Correcting estimations of copepod volume from 2-dimensional images
Supporting Information

Cédric Dubois*[1], Jean-Olivier Irisson[2], and Eric Debreuve[1]

[1]Université Côte d'Azur, CNRS, Inria, Équipe Morpheme, Laboratoire I3S, Sophia Antipolis, France. *Corresponding author: cedric.dubois@univ-cotedazur.fr
[2]Sorbonne Université, CNRS, Équipe COMPLEx, Laboratoire LOV, Institut IMEV, Villefranche-sur-mer, France.

Keywords— Zooplankton, Volume, Biomass, Copepod, Geometrical model, *in situ* imaging.

S1 Projection of an ellipsoid

S1.1 Geometrical setup

A centered ellipsoid is defined by all 3-dimensional vectors verifying

$$x^\intercal M x = 1 \tag{14}$$

where M is a positive definite[1] 3×3-matrix whose elements are denoted by m_{ij}. Let (i, j, k) denote an orthonormal basis, and let O denote the origin. To study how this ellipsoid projects onto a plane using perspective projection, let us define (i) an optical center e

$$e = \begin{bmatrix} 0 \\ 0 \\ -\epsilon \end{bmatrix} \tag{15}$$

where $\epsilon > 0$ is such that e is outside the ellipsoid, and (ii) a projection plane Π described by its normal

$$n = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \tag{16}$$

and its distance δ, $\epsilon > \delta > 0$, to the origin such that Π does not intersect the ellipsoid. The plane Π is equipped with the orthonormal basis (u, v) where u and v correspond to i and j respectively. Its origin O_Π is located at the intersection between Π and the segment linking e to O. All these elements are illustrated on Fig. S1.1.

S1.2 Ellipsoid silhouette in 3-D

For some unit vector d, let x be defined as

$$x = e + \tau d \tag{17}$$

with $\tau > 0$ and $d \cdot n > 0$. The ellipsoid silhouette as seen from e is given by the set of vectors d such that the half-line described by x when τ varies is tangent to the ellipsoid[2].
The point x is on the ellipsoid if and only if

$$(d^\intercal M d)\tau^2 + (2d^\intercal M e)\tau + (e^\intercal M e - 1) = 0, \tag{18}$$

[1] positive definite matrices are, by definition, symmetric
[2] for such a vector d, there is a corresponding value for τ

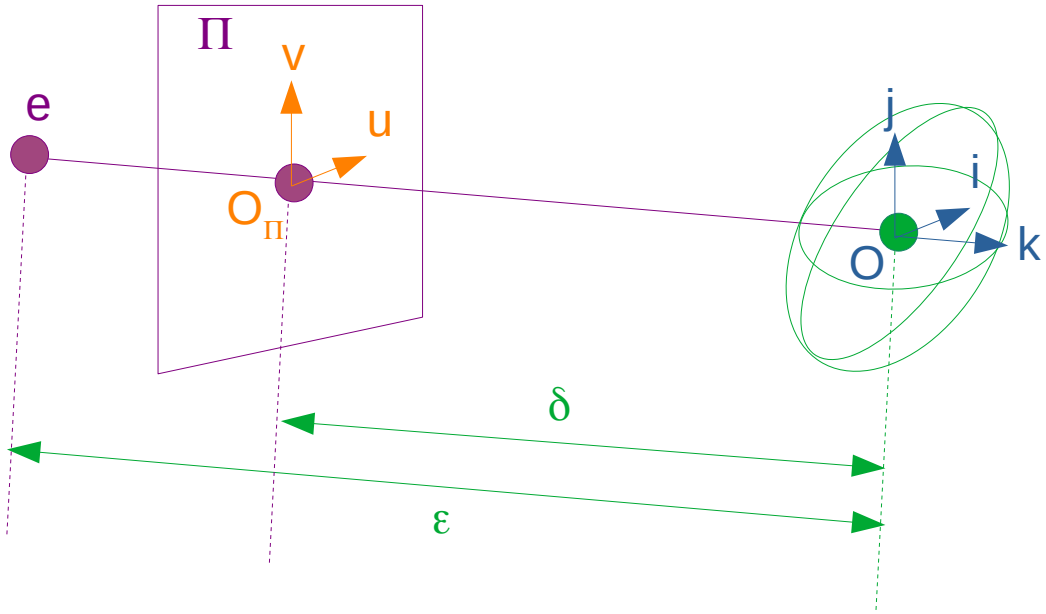

Figure S1.1: Geometrical setup of the ellipsoid model. The camera is represented by its optical center e, the sensor plane Π, and the sensor orthonormal coordinate system (O_Π, u, v). The global orthonormal coordinates system is represented by (O, i, j, k). The axes i and u are parallel, and so are j and v. The ellipsoid center is at distance δ from Π and ϵ from e. Without loss of generality (for our problem), the ellipsoid center is at O, and e is on the axis k with the optical axis aligned with k. Consequently, O_Π is also on the axis k.

which is of the form

$$\alpha\tau^2 + \beta\tau + \gamma = 0. \tag{19}$$

Therefore, the half-line described by x is tangent to the ellipsoid if and only if Eq. (19) has a unique solution[3], that is if and only if $\beta^2 - 4\alpha\gamma = 0$, which is equivalent to

$$d^\intercal S d = 0 \tag{20}$$

where S is equal to

$$S = Me\, e^\intercal M + (1 - e^\intercal Me)M. \tag{21}$$

The ellipsoid silhouette is defined by the solutions of Eq. (20) respecting, as mentioned earlier, the following two conditions: $|d| = 1$ and $d \cdot n > 0$.

