



# Bayesian inference of a parametric random ellipsoid from its orthogonal projections

Mathieu de Langlard, Fabrice Lamadie, Sophie Charton, Johan Debayle

## ► To cite this version:

Mathieu de Langlard, Fabrice Lamadie, Sophie Charton, Johan Debayle. Bayesian inference of a parametric random ellipsoid from its orthogonal projections. *Methodology and Computing in Applied Probability*, 2021, 23, pp.549 à 567. 10.1007/s11009-020-09806-w . hal-02927445

**HAL Id: hal-02927445**

**<https://hal.science/hal-02927445>**

Submitted on 7 Sep 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Bayesian inference of a parametric random spheroid from its orthogonal projections

Mathieu de Langlard\* · Fabrice Lamadie ·  
Sophie Charton · Johan Debayle

Received: date / Accepted: date

**Abstract** The paper focuses on a new method for the inference of a parametric random spheroid from the observations of its 2D orthogonal projections. Such a stereological problem is well-known from the literature when the projections come from only one deterministic spheroid. Nevertheless, when the spheroid is random itself, the estimation of its distribution is not straightforward. From a theoretical viewpoint, it is shown that the semi-axes of the spheroid and the ones of the projected ellipses are linked through a random polynomial of degree two which admits two real random positive roots. The likelihood can be formulated in terms of the coefficients of the random polynomial, but is not analytically tractable. Assuming that the random spheroid is parameterized by a set of parameters  $\theta_{\text{real}}$ , an approximation of the maximum a posteriori is used to estimate  $\theta_{\text{real}}$ . The estimator is based on the so-called approximate Bayesian computation method and a kernel density technique. As an illustration, the case of a spheroids population, whose semi-major axis follows a gamma distribution and the flattening coefficient a truncated normal distribution, is studied. The numerical results demonstrate that the bias of the estimator is very low, with a reasonable variance, both for the first and the second order moments of the semi-axes. The proposed method enables to recover some 3D morphological characteristics of a population of independent and

---

Mathieu de Langlard  
CEA, DES, ISEC, DMRC, Univ. Montpellier, Marcoule, France  
\*Current affiliation: INRIA Paris, France  
Tel.: (+49) 731 502 3529  
E-mail: mathieu.de-langlard@inria.fr

Fabrice Lamadie  
CEA, DES, ISEC, DMRC, Univ. Montpellier, Marcoule, France  
E-mail: fabrice.lamadie@cea.fr

Sophie Charton  
CEA, DES, ISEC, DMRC, Univ. Montpellier, Marcoule, France  
E-mail: sophie.charton@cea.fr

Johan Debayle  
Mines Saint-Etienne, CNRS, UMR 5307 LGF, Centre SPIN, 158 cours Fauriel, 42023 Saint-Étienne Cedex 2, France  
E-mail: debayle@emse.fr

identically distributed spheroids thanks to the only observations of its projected ellipses.

**Keywords** Bayesian inference · random spheroid · orthogonal projection · stereology · morphological characterization · approximate Bayesian computation · Markov Chain Monte Carlo method

## 1 Introduction

In many situations, ellipsoids arise as a simple, but realistic, model for given objects. For example, ellipsoidal models are frequently encountered in medical areas (Fessler and Macovski 1991; Jaggi et al. 1995; Noumeir 1999), and they are also used as generalized model for droplets and/or bubbles in the field of chemical engineering (de Langlard et al. 2018b; Kracht et al. 2013). More generally, they are fundamental in 3D imaging (Merola et al. 2013; Liu et al. 2006; Ozturk-Isik et al. 2009) for which morphological characteristics (volume, surface, eccentricity, etc.) of the ellipsoids are essential as they inform on the behavior/properties of the system.

Regardless of the application, the modeled ellipsoidal objects are usually observed only through sections (Launeau and Robin 2005; Sahagian and Prousevitich 1998) or projections (de Langlard et al. 2018a). In this paper, the focus is on orthogonal projection, sometimes called *silhouette projection* (Karl et al. 1994), for which solely the elliptical shadow resulting from the orthographic projection is observed. Methods to reconstruct a single ellipsoidal object from its projections are available in the literature (Bresler and Macovski 1987; Noumeir 1999). Karl et al. (1994) reported an important results, in particular that three projections on different angles are sufficient to uniquely reconstruct the underlying ellipsoid. However, this type of reconstruction methods can hardly be implemented for a population of ellipsoids for which various projections at different angles are observed. It then becomes necessary to match the projections with their original ellipsoids, which may prove difficult or even impossible in some applications (de Langlard et al. 2018a). The proposed method aims at addressing such situation by adopting a probabilistic approach.

We consider a population of ellipsoidal objects which are independent and identically distributed (i.i.d) realizations of a random spheroid, meaning that the axes and the orientation of each spheroid realization follow the same probability laws. The observed projections, called *projected ellipses*, can be viewed as i.i.d realizations of the projected random spheroid. In such probabilistic setting, the problem is not a reconstruction problem from the projections, but rather an inference problem on the probability laws of the random spheroid characteristics. Interestingly, the link between the 3D properties of the random spheroid and those 2D of the projected ellipses is established by a random polynomial of degree two. Assuming the random spheroid is parametrized by the vector  $\theta_{\text{real}}$ , the inference is made with a Bayesian approach. An estimator of an approximation of the maximum *a posteriori* (MAP), called *approximate maximum a posteriori* (AMAP), is proposed. The estimation procedure is based on the approximate Bayesian computation (ABC) method to provide samples from an approximation of the posterior density, and then on a kernel density method applied on these samples. The AMAP

estimator is compared to the approximate posterior mean (APM) estimator which consists of averaging the samples from the approximation of the posterior density. The computation of the AMAP estimator is theoretically justified by the work of Rubio et al. (2013), and its direct relationship with the maximum likelihood estimator, while the APM estimator is only used here as a benchmark.

