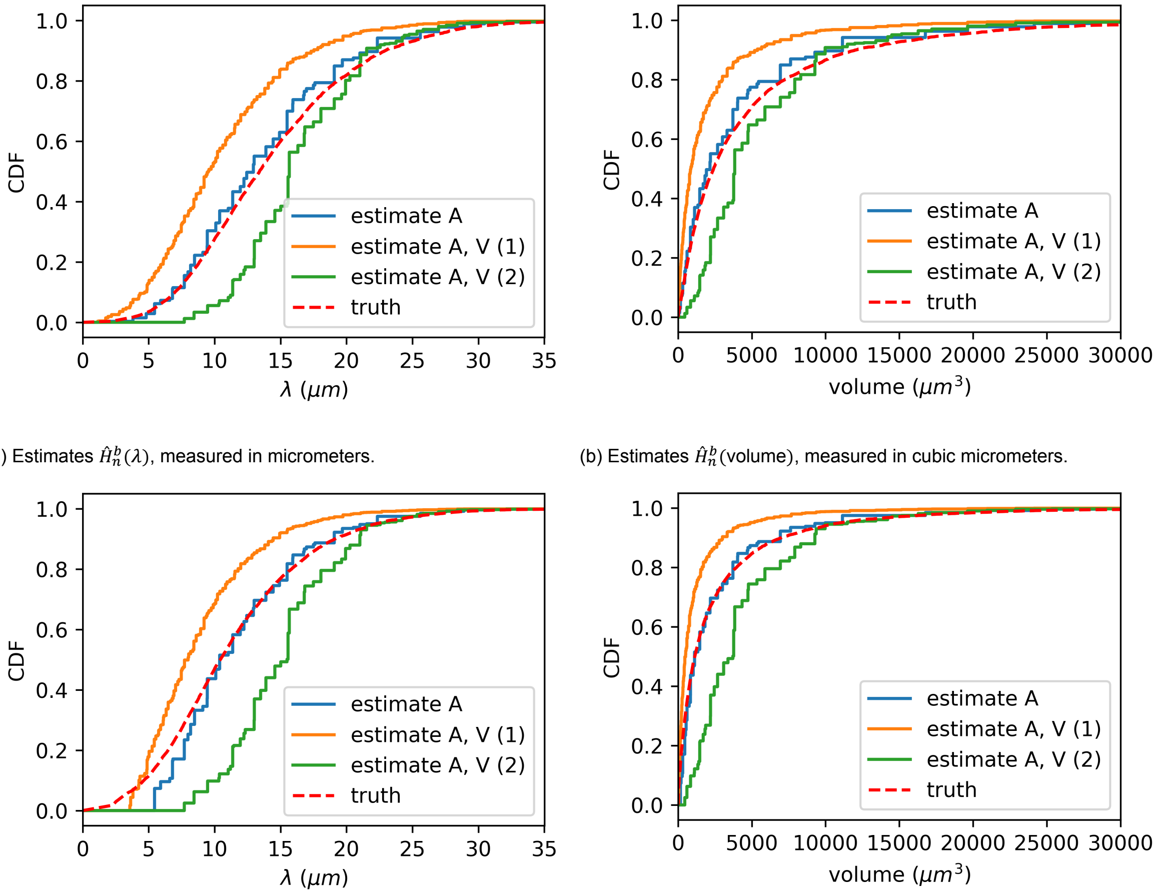
(e) Estimates $\hat{H}^b_{n,A}$ (blue), true $H^b$ (red) is a log-normal distribution.

(f) Estimates $\hat{H}^b_{n,A,V}$ (blue), true $H^b$ (red) is a log-normal distribution.

(g) Estimates $\hat{H}_{n,A}$ (blue), true $H$ (red) is a log-normal distribution.

(h) Estimates $\hat{H}_{n,A,V}$ (blue), true $H$ (red) is a log-normal distribution.

Figure A.2: Maximum likelihood estimates of distribution $H^b$ and corresponding distribution $H$ for the tetrahedron particle $K$.