Let S_ϵ be defined as

$$S_\epsilon = \frac{1}{\epsilon^2} S. \tag{22}$$

Note that S_ϵ can be used in Eq. (20) in place of S. To propose a more explicit form of S_ϵ, let us use the following block matrix formulation

$$M = \left[\begin{array}{c|c} M_{11} & M_{21}^\intercal \\ \hline M_{21} & m_{33} \end{array}\right] \tag{23}$$

where M_{11} is a 2×2-matrix, M_{21} is a 1×2-vector, and m_{33} is a scalar. Using such a block formulation, we have

$$e\, e^\intercal = \left[\begin{array}{c|c} 0_{11} & 0_{21}^\intercal \\ \hline 0_{21} & \epsilon^2 \end{array}\right] \tag{24}$$

where 0_{ij} denotes a matrix of zeros matching the dimension of M_{ij}. Then

$$Me\, e^\intercal M = \epsilon^2 \left[\begin{array}{c|c} M_{21}^\intercal M_{21} & m_{33} M_{21}^\intercal \\ \hline m_{33} M_{21} & m_{33}^2 \end{array}\right]. \tag{25}$$

Similarly

$$e^\intercal Me = \epsilon^2 m_{33}. \tag{26}$$

Therefore

$$S_\epsilon = \left[\begin{array}{c|c} M_{21}^\intercal M_{21} & m_{33} M_{21}^\intercal \\ \hline m_{33} M_{21} & m_{33}^2 \end{array}\right] - \left(m_{33} - \frac{1}{\epsilon^2}\right) M \tag{27}$$

$$= \left[\begin{array}{c|c} M_{21}^\intercal M_{21} - m'_{33} M_{11} & (m_{33} - m'_{33}) M_{21}^\intercal \\ \hline (m_{33} - m'_{33}) M_{21} & (m_{33} - m'_{33}) m_{33} \end{array}\right] \tag{28}$$

[3]for completeness, note that this unique solution is $\tau = -\frac{\beta}{2\alpha}$.

where m'_{33} is defined as

$$m'_{33} = m_{33} - \frac{1}{\epsilon^2}. \tag{29}$$

So finally

$$S_\epsilon = \left[\begin{array}{c|c} M_{21}^\intercal M_{21} - m'_{33} M_{11} & (1/\epsilon^2) M_{21}^\intercal \\ \hline (1/\epsilon^2) M_{21} & (1/\epsilon^2) m_{33} \end{array}\right]. \tag{30}$$

S1.3 2-D silhouette

As mentioned earlier, Eq. (20) defines the silhouette of the ellipsoid in a perspective projection setup. On plane Π, this silhouette is defined by points p such that, for all solutions d to Eq. (20),

$$\begin{cases} p = e + \tau' d \\ (p - e) \cdot n = \epsilon - \delta \end{cases} \tag{31}$$

where τ' is a scalar[4]. The first equation of (31) is equivalent to

$$d = \frac{1}{\tau'}(p - e). \tag{32}$$

Equation (20) can now be rewritten in terms of p (and S_ϵ as noted earlier) as follows

$$(p - e)^\intercal S_\epsilon (p - e) = 0. \tag{33}$$

The point p can be written as

$$p = O_\Pi + q \tag{34}$$

where q is a vector whose third component is equal to 0. Using a block formulation, we have

$$q = \left[\begin{array}{c} q_1 \\ \hline 0 \end{array}\right] \tag{35}$$

and

$$S_\epsilon = \left[\begin{array}{c|c} S_{11} & S_{21}^\intercal \\ \hline S_{21} & s_{33} \end{array}\right]. \tag{36}$$

Then, Eq. (33) is equivalent to

$$q_1^\intercal P q_1 + Q q_1 = r \tag{37}$$

[4]combining the two equations of (31) together, one gets $\tau' = (\epsilon - \delta)/(d \cdot n)$

where [5]

$$P = S_{11}, \tag{38}$$
$$Q = 2(\epsilon - \delta)S_{21}, \tag{39}$$
$$r = -(\epsilon - \delta)^2 s_{33}. \tag{40}$$

One recognizes the equation of an ellipse which can be put into the following standard form

$$(q_1 - c)^\intercal \left(\frac{1}{r - Qc/2} P \right) (q_1 - c) = 1 \tag{41}$$

where $c = -P^{-1}Q^\intercal/2$ is the center of the ellipse.

S1.4 Semi-axes for perspective projection

Let λ_1 and λ_2 be the two (positive) eigenvalues of P, $\lambda_1 \leq \lambda_2$. Then the semi-minor and semi-major axes of the ellipse defined by Eq. (41) are

$$\rho_i = \sqrt{\frac{r - Qc/2}{\lambda_i}}, i \in \{1, 2\}. \tag{42}$$

Gathering everything together, ρ_i can be rewritten in terms of M as follows

$$r = -(1 - \delta/\epsilon)^2 m_{33}, \tag{43}$$
$$Q = 2\frac{\epsilon - \delta}{\epsilon^2} M_{21}, \tag{44}$$
$$P = M_{21}^\intercal M_{21} - \left(m_{33} - \frac{1}{\epsilon^2} \right) M_{11}, \tag{45}$$
$$c = -P^{-1}Q^\intercal/2, \tag{46}$$
$$\lambda_i = (\mathrm{tr}(P) + \sigma_i \sqrt{\Delta})/2, \tag{47}$$
$$|\sigma_i| = 1 \text{ and } \sigma_1 \sigma_2 = -1, \tag{48}$$
$$\Delta = \mathrm{tr}(P)^2 - 4\det(P) \tag{49}$$

where $\mathrm{tr}(P)$ is the trace of P, $\det(P)$ is its determinant, and the σ_i's are chosen so that $\rho_1 \geq \rho_2$.

[5] in case P is definite negative, P, Q and r must be replaced with their opposite

S1.5 Semi-axes for parallel projection

From the previous section, it follows that, for a parallel projection (i.e. $\epsilon = \infty$), the semi-minor and semi-major axes have the following simpler expression

$$\rho_i = \sqrt{\frac{m_{33}}{\lambda_i}}, i \in \{1, 2\} \tag{50}$$

and

$$P = m_{33} M_{11} - M_{21}^{\intercal} M_{21}, \tag{51}$$

while λ_i, σ_i, and Δ are unchanged. Note that δ no longer appears in the equations.

S2 Invariance of volume estimation error to scaling

S2.1 Individual volume

This section has no direct practical application. It only presents developments useful in Supporting Information S2.2.

For an axis-aligned ellipsoid, M is diagonal. The diagonal components are related to the semi-axes r_i as follows

$$m_{ii} = r_i^{-2}, i \in \{1, 2, 3\}. \tag{52}$$

The ellipsoid volume is then classically given by

$$V = \frac{4}{3}\pi r_1 r_2 r_3 \tag{53}$$

$$= \frac{4}{3}\frac{\pi}{\sqrt{m_{11}m_{22}m_{33}}}. \tag{54}$$

For a general ellipsoid (i.e., any orientation), the volume is

$$V = \frac{4}{3}\frac{\pi}{\sqrt{\det(M)}} \tag{55}$$

where $\det(M)$ is the determinant of M. Let V_* denote an estimation of the true volume V, where $*$ is ESD or ELL here. The relative error in volume estimation is defined as

$$\mathcal{E}_* = \frac{V_*}{V}. \tag{56}$$

For writing Eq. (56) for the $\mathcal{M}_{\text{ESD}}$, it should be reminded that, since the projection silhouette is an ellipse of area $\pi\rho_1\rho_2$, the equivalent radius is equal to $\sqrt{\rho_1\rho_2}$. Then, the relative errors of the $\mathcal{M}_{\text{ESD}}$ or $\mathcal{M}_{\text{ELL}}$ methods are

$$\mathcal{E}_{\text{ESD}} = (\rho_1\rho_2)^{3/2}\sqrt{\det(M)} \qquad \text{See Eq. (1)} \tag{57}$$

$$\mathcal{E}_{\text{ELL}} = \rho_1\rho_2^2\sqrt{\det(M)}. \qquad \text{See Eq. (2)} \tag{58}$$

The following section demonstrates the invariance of these individual volume estimation errors to scaling, a mathematical result that strongly impact the practical procedure (see Section "Simulation of copepod bodies").