The proposed inference procedure allows for the estimation of the 3D morphological characteristics of a population of spheroids, assuming that they are independent and identically distributed. We have not encountered other methods which try to fill the gap between the observed 2D projected ellipses semi-axes and the 3D geometrical properties of the population of spheroids. A direct application is the morphological characterization of multiphase flows using image processing, where the images of particles (bubbles or droplets) are often 2D projections coming from 3D spheroid-like particles (see de Langlard et al. (2018a)).

This paper is divided in five sections. Section 2 formulates mathematically the problem and establishes the form of the likelihood. Section 3 focuses on the sampling scheme from the approximation of the posterior distribution, and on the computation of the APM estimator. The AMAP estimator is introduced in Section 4. Finally, some numerical results and a comparison between the APM and the AMAP estimators are presented in Section 5, before drawing the conclusion in Section 6.

## 2 Probabilistic relationship between a random spheroid and its projections

### 2.1 Problem formulation

We consider a random ellipsoid of revolution around its semi-major axis  $a$ , called *random spheroid*, which is centered at the origin  $O$  in  $\mathbb{R}^3$  and whose major axis is oriented along  $(Oz)$ , where  $z$  is the point of coordinates  $(0, 0, 1)$ . The semi-axes,  $a$  and  $b$ , are random variables with  $b = \epsilon a$  where  $\epsilon$ , the random flattening coefficient, takes values in the unit interval  $[0, 1]$ . Moreover, it is assumed that  $a \sim \mathcal{F}_1(\boldsymbol{\theta}_1)$  and  $\epsilon \sim \mathcal{F}_2(\boldsymbol{\theta}_2)$ , where  $\mathcal{F}_1$  and  $\mathcal{F}_2$  are probability laws which depend respectively on the parameter vector  $\boldsymbol{\theta}_1 \in \mathbb{R}^{p_1}$  and  $\boldsymbol{\theta}_2 \in \mathbb{R}^{p_2}$ ,  $p_1$  and  $p_2$  being their respective number of parameters. Let  $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_k)$  be the random vector such that  $\mathbf{Z}_i = (a_i, \epsilon_i)$  where  $a_i \sim \mathcal{F}_1$  and  $\epsilon_i \sim \mathcal{F}_2$  for all  $i \in \llbracket 0, k \rrbracket$ . The  $\mathbf{Z}_i$  are assumed to be independent and identically distributed (i.i.d) and their realizations are not observed: they are *hidden states*.

Besides, it is assumed that a set of ellipses  $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_k)$  is observed. Each ellipse corresponds to an orthogonal projection of a realization of the spheroid in an Euclidean plane. The projection plane is orthogonal to a random vector  $\mathbf{u}$  uniformly distributed on the unit sphere  $\mathbb{S}^2$ . The observations can then be rewritten as  $\mathbf{y}_i = (a'_i, b'_i)$ , where  $a'_i$  and  $b'_i$  are the semi-major and semi-minor axis of the corresponding ellipse. We denote by  $\mathbf{Y}_i$ 's the associated random variables that are also i.i.d.

The aim of the proposed method is to recover the parameter vector  $\boldsymbol{\theta}_{\text{real}} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \mathbb{R}^d$ , with  $d = p_1 + p_2$ , from  $\mathbf{y}$ , the observations of the projected ellipses. A schematic representation of the problem is shown in Figure 1.

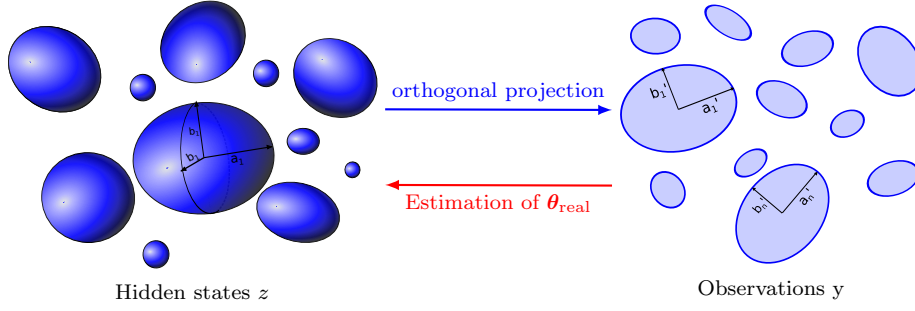


Fig. 1: Representation of the inference problem: the projected ellipses are observed and the goal is to recover the vector of parameters  $\theta_{\text{real}}$ .