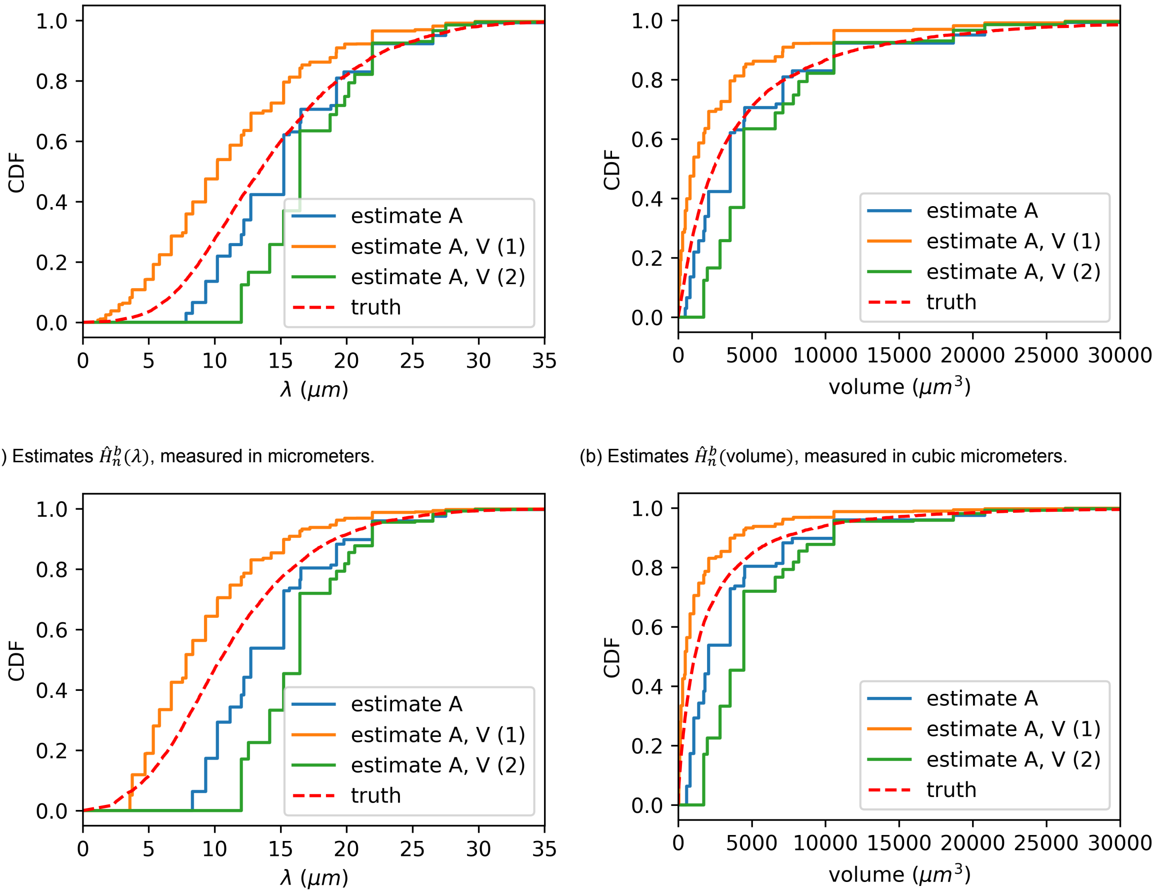
(a) Estimates $\hat{H}_n^b(\lambda)$, measured in micrometers.

(b) Estimates $\hat{H}_n^b$(volume), measured in cubic micrometers.

(c) Estimates $\hat{H}_n(\lambda)$, measured in micrometers.

(d) Estimates $\hat{H}_n^b$(volume), measured in cubic micrometers.

Figure A.3: Estimates of the true $H$ and length-biased $H^b$ grain size and volume distributions (red), computed by all methods using the cube reference particle $K$. Observations are adjusted according to the first interpretation in estimates (1), shown in orange, and according to the second interpretation in estimates (2), shown in green.



(a) Estimates $\hat{H}_n^b(\lambda)$, measured in micrometers.

(b) Estimates $\hat{H}_n^b$(volume), measured in cubic micrometers.

(c) Estimates $\hat{H}_n(\lambda)$, measured in micrometers.

(d) Estimates $\hat{H}_n^b$(volume), measured in cubic micrometers.

Figure A.4: Estimates of the true $H$ and length-biased $H^b$ grain size and volume distributions (red), computed by all methods using the tetrahedron reference particle $K$. Observations are adjusted according to the first interpretation in estimates (1), shown in orange, and according to the second interpretation in estimates (2), shown in green.

# Bibliography

Arratia, R., Goldstein, L., & Kochman, F. (2019). Size bias for one and all. *Probability Surveys*, *16*, 1–61. https://doi.org/10.1214/13-PS221

Jongbloed, G. (1998). The iterative convex minorant algorithm for nonparametric estimation. *Journal of Computational and Graphical Statistics*, *7*(3), 310–321. https://doi.org/10.1080/10618600.1998.10474778

Schuster, E. F. (1985). Incorporating support constraints into nonparametric estimators of densities. *Communications in Statistics - Theory and Methods*, *14*(5), 1123–1136. https://doi.org/10.1080/03610928508828965

Sheather, S. J., & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, *53*(3), 683–690. https://doi.org/10.1111/j.2517-6161.1991.tb01857.x

Van der Jagt, T., Jongbloed, G., & Vittorietti, M. (2023). Existence and approximation of densities of chord length- and cross section area distributions. *Image Analysis and Stereology*, *42*(3), 171–184. https://doi.org/10.5566/ias.2923

Van der Jagt, T., Jongbloed, G., & Vittorietti, M. (2024). Stereological determination of particle size distributions for similar convex bodies. *Electronic Journal of Statistics*, *18*(1), 742–774. https://doi.org/10.1214/24-EJS2215

Van der Jagt, T., Vittorietti, M., Sedighiani, K., Bos, C., & Jongbloed, G. (2025). Estimation of 3d grain size distributions from 2d sections in real and simulated microstructures. *Computational Materials Science*, *256*(1), 113949. https://doi.org/10.1016/j.commatsci.2025.113949

Wellner, J. A., & Zhan, Y. (1997). A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data. *Journal of the American Statistical Association*, *92*(439), 945–959. https://doi.org/10.1080/01621459.1997.10474049

Wicksell, S. D. (1925). The corpuscle problem: A mathematical study of a biometric problem. *Biometrika*, *17*(1-2), 84–99. https://doi.org/10.1093/biomet/17.1-2.84