S2.2 Invariance of individual volume estimation error to scaling

This section has no direct practical application. It only presents developments useful in Supporting Information S2.3.

S2.2.1 Common remarks

The purpose is to show that $\mathcal{E}_{\text{ESD}}$ and $\mathcal{E}_{\text{ELL}}$ do not depend on the absolute volume of the ellipsoid, which is a function of (r_1, r_2, r_3), but rather on the ellipsoid proportions $(r_2/r_1, r_3/r_1)$. One way to prove this statement is to show that $\mathcal{E}_*(\alpha M) = \mathcal{E}_*(M)$ for any $\alpha > 0$. Indeed, if this holds, then choosing α equal to r_1^2 implies that αM is defined by the triplet $(1, r_2/r_1, r_3/r_1)$.

Let ρ, resp. λ, be a generic notation for ρ_1 and ρ_2, resp. λ_1 and λ_2. The other useful reminders are

$$\rho = \sqrt{\frac{m_{33}}{\lambda}} \tag{59}$$

$$\lambda : \text{eigenvalue of } P \tag{60}$$

$$P = m_{33} M_{11} - M_{21}^{\intercal} M_{21}. \tag{61}$$

Let us add a subscript α to these quantities to denote their expressions when M is replaced with αM. We have

$$m_{33,\alpha} = \alpha m_{33} \tag{62}$$

$$P_\alpha = \alpha^2 P. \tag{63}$$

It is also clear that if λ is an eigenvalue of P, then $\beta\lambda$ is an eigenvalue of βP ($Px = \lambda x \Rightarrow \beta P x = \beta\lambda x$) for any $\beta \neq 0$. Therefore,

$$\lambda_\alpha = \alpha^2 \lambda. \tag{64}$$

Hence, it can be concluded that

$$\rho_\alpha = \frac{\rho}{\sqrt{\alpha}}. \tag{65}$$

Finally, note that we have the following property on the matrix determinant

$$\det(\alpha M) = \alpha^3 \det(M) \tag{66}$$

if M is a 3×3-matrix.

S2.2.2 $\mathcal{M}_{\text{ESD}}$ method

As a reminder, the relative error in volume estimation of the $\mathcal{M}_{\text{ESD}}$ method is

$$\mathcal{E}_{\text{ESD}}(M) = \mathcal{E}_{\text{ESD}}(r_1, r_2, r_3, \xi) \tag{67}$$

$$= (\rho_1 \rho_2)^{3/2} \sqrt{\det(M)}. \tag{68}$$

Then

$$\mathcal{E}_{\text{ESD}}(\alpha M) = (\rho_{1,\alpha}\rho_{2,\alpha})^{3/2}\sqrt{\det(\alpha M)} \tag{69}$$

$$= \frac{(\rho_1\rho_2)^{3/2}}{\sqrt{\alpha^3}}\sqrt{\alpha^3\det(M)} \tag{70}$$

$$= \mathcal{E}_{\text{ESD}}(M). \tag{71}$$

S2.2.3 $\mathcal{M}_{\text{ELL}}$ method

As a reminder, the relative error in volume estimation of the $\mathcal{M}_{\text{ELL}}$ method is

$$\mathcal{E}_{\text{ELL}}(M) = \mathcal{E}_{\text{ELL}}(r_1, r_2, r_3, \xi) \tag{72}$$

$$= \rho_1\rho_2^2\sqrt{\det(M)}. \tag{73}$$

Then

$$\mathcal{E}_{\text{ELL}}(\alpha M) = \rho_{1,\alpha}\rho_{2,\alpha}^2\sqrt{\det(\alpha M)} \tag{74}$$

$$= \frac{\rho_1\rho_2^2}{\alpha\sqrt{\alpha}}\sqrt{\alpha^3\det(M)} \tag{75}$$

$$= \mathcal{E}_{\text{ELL}}(M). \tag{76}$$

S2.3 Invariance of total volume estimation error to scaling

Let V^i, $i \in [1..n]$, be a set of (true) ellipsoid volumes, and let V_*^i and $\mathcal{E}_*^i$ be some corresponding estimated volumes by the method "$*$" and the associated individual volume estimation errors, respectively. The total volume estimation error is

$$\mathcal{T}_* = \frac{\sum_i V_*^i}{\sum_i V^i} = \frac{\bar{V}_*}{\bar{V}} \tag{77}$$

where $\bar{X}$ denotes the average of X. Now, suppose that each ellipsoid volume is scaled by a factor α_i ($U^i = \alpha_i V^i$), for example as a result of the normalization of the ellipsoids by dividing their semi-axes r_1^i, r_2^i, and r_3^i by r_1^i. How will the estimated volumes U_*^i vary with respect to V_*^i? From Supporting Information S2.2, we know that $\mathcal{E}_*^i = V_*^i/V^i$ is invariant to ellipsoid scaling when "$*$" is ESD or ELL. Therefore, U_*^i/U^i must still be equal to $\mathcal{E}_*^i$, which implies that $U_*^i = \alpha_i V_*^i$. Hence, the total volume estimation error after scaling is

$$\mathcal{T}'_* = \frac{\sum_i U_*^i}{\sum_i U^i} = \frac{\sum_i \alpha_i V_*^i}{\sum_i \alpha_i V^i} = \frac{\overline{\alpha V_*}}{\overline{\alpha V}}. \tag{78}$$

The covariance between two random variables, say a factor α and a volume V, can be written in terms of expected values as follows

$$\mathrm{Cov}(\alpha, V) = \mathrm{E}[\alpha V] - \mathrm{E}[\alpha]\,\mathrm{E}[V]. \tag{79}$$