A Bayesian framework is adopted, in which the parameter vector to be recovered  $\theta_{\text{real}}$  is considered as a realization of a random variable  $\Theta$ . The method then consists in finding the maximum *a posteriori* (MAP), i.e. the point value  $\theta_{\text{MAP}}$  at which the posterior density function  $f(\cdot|\mathbf{y})$  is maximal. This  $\theta_{\text{MAP}}$  value is then used as an estimator of  $\theta_{\text{real}}$ . According to the Bayes formula, we have

$$f(\theta|\mathbf{y}) \propto l(\mathbf{y}|\theta)f(\theta), \quad (1)$$

where  $l(\cdot|\theta)$  is the probability density function of the sample  $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_k)$  given a realization of the set of parameters  $\theta$ , and  $f(\cdot)$  the *a priori* marginal probability density function, called the *prior density*, of the random parameter vector  $\Theta$ . We denote by  $\mathcal{L}(\theta; \mathbf{y}) = \prod_{i=1}^k l(\mathbf{y}_i|\theta)$  the corresponding likelihood function. The marginal distribution,  $f$ , and its support,  $D$ , are fixed according to prior information on the set of parameters, while the likelihood is problem specific, and has to be determined.

## 2.2 Likelihood determination

Vickers (1996) proved that the orthogonal projection of a revolution ellipsoid of Cartesian equation

$$a \geq c \geq b > 0, \quad \frac{x^2}{c^2} + \frac{y^2}{b^2} + \frac{z^2}{a^2} = 1 \quad (2)$$

on the plane orthogonal to the unit vector  $\mathbf{u} = (u_x, u_y, u_z)$  is an ellipse whose squared semi-axes  $(a'^2, b'^2)$  are the roots of the polynomial

$$\mathcal{X}(\lambda) = \lambda^2 - \alpha\lambda + \beta, \quad (3)$$

where the coefficients  $\alpha$  and  $\beta$  are defined by

$$\begin{cases} \alpha = u_x^2(b^2 + a^2) + u_y^2(a^2 + c^2) + u_z^2(c^2 + b^2) \\ \beta = u_x^2 b^2 a^2 + u_y^2 c^2 a^2 + u_z^2 c^2 b^2. \end{cases} \quad (4)$$

Here, the unit vector  $u$  is uniformly distributed on the unit sphere  $\mathbb{S}^2$  of  $\mathbb{R}^3$ , thus guaranteeing the random nature of the projection plane. Note that it is equivalent to assuming either that the random spheroid is uniformly oriented, with the

projections belonging to a fixed plane, or that the random spheroid is fixed and the projection plane is uniformly oriented. Moreover, the coefficients  $\alpha$  and  $\beta$  are random variables, as they are measurable as a product and sum of measurable functions of the random variables  $a$ ,  $\epsilon$  and  $\mathbf{u}$ .  $\mathcal{X}$  is therefore called a *random polynomial*.

With the assumption  $b = c = \epsilon a$  where  $\epsilon \in [0, 1]$ , the set of equations simplifies to

$$\begin{cases} \alpha = a^2(1 + \epsilon^2 + u_z^2(\epsilon^2 - 1)) \\ \beta = \epsilon^2 a^2(\alpha - \epsilon^2 a^2). \end{cases} \quad (5)$$

The roots  $\lambda_1$  and  $\lambda_2$  of  $\mathcal{X}$  are also random variables because they are measurable as continuous functions of  $\alpha$  and  $\beta$ . In the next proposition, it is shown that the roots of  $\mathcal{X}(\lambda)$  are real random variables.

**Proposition 1** *Let  $\Delta = \alpha^2 - 4\beta$  be the real random variable where  $\alpha$  and  $\beta$  are the random variables defined in Eq. (5). Then,  $\Delta$  is positive almost surely.*

*Proof* We have

$$\begin{aligned} \Delta \geq 0 &\Leftrightarrow \alpha^2 \geq 4\beta \\ &\Leftrightarrow a^4(1 + \epsilon^2 + u_z^2(\epsilon^2 - 1))^2 \geq 4\epsilon^2 a^2(a^2(1 + \epsilon^2 + u_z^2(\epsilon^2 - 1)) - \epsilon^2 a^2) \\ &\Leftrightarrow (\epsilon^2 + 1 + u_z^2(\epsilon^2 - 1))^2 \geq 4\epsilon^2(1 + u_z^2(\epsilon^2 - 1)). \end{aligned}$$

All inequalities are true almost surely. It then suffices to note that the last inequality is true because

$$\forall (x, y) \in \mathbb{R}^2, (x + y)^2 \geq 4xy.$$

Choosing  $x = \epsilon^2$  and  $y = 1 + u_z^2(\epsilon^2 - 1)$  leads to the result.  $\square$

From this proposition, we deduce that the roots can be expressed as

$$\lambda_1 = \frac{\alpha + \sqrt{\alpha^2 - 4\beta}}{2} \text{ and } \lambda_2 = \alpha - \lambda_1. \quad (6)$$

Moreover,  $\lambda_1$  is clearly positive almost surely. And, it is easy to show that  $\lambda_2$  also is.