The situation of practical interest here is when α is related to the normalization of the ellipsoids, in which case V and α are not independent (which would otherwise immediately guarantee that $\mathrm{Cov}(\alpha, V) = 0$). Indeed,

$$\alpha = \frac{1}{r_1^3} \tag{80}$$

$$V = \frac{4}{3}\pi r_1 r_2 r_3. \tag{81}$$

However, they are not correlated (i.e., there relationship is not linear). Consequently, their covariance is equal to zero and

$$\mathrm{E}[\alpha V] = \mathrm{E}[\alpha]\,\mathrm{E}[V]. \tag{82}$$

The same conclusion can be drawn regarding the estimated volume V_*. If the number of samples n is large enough, these results can be safely extended to average values so that

$$\overline{\alpha V} \simeq \bar{\alpha}\bar{V} \tag{83}$$

$$\overline{\alpha V_*} \simeq \bar{\alpha}\bar{V}_*. \tag{84}$$

In the end,

$$\mathcal{T}_*' \simeq \frac{\bar{\alpha}\bar{V}_*}{\bar{\alpha}\bar{V}} = \mathcal{T}_*. \tag{85}$$

This concludes the proof that, if the number of involved volumes is high enough, the total volume estimation error is largely invariant to scaling. As a consequence, the randomly generated ellipsoids used to determine the total volume correction factor (see Sections "Simulation of copepod bodies" and "Corrected total volume", and, in particular, Eq. (12)) can be safely normalized, for example by dividing their semi-axes by their largest one.

S3 Improved surface estimation and ellipse fitting

S3.1 Common steps

The general idea of the improved methods proposed in this appendix is to get rid of the antennae (and tail) before measuring the copepod silhouette surface or fitting an ellipse onto it. It is assumed that the binary mask of the copepod has been determined previously. We propose to compute the Inner Distance Map (IDM) of this mask and to erode it using a threshold. We fix the threshold to $\max(\text{IDM})/2.1$ in our experiments. This step allows to get rid of the antennae. Note that the binary mask could have been eroded directly using mathematical morphology. However, it would make use of a discrete so-called structural element (typically a discretized disk), which would lead to a coarser eroded shape. Given the small size of a copepod in our images, this could have a negative impact on the subsequent steps. Next, to recover the original copepod body size, the outer distance map of the eroded mask is computed and thresholded using the same threshold as the one used for erosion. This amounts to dilate the eroded mask, but again in a finer way than if using mathematical morphology. The various steps are illustrated in Fig. S3.1. A Python implementation is available at: `https://gitlab.inria.fr/cedubois/Copepod-Volume-Correction`.

S3.2 Surface estimation

The copepod surface estimation is performed by counting the number of pixels of its binary mask. The improved version simply counts the pixels of the mask obtained in Supporting Information S3.1 as opposed to counting the pixels in the original binary mask which includes the antennae.

S3.3 Ellipse fitting

When an object is described by a binary mask of pixels, the most classical ellipse fitting method interprets the pixels as the samples of a point cloud. The covariance matrix of the cloud is computed. Its eigenvectors represent the best fitting ellipse orientation while its eigenvalues represent the semi-axes of the ellipse. A simple improvement of this method (or any other ellipse fitting method as a matter of fact) consists in rescaling the ellipse so that its area matches the object area. This is implemented by the software ImageJ that ZooProcess (ZP), the software used to process UVP images,

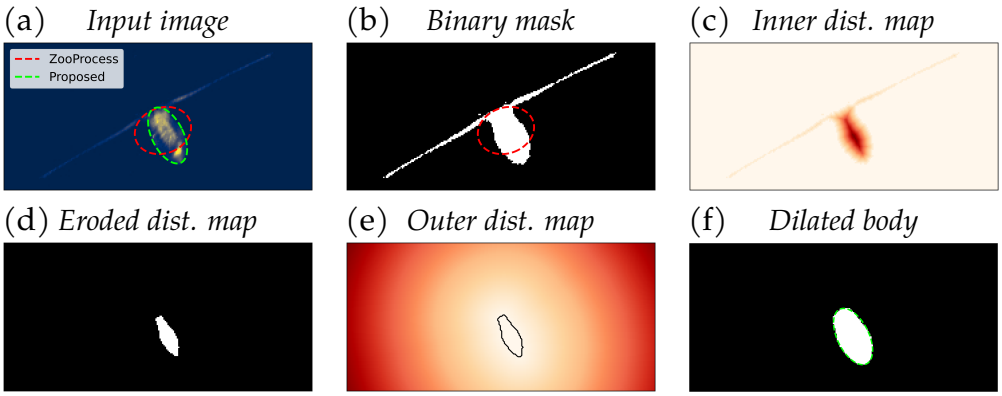

Figure S3.1: Copepod body mask computation as a common preliminary step for surface estimation and ellipse fit. Reading the figure in lexicographical order, each image is the result of the processing of the previous one. They are: (a) the input grayscale image, (b) the binary mask obtained by thresholding, (c) the inner distance map, (d) the eroded mask obtained by thresholding, (e) the outer distance map, and (f) the dilated mask obtained by thresholding (same threshold).

uses. However, if this improvement allows to correct the fitted ellipse surface (which can be enough for some applications), it does not help that much for copepod volume estimation. Indeed, the precision of the small semi-axis is crucial, and it is not improved by the surface adjustment. As a reminder, the $\mathcal{M}_{\text{ELL}}$ estimation of the volume is:

$$V_{\text{ELL}} = \frac{4}{3}\pi\rho_1\rho_2^2 = \frac{4}{3}\underbrace{\pi\rho_1\rho_2}_{\text{Surface}}\ \underbrace{\rho_2}_{\substack{\text{Minor}\\ \text{semi-axis}}}. \tag{86}$$

Alternatively, the ellipse could be fitted on the grayscale version of the object, that is using the pixel intensities as sample weights when computing the covariance matrix. However, we found that this alternative does not work well on the copepod images of our data set.

Whatever the ellipse fitting method is, the starting point is the copepod mask. The fitting methods get distracted by the antennae, which can result in very bad ellipses (see the red ellipses in Fig. S3.2). Therefore, we proposed to fit an ellipse on the mask obtained in Supporting Information S3.1 instead of the original binary mask which includes the antennae (see the green ellipses in Fig. S3.2).