We note  $S = \{(v, w) \in \mathbb{R}_+^2, v \geq w\}$  and  $T = \{(v, w) \in \mathbb{R}_+^2, v \geq 2\sqrt{w}\}$  two subsets of  $\mathbb{R}_+^2$ . Let  $g_{a'b'}^\theta : S \rightarrow \mathbb{R}_+$  be the joint probability density function of the semi-major axis  $a'$  and semi-minor axis  $b'$  of the random projected ellipse,  $g_{\lambda_1\lambda_2}^\theta : S \rightarrow \mathbb{R}_+$  the joint probability density function of the roots  $\lambda_1$  and  $\lambda_2$  of the random polynomial  $\mathcal{X}(\lambda)$ , and  $g_{\alpha\beta}^\theta : T \rightarrow \mathbb{R}_+$  the one of the  $\mathcal{X}$  coefficients in Eq. (3). In Proposition 2 below, we express the analytical link between  $g_{a'b'}^\theta$  and  $g_{\alpha\beta}^\theta$ .

**Proposition 2** *Let  $\mathcal{X}(\lambda)$  be the random polynomial defined in Eq. (3) whose roots  $(\lambda_1, \lambda_2)$  correspond to the squared semi-axes  $(a'^2, b'^2)$  of the random projected ellipse from the random spheroid defined in Section 2.1. Then, the joint density function  $g_{a'b'}^\theta$  of  $(a', b')$  is given by*

$$g_{a'b'}^\theta(v, w) = g_{\alpha\beta}^\theta(v^2 + w^2, v^2 w^2)(v^2 - w^2)4vw, \quad \forall (v, w) \in S, \quad (7)$$

where  $S = \{(v, w) \in \mathbb{R}_+^2, v \geq w\}$ .

*Proof* By definition, the support of the random vectors  $(\lambda_1, \lambda_2)$  and  $(a', b')$  is the set  $S = \{(v, w) \in \mathbb{R}_+^2, v \geq w\}$ . Let  $h_1 : S \rightarrow S$  be the function defined as

$$h_1 : (v, w) \mapsto (\sqrt{v}, \sqrt{w}). \quad (8)$$

Note that  $(a', b') = h_1(\lambda_1, \lambda_2)$ . Moreover, the function  $h_1$  is a  $C^1$ -diffeomorphism. The inverse function  $h_1^{-1} : S \rightarrow S$  is given by

$$h_1^{-1} : (v, w) \mapsto (v^2, w^2). \quad (9)$$

By applying a change of variables (see Cramér (1946), page 293) we have

$$g_{a'b'}^\theta(v, w) = g_{\lambda_1\lambda_2}^\theta(h_1^{-1}(v, w)) \times |\det(J_1)|, \quad \forall (v, w) \in S, \quad (10)$$

where  $J_1$  is the Jacobian matrix of the inverse transformation given by

$$\begin{pmatrix} 2v & 0 \\ 0 & 2w \end{pmatrix}$$

and  $\det(J_1)$  is its determinant. As  $\det(J_1) = 4vw$ , it reads

$$g_{a'b'}^\theta(v, w) = g_{\lambda_1\lambda_2}^\theta(v^2, w^2)4vw, \quad \forall (v, w) \in S. \quad (11)$$

From Proposition 1, we notice that the support of the random vector  $(\alpha, \beta)$  is the set  $T = \{(v, w) \in \mathbb{R}_+^2, 1 \leq v \leq 2\sqrt{w}\}$ . Let  $h_2 : T \rightarrow S$  be the function defined as

$$h_2 : (v, w) \mapsto \left( \frac{v + \sqrt{v^2 - 4w}}{2}, \frac{v - \sqrt{v^2 - 4w}}{2} \right). \quad (12)$$

Note that  $(\lambda_1, \lambda_2) = h_2(\alpha, \beta)$ . The inverse function  $h_2^{-1} : S \rightarrow T$  is given by

$$h_2^{-1} : (v, w) \mapsto (v + w, vw). \quad (13)$$

Again, following the change of variables defined by the function  $h_2$ , it comes

$$g_{\lambda_1\lambda_2}^\theta(v, w) = g_{\alpha\beta}^\theta(h_2^{-1}(v, w)) \times |\det(J_2)|, \quad \forall (v, w) \in S, \quad (14)$$

where  $J_2$  is the Jacobian matrix of the transformation  $h_2^{-1}$ . Its determinant is  $\det(J_2) = (v - w)$ , and thus

$$g_{\lambda_1\lambda_2}^\theta(v, w) = g_{\alpha\beta}^\theta(v + w, vw)(v - w), \quad \forall (v, w) \in S. \quad (15)$$

The result is obtained by replacing Eq. (15) in Eq. (11).

□

This result demonstrates the equivalence between the joint probability law of  $a'$  and  $b'$  and that of the coefficients of the random polynomial  $\mathcal{X}$  defined in Eq. (3). The following corollary specifies the likelihood  $\mathcal{L}(\theta; \mathbf{y})$ .

**Corollary 1** *Let  $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_k)$  be the set of observed orthogonal projections of the random spheroid with parameter vector  $\theta$ . Then, the likelihood  $\mathcal{L}(\theta; \mathbf{y})$  is given by*

$$\mathcal{L}(\theta; \mathbf{y}) = \prod_{i=1}^k g_{\alpha\beta}^\theta(a_i'^2 + b_i'^2, a_i'^2 b_i'^2)(a_i'^2 - b_i'^2)4a_i' b_i'. \quad (16)$$

where  $g_{\alpha\beta}^\theta$  is the joint probability density function of the coefficients of the random polynomial  $\mathcal{X}$  defined by Eq. (3).