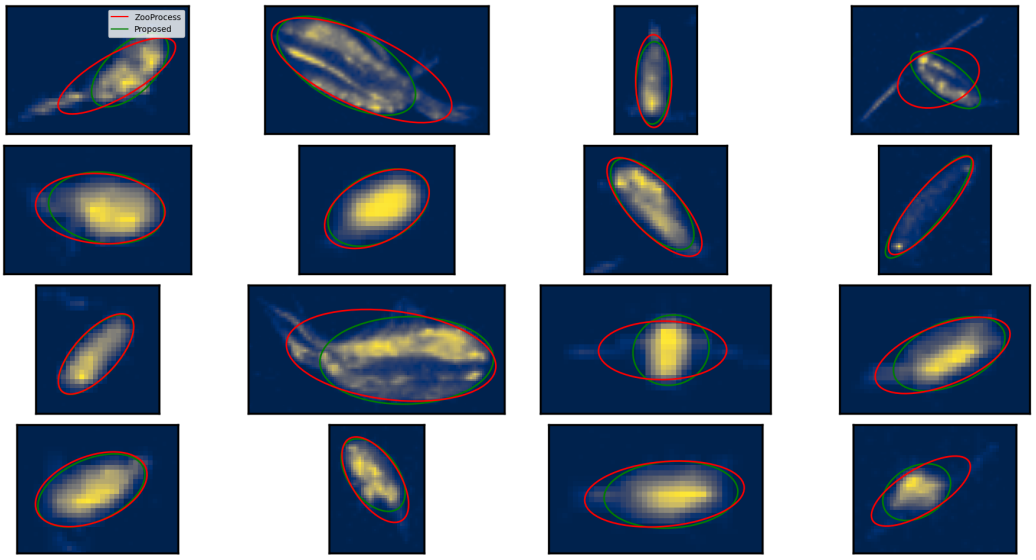

Figure S3.2: Multiple examples of ellipse fit based on the original mask in green and for the new proposed method in red. We see that when antennae are not visible, the result is almost the same, but when they are visible, the classic method is not appropriate.

S4 The proposed method, step-by-step

This section gathers the results of the different sections into a step-by-step procedure for estimating the total volume of copepods given a data set of 2-D views. It is composed of two stages: a learning stage which has to be performed once for all, or whenever the expert thinks the proposed simulation procedure must be adapted to the data, and a "usage" stage which can be applied at will.

S4.1 Learning stage

1. Generate random ellipsoid samples that realistically represent a generic population of copepods, or a population following some characteristics inferred from the data set. The randomness must be constrained by the expert knowledge in the form of specific simulation parameters.
2. Compute the total volume of the ellipsoid samples. This represents the true total volume. See Eqs. (4) and (12).
3. For each ellipsoid sample, compute the projection ellipse (see Eq. (7)) and the estimated volume using either the $\mathcal{M}_{\text{ESD}}$ (see Eq. (1)) or the

$\mathcal{M}_{\text{ELL}}$ method (see Eq. (2)).

4. Sum all the estimated volumes to get the estimated total volume.
5. Compute the total volume estimation error $\mathcal{T}_*$ from the true and estimated total volumes (see Eq. (12)). This is the final product of the learning stage.

S4.2 "Usage" stage

1. For each copepod image of a data set, determine the copepod silhouette using an image segmentation method. On UVP images, a simple binarization using a fixed threshold is enough.
   1.a. For the $\mathcal{M}_{\text{ESD}}$ method, compute the silhouette area A (see Supporting Information S3.2) and the corresponding estimated volume (see Eq. (1)).
   1.b. For the $\mathcal{M}_{\text{ELL}}$ method, fit an ellipse onto the silhouette (see Supporting Information S3.3). Let ρ_1 and ρ_2 be the semi-major and semi-minor axes, respectively. Then compute the corresponding estimated volume (see Eq. (2)).
2. Sum all the estimated volumes to get the estimated total volume $\tilde{W}_*$ where $*$ is either ESD or ELL.
3. Compute the corrected total volume estimation $\hat{W}_*$ by dividing $\tilde{W}_*$ with $\mathcal{T}_*$ from the learning stage (see Eq. (13)).

S4.3 Algorithmic details

S4.3.1 Uniformly random rotations

This section defines the rotation matrices used to simulate random orientations of ellipsoids.

In order to generate an ellipsoid with a uniformly random orientation, we generate a random rotation matrix R and rotate an axis-aligned ellipsoid with it. The generation of an axis-aligned ellipsoid is described in Section "Simulation of copepod bodies". If the axis-aligned ellipsoid is represented by a matrix M (see Section "Principle for correction of total volume"), then the rotated ellipsoid is represented by the matrix

$$M_{\text{rot}} = RMR^{\intercal}. \tag{87}$$

A general rotation matrix can be defined using three elementary rotation

matrices

$$R = R_z(\Phi)R_y(\Theta)R_x(\Psi) \tag{88}$$

with $R_i(\alpha)$, $i \in \{x, y, z\}$, defines the rotation by angle α around axis i. To generate a random rotation matrix, one has to randomly choose the angle triplet (Ψ, Θ, Φ). In order to guarantee the uniformity of the ellipsoid orientations, the angles Ψ, Θ, and Φ must be distributed adequately, that is

$$\Psi = U[0, 2\pi[\tag{89}$$

$$\Theta = \arccos(1 - 2U[0, 1[) - \frac{\pi}{2} \tag{90}$$

$$\Phi = U[0, 2\pi[\tag{91}$$

where $U[a, b[$ is the uniform distribution between a (included) and b (excluded).

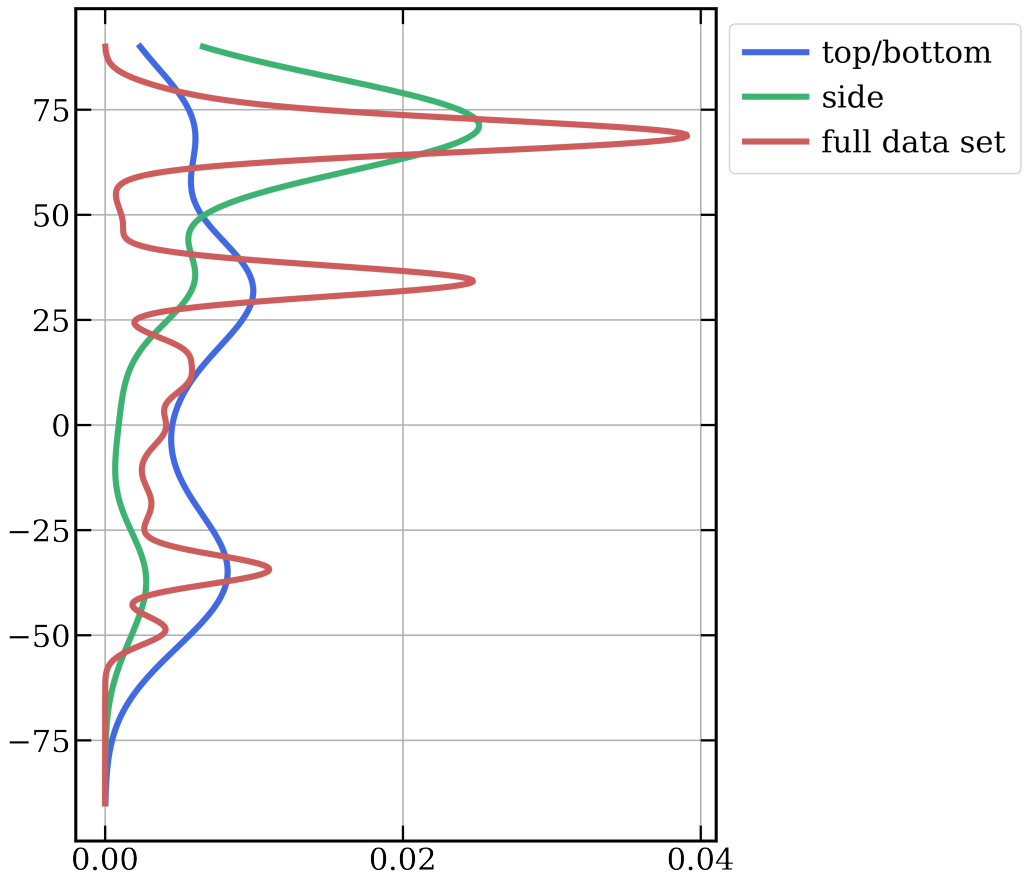

Figure S5.1: Kernel density estimate of the latitudinal distribution of the images of all copepods and of the side or top/bottom views.

S5 Distribution of selected sample images

To define the real-world distribution of the semi axes of the ellipsoid representing the body of copepods (r_1, r_2, and r_3) as well as a the ratios between them, defining the shape of the ellipsoid (r_2/r_1 and r_3/r_1), 295 copepods seen from the side (on which r_1 and r_2 are measurable) and 265 copepods seen from the top or bottom (on which r_1 and r_3 are measurable) were manually curated from a collection of >150k images. To make sure that these small samples were representative of the whole data set, we checked their latitudinal and size (i.e. r_1) distributions. The shape of the latitudinal distribution of the side and top/bottom views matches well that of the total data set (Fig. S5.1). The side views shows and excess at high latitude, likely linked with a bias in the size distribution (see below; copepods are larger at

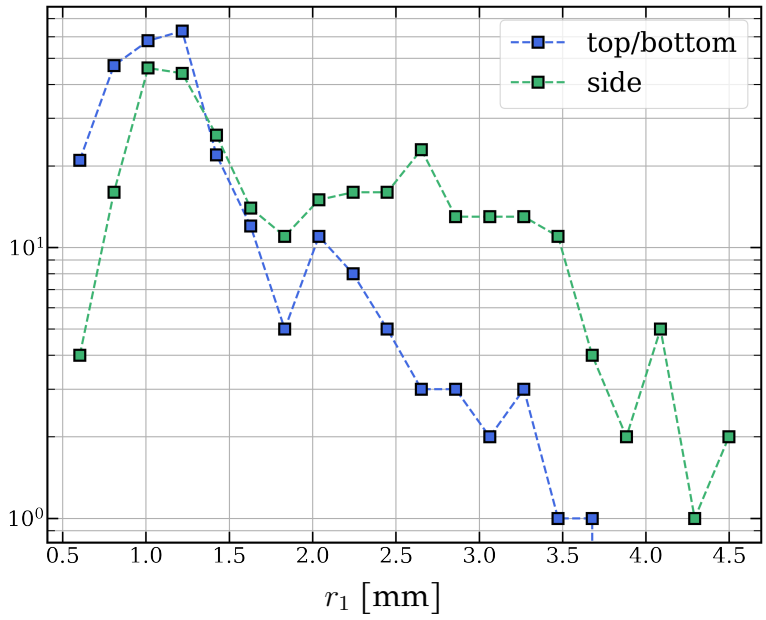

Figure S5.2: Distribution of the length of the semi-major axis of the ellipse fitted in the two views of the copepods. The vertical axis is the number of observations, in $\log_{10}$ scale. The horizontal axis is the semi-major axis r_1, which is equal to ρ_1 in these viewpoints and approximates the half of the prosome length, in millimeters.

high latitudes), and a linked under-representation elsewhere. The pattern is opposite for the top-bottom views. However, no region is completely missed in the samples and even some details of the distribution (such as the two peaks around -40°) are captured. Therefore, we consider them representative enough. The theoretical expectation for the length distribution is an exponential decay (Sprules and Barth, 2016), i.e. a linear decrease, in log-scale. This is approximately true once the lower detection limit of the camera is passed, after $\sim$ 1 mm (Fig. S5.2). However, the distribution of side views shows an excess in the size range 2 to 3.5 mm. This is likely due to the fact that telling that a copepod is viewed from the top/bottom can be determined from the geometry of its antennae relative to its body, no matter its size; making sure that a copepod is viewed from the side requires additional details, which are easier to asses on larger individuals, inducing a bias in the manual selection of images. As explained in the main text, this has little consequence on the estimation of the distribution of the semi-axes ratios (r_2/r_1 and r_3/r_1) but does now allow the estimation of the distribution of r_1 from these samples only.

S6 Sensitivity of the correction factors to shape and orientation

S6.1 Shape

In our dataset, the distribution of semi axes ratios was unimodal (see Fig. 4). Nevertheless, the proposed method can accommodate any distribution thanks to the use of the KDE approach (Parzen, 1962; Scott, 1979). Thus, it is interesting to verify how the correction factors vary in a multimodal scenario, for example when two populations of copepods with different shapes are present. Figure S6.1 shows some examples of synthetic PDFs of r_3/r_1 for a mixture of two body shape distributions with varying proportions, and the effect on the correction factors, the other distributions (r_2/r_1 and body orientation) being fixed. The top-left panel represents organisms with a round cross-section, that is to say $r_3 \sim r_2$ (i.e. calanoid-like). Therefore, we used the same distributions for r_3/r_1 and r_2/r_1: a Normal law with mean 0.41 and variance 0.0076 (values based on the fit to the UVP5 data). The second distribution represents organisms with a flatter cross-section i.e. $r_3 \ll r_2$ (i.e. harpacticoid-like). We used a Beta distribution of parameters $a = 2$ and $b = 15$. The parameter $\alpha \in [0, 1]$ determines the proportions of samples from the two populations ($\alpha = 0$: 100% of the samples are from the first population; $\alpha = 1$: 100% are from the second one). The computed correction factors increase with α, i.e. as the proportion of flatter organisms increases. This is to be expected since it is for this kind of shape that the viewing angle has the most consequences on the appearance of the organism. This illustrates that the correction factors strongly depend on community composition (and therefore on a correct modeling of the data). In particular, the correction factors obtained in the paper for the global UVP5 dataset should not be used blindly on other datasets. Instead, the required PDFs should be estimated from the data.