*Proof* Let  $\theta \in \mathbb{R}^d$  be a realization of the parameter vector  $\Theta$ . By definition, we have

$$l(\mathbf{y}_i|\theta) = g_{a'b'}^{\theta}(a'_i, b'_i). \quad (17)$$

From Proposition 2, it comes

$$l(\mathbf{y}_i|\theta) = g_{\alpha\beta}^{\theta}(a_i'^2 + b_i'^2, a_i'^2 b_i'^2)(a_i'^2 - b_i'^2)4a'_i b'_i. \quad (18)$$

As the observations  $\mathbf{Y}_i$  are i.i.d, the likelihood  $\mathcal{L}(\theta; \mathbf{y})$  is then given by

$$\mathcal{L}(\theta; \mathbf{y}) = \prod_{i=1}^k l(\mathbf{y}_i|\theta) \quad (19)$$

$$= \prod_{i=1}^k g_{\alpha\beta}^{\theta}(a_i'^2 + b_i'^2, a_i'^2 b_i'^2)(a_i'^2 - b_i'^2)4a'_i b'_i. \quad (20)$$

□

Consequently, the likelihood  $\mathcal{L}(\theta; \mathbf{y})$  is completely determined by the joint probability density  $g_{\alpha\beta}^{\theta}$  of the coefficients of the random polynomial of Eq. (3). However, the joint law of  $\alpha$  and  $\beta$  is not explicitly known. These coefficients are non-linear functions of both the semi-major axis, the flattening coefficient of the random spheroid, and on the projection plane's orientation (see Eq. (5)). Therefore, analytical expression of their joint probability density function is difficult to obtain. The posterior density is therefore not analytically tractable.

### 3 The approximate posterior mean estimator

Pritchard et al. (1999); Tavaré et al. (1997) were first to introduce the ABC methods when the likelihood is not available in analytical form in order to sample from an approximation of the posterior distribution. The simplest approach of ABC is based on a rejection technique by bypassing the calculation of the likelihood function. Only samples according to the conditional distribution of the observations given the set of parameters are required.

If  $y \sim l(\cdot|\theta)$  and if  $f(\cdot)$  is the prior density then the proposed rejection method follows the three steps below:

1. sample  $\theta' \sim f(\cdot)$  ;
2. sample  $\mathbf{y}' \sim l(\cdot|\theta')$  ;
3. accept  $\theta'$  if  $\rho(S(\mathbf{y}'), S(\mathbf{y})) \leq \delta$ .

In the third step,  $S : \mathbb{R}^{2 \times k} \rightarrow \mathbb{R}_+^s$  is a function that calculates a set of  $s$  statistics on the  $\mathbf{y}$  data (*e.g.* mean, standard deviation, etc.) and the function  $\rho : \mathbb{R}^s \times \mathbb{R}^s \rightarrow \mathbb{R}^+$  measures the difference between the statistics of the simulated data  $\mathbf{y}'$  and those observed  $\mathbf{y}$ . In other words,  $\theta'$  is considered as a representative sample if the distance between the statistics of the simulated data according to the likelihood and the observed ones is small enough (*i.e.* smaller than a chosen tolerance value  $\delta$ ). If  $\theta'$  is accepted, then it is a sample from the posterior distribution conditionally to  $\rho(S(\mathbf{y}'), S(\mathbf{y})) \leq \delta$ , *i.e.* from  $f(\cdot|\rho(S(\mathbf{y}'), S(\mathbf{y})) \leq \delta)$ . When  $\delta \rightarrow 0$ , we sample conditionally to the summary statistics of the observed data, *i.e.* from  $f(\cdot|S(\mathbf{y}))$ .



Moreover, when the summary statistics  $S$  is *Bayes sufficient* (see Sisson et al. (2018), page 128), the approximate posterior density  $f(\cdot | \rho(S(\mathbf{y}'), S(\mathbf{y})) \leq \delta)$  converges to the true posterior  $f(\cdot | \mathbf{y})$  when  $\delta \rightarrow 0$ . However, the algorithm becomes ineffective in practice if  $\delta$  is too small, resulting in a systematic rejection of all the proposals  $\boldsymbol{\theta}'$ .

The three-step acceptance-rejection procedure is a very general method, but naive in the multidimensional case as it is really inefficient due to the curse of dimensionality. Other sampling methods were proposed to circumvent this problem. Beaumont et al. (2002) considered a *conditional density estimate* based on a regression technique (Ruppert and Wand 1994), Marjoram et al. (2003); Ratmann et al. (2007); Becquet and Przeworski (2007) adopted a *Markov chain Monte Carlo* (MCMC) version, and similarly, Sisson et al. (2007); Beaumont et al. (2009); Toni et al. (2009) implemented a *sequential Monte Carlo* version. An exhaustive summary of these developments falls outside the scope of the present paper. We refer to Beaumont (2010); Sisson et al. (2018) for more details about ABC methods.