S6.2 Orientation

In our simulations, the orientation of copepods relative to the camera was considered uniformly random. Nevertheless, with other imaging instruments, the orientation may be more constrained (e.g. with scanners like the ZooScan or in-flow imagers such as the FlowCam). It is possible to relax the uniformity assumption and verify how the correction factors vary with different degrees of constraint on the orientation.

Rotation can be performed around the x-axis of the copepod/ellipsoid

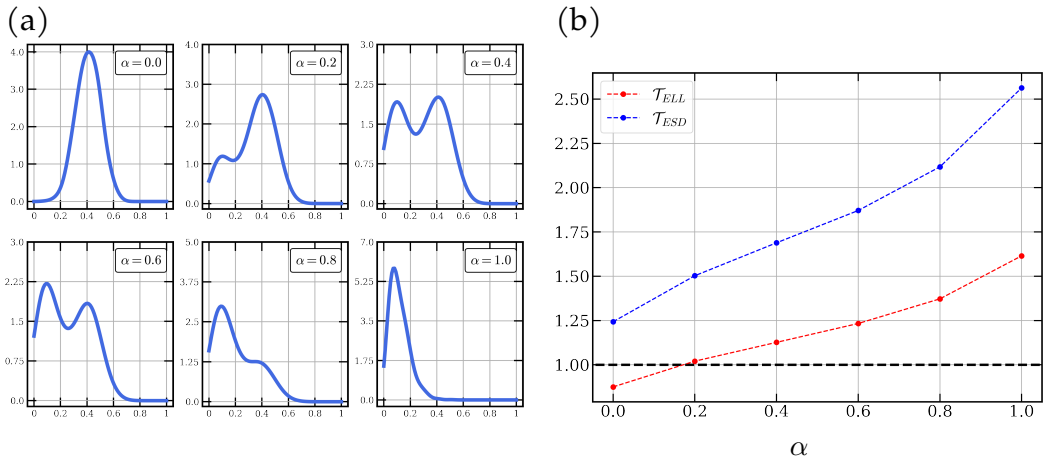

Figure S6.1: (a) Distribution of r_3/r_1 for various mixtures of two subpopulations, from $\alpha = 0$ (Normal distribution fitted on the r_2/r_1 data of the UVP5 experiment), to $\alpha = 1$, Beta distribution with parameters $a = 2$ and $b = 15$. (b) The corresponding correction factors for $N = 10^6$ ellipsoids. The black dashed line indicates a correction factor of one, i.e. no error.

(i.e. the length) and determines whether we get a dorsal, side or ventral view (or something in-between); along the y-axis (i.e. the width) and, in the case of the UVP, this changes the vertical tilt of the organism; and along the z-axis, normal to the view plane, which, in the case of the UVP, changes the "cardinal" orientation of the organism. In the simulations, rotation along the z-axis is set to the identity (i.e. no rotation) since this does not influence the results at all; the x-axis rotation is free and uniform in $[0, \pi]$ (so that $\rho_2 \in [r_3, r_2]$); and the y-axis rotation is uniform in the interval $[0, \theta_{\max}]$: the higher $\theta_{\max}$, the more the ellipsoid can rotate vertically. Figure S6.2 shows the correction factors obtained. When $\theta_{\max} = 0$, all copepods are aligned on a plane (a "Zooscan-like" scenario). When $\theta_{\max} = \pi/2$, the results are the same as with a random uniform distribution. It is interesting to note that: (i) the correction factor for the $\mathcal{M}_{\text{ELL}}$ method is relatively stable, while it varies more significantly for the $\mathcal{M}_{\text{ESD}}$ method; (ii) the variation among the different orientation scenarii is much lower than for the different shapes (Figure S6.1).