In this study, the MCMC-version is used to sample from  $f(\cdot | \rho(S(\mathbf{y}'), S(\mathbf{y})) \leq \delta)$ . The algorithm (see Algorithm 1) is the following: first, an initial set of parameters is considered (for example, a sample from the prior distribution). A Markov chain  $\Theta_1, \Theta_2, \dots$  is then built according to a random walk: at each iteration  $t$ , a new set of parameters,  $\boldsymbol{\theta}'$ , is proposed from an independent perturbation  $q$  of the previous state  $\boldsymbol{\theta}$ . The search space  $D$  is assumed to be a bounded set of  $\mathbb{R}^d$  which contains  $\boldsymbol{\theta}_{\text{real}}$ . The perturbation  $q$  is chosen to be a uniform random vector belonging to an appropriate subset  $P$  centered at the origin. Then,  $\boldsymbol{\theta}'$  is uniformly generated over  $P_{\boldsymbol{\theta}}$ , where  $P_{\boldsymbol{\theta}}$  is the translation of  $P$  by the vector  $\boldsymbol{\theta}$ . Note that for states  $\boldsymbol{\theta}$  close to the boundary of  $D$ , the new proposal  $\boldsymbol{\theta}'$  can fall outside of  $D$  and must therefore be corrected to belong to the space of all possible states, while preserving the symmetry of the proposal density. Then, a set of projected axes  $\mathbf{y}'$  is sampled from  $l(\cdot | \boldsymbol{\theta}')$ . Two scenarios are then possible: i) if the distance  $\rho(S(\mathbf{y}'), S(\mathbf{y}))$  is lower than the chosen tolerance  $\delta$ , the proposition  $\boldsymbol{\theta}'$  is accepted as a new sample from  $f(\cdot | \rho(S(\mathbf{y}'), S(\mathbf{y})) \leq \delta)$  with probability  $\min \left\{ \frac{f(\boldsymbol{\theta}')}{f(\boldsymbol{\theta})}, 1 \right\}$ , ii) otherwise, the previous state  $\boldsymbol{\theta}$  is considered as a new sample. As shown by Marjoram et al. (2003), the stationary distribution of the Markov chain  $\Theta_1, \Theta_2, \dots$  in Algorithm 1 is  $f(\cdot | \rho(S(\mathbf{y}'), S(\mathbf{y})) \leq \delta)$ .

Note that a new proposal of parameter vector is obtained from the current state  $\boldsymbol{\theta}$  by an independent uniform perturbation within  $P_{\boldsymbol{\theta}}$ . This enables to not be too restrictive on the part of the space to explore providing that the size of  $P_{\boldsymbol{\theta}}$  is not too small. However, it is difficult to give a general rule of thumb for the choice of  $P$  as it may influence the autocorrelation and the mixing of the chain. Regarding the domain  $D$  of all possible states, it should be fixed as small as possible while ensuring  $\boldsymbol{\theta}_{\text{real}} \in D$ . This will lead to a better convergence of the chain, see also Rubio et al. (2013).

The APM estimator is computed by averaging the samples from the approximation of the posterior density. Note that other statistics such as the approximate posterior median could also be computed directly from these samples. The APM estimator is only used here as a reference with respect to the AMAP which requires two additional calculation steps described below.

**Algorithm 1** The APM estimator.**Initialization:**

Choose a vector of initial parameters  $\theta_0$ .

**Repeat:** At iteration  $t \geq 1$ ,  $\Theta_t = \theta$ ,

1) Sample the parameters according to a random walk:

$$\theta' = \theta + q$$

where  $q$  is random number uniformly distributed over  $P$ . A correction is made if  $\theta' \notin D$ .

2) Sample  $\mathbf{y}' \sim l(\cdot|\theta')$  using Eq. (5) and (6).

3) If  $\rho(S(\mathbf{y}'), S(\mathbf{y})) \leq \delta$  then:

– generate  $u \sim \mathcal{U}(0, 1)$ ,

– compute  $h = \frac{f(\theta')}{f(\theta)}$  and accept  $\theta'$  if  $u \leq h$ , i.e.  $\Theta_{t+1} = \theta'$ . Otherwise, stay at  $\theta$ , i.e.  $\Theta_{t+1} = \theta$ .

4) If  $\rho(S(\mathbf{y}'), S(\mathbf{y})) \geq \delta$ , keep  $\theta$ , i.e.  $\Theta_{t+1} = \theta$ .

**Return:** the APM estimator is obtained by averaging these samples.

**4 The approximate maximum a posteriori estimator**

We recall that the MAP,  $\theta_{\text{MAP}}$ , is the value which maximizes the posterior density for the realized data, i.e.  $f(\cdot|\mathbf{y})$ . Rubio et al. (2013) proposed a natural approach to compute an approximation of the MAP which consists of using a kernel density with previous samples to get a parametric approximation of the posterior density and then maximizing this kernel density estimator. Note that if the prior  $f(\cdot)$  is chosen to be uniform on a compact set  $D$ , then the MAP estimator coincides with the maximum likelihood estimator. Therefore, the following approach can also serve as a method to get an approximation of the maximum likelihood estimate.

A kernel density estimator  $\hat{f}_n(\cdot|\rho(S(\mathbf{y}'), S(\mathbf{y})) \leq \delta)$  of  $f(\cdot|\rho(S(\mathbf{y}'), S(\mathbf{y})) \leq \delta)$  can be defined from the  $n$  samples  $\theta_1, \dots, \theta_n$  by

$$\hat{f}_n(\theta|\rho(S(\mathbf{y}'), S(\mathbf{y})) \leq \delta) = \frac{1}{n} |\det(H)|^{-1/2} \sum_{i=1}^n K\left(H^{-1/2}(\theta - \theta_i)\right), \forall \theta \in D, \quad (21)$$

where  $K$  is a kernel (a symmetric non-negative real-valued integrable function whose integral equals 1 on  $\mathbb{R}^d$ ) and  $H$  is a  $4 \times 4$  symmetric and positive definite matrix.  $H$  is called the *bandwidth matrix* of the kernel density estimator. Note that it is assumed for simplicity that  $H$  is diagonal, i.e.  $H = \text{diag}(h^{(1)}, \dots, h^{(4)})$ . The estimator  $\hat{f}_n(\cdot|\rho(S(\mathbf{y}'), S(\mathbf{y})) \leq \delta)$  in Eq. (21) is only an approximation of the true posterior density. However, as it will be explained thereafter,  $\hat{f}_n(\cdot|\rho(S(\mathbf{y}'), S(\mathbf{y})) \leq \delta)$ , under certain conditions, is an appropriate approximation of the true posterior density.

The non-parametric estimator in Eq. (21) can be maximized. Its maximum,  $\theta_{\text{AMAP}}$ , is the *approximate maximum a posteriori* (AMAP). It serves as an approximation of  $\theta_{\text{MAP}}$ . In practice, the maximization of the kernel density is achieved numerically to obtain an estimator  $\hat{\theta}_{\text{AMAP}}$  of  $\theta_{\text{AMAP}}$ . Indeed, as shown by Rubio et al. (2013) (Corollary 2), for a set of sufficient statistics  $S$ , and any level of precision  $\gamma > 0$ , there exists  $\delta$  such that, almost surely,

$$\lim_{n \rightarrow \infty} \left| \hat{f}_n(\hat{\theta}_{\text{AMAP}}|\rho(S(\mathbf{y}'), S(\mathbf{y})) \leq \delta) - f(\theta_{\text{MAP}}|\mathbf{y}) \right| \leq \gamma. \quad (22)$$

This result holds if  $\hat{\boldsymbol{\theta}}_{\text{AMAP}} \xrightarrow[n \rightarrow \infty]{} \boldsymbol{\theta}_{\text{AMAP}}$  almost surely and under some mild continuity assumptions of the approximate posterior distribution. Obviously, one crucial assumption is that  $\boldsymbol{\theta}_{\text{MAP}}$  exists, meaning that the posterior density has a unique maximizer. Here, the existence is guaranteed as the posterior density is assumed to be continuous on the compact set  $D$ . However, uniqueness is not ensured in general. Under these conditions,  $\hat{\boldsymbol{\theta}}_{\text{AMAP}}$  is close to  $\boldsymbol{\theta}_{\text{MAP}}$  for an increasing number of samples  $n$  and sufficiently small  $\delta$ . This result justifies the use of such an estimator  $\hat{\boldsymbol{\theta}}_{\text{AMAP}}$  for the MAP.

The estimation of the AMAP is detailed in Algorithm 2. Step 1) has been described in Algorithm 1. Regarding step 2), a fundamental parameter to be determined is the bandwidth matrix  $H$  (see Eq. (21)).  $H$  accounts for the degree of smoothing: the higher the value of the diagonal elements, the smoother the function determined by the kernel density estimation. Several methods exist for bandwidth selection (Duong and Hazelton 2003; Zhang et al. 2006; Zougab et al. 2014). Among them, the *Silverman's rule of thumb* (Silverman 2018) gives

$$h^{(i)} = \left( \frac{4}{(d+2)n} \right)^{\frac{1}{d+4}} \sigma^{(i)}, \quad (23)$$

where  $\sigma^{(i)}$  stands for the standard deviation of the  $i$ th element of  $\boldsymbol{\theta}$ . In practice,  $\sigma^{(i)}$  is the empirical standard deviation estimated from the samples. This model is especially relevant when the true probability distribution is close to Gaussian. In more general cases, the Silverman rule of thumb may lead to a poor kernel density estimator, and a more general method has to be chosen such as the so-called *likelihood cross-validation* (Brewer 2000; Zhang et al. 2006; Sugiyama 2015). The later consists in maximizing the pseudo-likelihood given by

$$L(\mathbf{x}_1, \dots, \mathbf{x}_n | H) = \sum_{i=1}^n \log \left( \frac{1}{n} \sum_{j=1}^n |H|^{-1/2} K \left( H^{-1/2} (\mathbf{x}_i - \mathbf{x}_j) \right) \right). \quad (24)$$

It is worth noting that the maximization of  $L$  in Eq. (24) leads to a zero bandwidth matrix  $H$ , as  $x_i$  can be equal to  $x_j$ . To overcome this problem, the  $k$ -fold cross-validation estimator of the pseudo-likelihood is preferred (Sugiyama 2015; Goodfellow et al. 2016), that also helps reducing the bias. Note that the correlation of the Markov chain may affect the estimation of the bandwidth matrix. In practice, it is not as problematic as might be expected, especially if the sample set is large enough (Rubio et al. 2013; Sköld and Roberts 2003). The choice of the kernel type is in general not crucial. Therefore, in the following, a multivariate Gaussian kernel is chosen. The last step of Algorithm 2 is also achieved by a direct search optimization algorithm called *pattern search* (Davidon 1991; Abramson 2002).