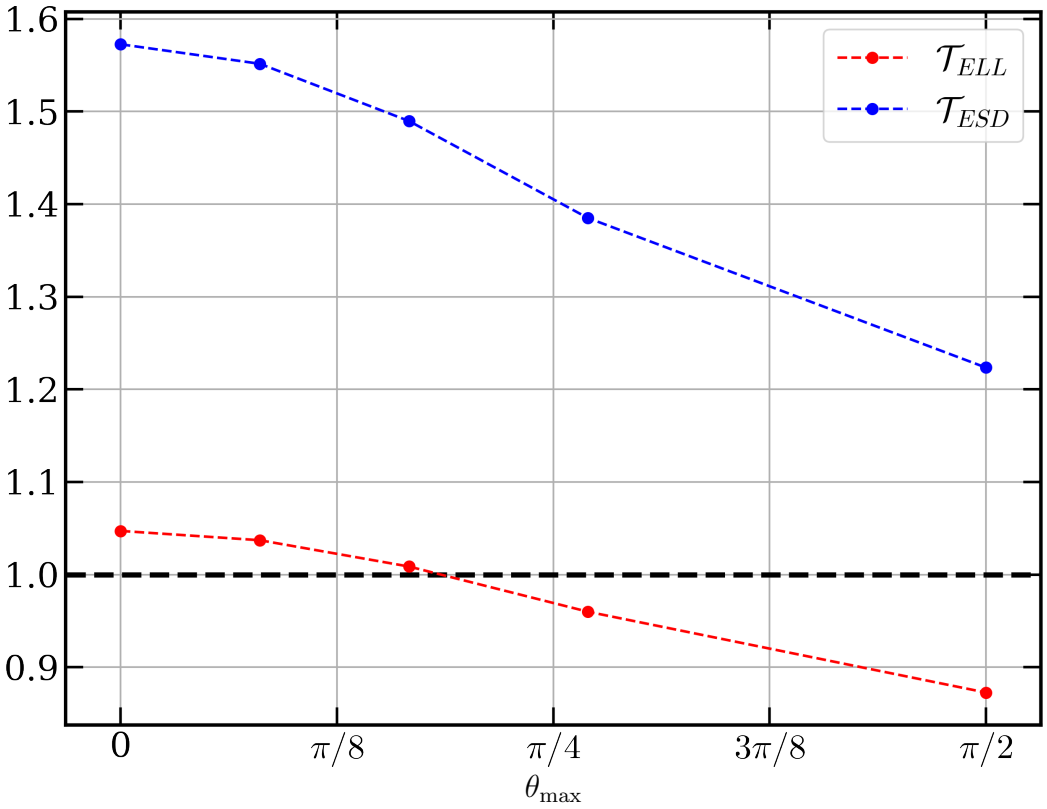

Figure S6.2: Evolution of the correction factors for varying ranges of allowed vertical rotation. The rotation angle around the y-axis, θ, is restricted to the interval $[0, \theta_{\max}]$. Each dot of the plots corresponds to a simulation for a particular $\theta_{\max}$. The black dashed line indicates a correction factor of one, i.e. no error.

S7 Sensitivity of the correction factors to the number of simulated ellipsoids

The number N of ellipsoids generated in the simulation only determines the precision of the estimation of those factors. As a matter of fact, the estimation of the correction factors tends to be perfect when N tends towards infinity. So this number has no reason to be related to the size of the dataset. Yet, it is still interesting to get an idea of the influence of N on the variance of the computed factors. Thus, we performed simulations with various numbers of ellipsoids $N_i = 10^i, i \in \{3, 4, 5, 6, 7\}$, lower than $N = 10^8$ used in the paper. Fig. S7.1 shows that the variance of the correction factors becomes negligible for $N_i \geq 10^6$.

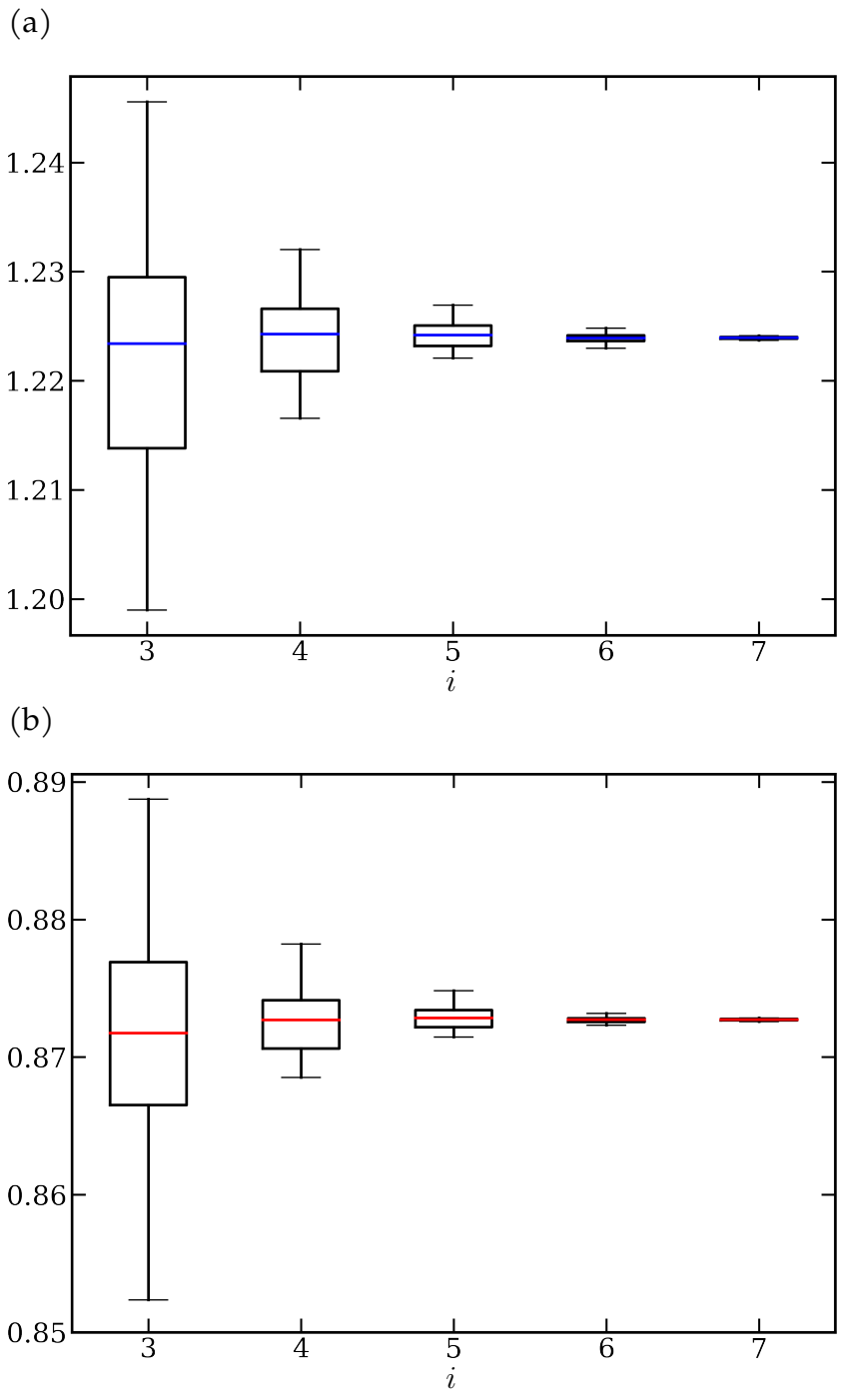

Figure S7.1: Corrections factors ((a) $\mathcal{M}_{\text{ESD}}$, (b) $\mathcal{M}_{\text{ELL}}$) for $N_i = 10^i, i \in \{3, 4, 5, 6, 7\}$. For each i, we computed the correction factors 50 times. The blue (red) line is the mean factor for the $\mathcal{M}_{\text{ESD}}$ ($\mathcal{M}_{\text{ELL}}$) method. The boxes and whiskers are drawn according to Tukey's definition (McGill et al., 1978).

References

McGill, R., Tukey, J. W., and Larsen, W. A. (1978). Variations of box plots. *The American Statistician*, 32(1):12–16.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.*, 33(3):1065–1076.

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, 66(3):605–610.

Sprules, W. G. and Barth, L. E. (2016). Surfing the biomass size spectrum: some remarks on history, theory, and application. *Canadian Journal of Fisheries and Aquatic Sciences*, 73(4):477–495.