## 5 Numerical Validation

Rubio et al. (2013) performed numerical experiments in simple cases to highlight the use of the AMAP estimator when the likelihood is intractable. Additional numerical experiments are introduced to further emphasize the performances of

**Algorithm 2** Estimation of the AMAP

- 
- 1) Obtain  $n$  samples  $\theta_1, \dots, \theta_n$  from  $f(\cdot | \rho(S(\mathbf{y}'), S(\mathbf{y})) \leq \delta)$  (see Algorithm 1).
  - 2) Compute the optimal bandwidth matrix  $H_0$  using a likelihood cross-validation procedure and construct the kernel density estimator  $\hat{f}_n(\cdot | \rho(S(\mathbf{y}'), S(\mathbf{y})) \leq \delta)$  using Eq. (21).
  - 3) Compute the maximum  $\hat{\theta}_{\text{AMAP}}$  of  $\hat{f}_n(\cdot | \rho(S(\mathbf{y}'), S(\mathbf{y})) \leq \delta)$  using an optimization algorithm.
- 

the AMAP estimator in our setting. Comparison with the APM estimator performances is also provided in a specific case study. The influence of three parameters is considered: the tolerance  $\delta$ , the number of samples  $n$ , and the number of observations  $k$ . Finally, a numerical experiment is performed to assess the influence of the set of statistics.

## 5.1 Assumptions and ground truth data

We consider a random spheroid with a semi-major axis following a gamma law  $a \sim \Gamma(4, 0.05)$  and a semi-minor axis  $b = \epsilon a$ , where the flattening coefficient  $\epsilon$  follows a truncated normal law  $\text{tr}\mathcal{N}(0.85, 0.1; 1/2, 1)$  restricted to the interval  $[1/2, 1]$ , of mean 0.85 and standard deviation 0.1. The true set of parameters is  $\theta_{\text{real}} = (\mu_1, \sigma_1, \mu_2, \sigma_2)$  where  $\mu_1 = 0.2$ ,  $\sigma_1 = 0.1$ ,  $\mu_2 = 0.85$  and  $\sigma_2 = 0.1$ . Note that, the choice of a gamma probability law can be used to describe the diameter distribution of spherical particles or the semi-major axis of spheroid-like particles with low elongation in the context of multiphase flows description.

A number of  $k$  realizations of the oriented random spheroid are sampled. For each realization, the semi-axes of the projected ellipses,  $\mathbf{a}' = (a'_1, \dots, a'_k)$  and  $\mathbf{b}' = (b'_1, \dots, b'_k)$ , are obtained from Eq. (6). The set of  $k$  projected ellipses constitutes the observations from which the AMAP estimator is constructed.

The vector of initial parameters  $\theta_0$  for the ABC sampler is calculated from the observations of the projected ellipses as follows:

$$\theta_0 = (\hat{\mu}(\mathbf{a}'), \hat{\sigma}(\mathbf{a}'), \hat{\mu}(\mathbf{b}'/\mathbf{a}'), \hat{\sigma}(\mathbf{b}'/\mathbf{a}')) , \quad (25)$$

where  $\hat{\mu}$  and  $\hat{\sigma}$  are the empirical mean and standard deviation, i.e.

$$\left\{ \begin{array}{l} \hat{\mu}(\mathbf{a}') = \frac{1}{k} \sum_{i=1}^k a'_i, \quad \hat{\sigma}(\mathbf{a}') = \sqrt{\frac{1}{k-1} \sum_{i=1}^k (a'_i - \hat{\mu}(\mathbf{a}'))^2}, \\ \hat{\mu}(\mathbf{b}'/\mathbf{a}') = \frac{1}{k} \sum_{i=1}^k \frac{b'_i}{a'_i} \text{ et } \hat{\sigma}(\mathbf{b}'/\mathbf{a}') = \sqrt{\frac{1}{k-1} \sum_{i=1}^k \left( \frac{b'_i}{a'_i} - \hat{\mu}(\mathbf{b}'/\mathbf{a}') \right)^2}. \end{array} \right. \quad (26)$$

As it is a numerical study, the prior density  $f : D \rightarrow \mathbb{R}^+$  is chosen as the probability density function of a continuous multivariate uniform distribution for convenience. A burn-in period of 100 steps is used to mitigate the choice of the vector of initial parameters. It is not necessary to have longer burn-in period as the initial state  $\theta_0$  defined in Eq. (25) is more likely to bring the Markov chain directly into a high probability region than a vector sampled from the multivariate uniform prior

over  $D$ . The support  $D$  of the parameters vector  $\boldsymbol{\theta}$  is defined, with respect to the observations, by

$$D = (0, \max_i(a'_i)) \times (0, 3\hat{\sigma}(\mathbf{a}')) \times (0.5, \max_i(b'_i/a'_i)) \times (0, 3\hat{\sigma}(\mathbf{b}'/\mathbf{a}')). \quad (27)$$

It is better to consider a small support  $D$ , while ensuring that the estimated parameter vector  $\boldsymbol{\theta}_{\text{real}}$  belongs to  $D$ . Note that, with a uniform prior, the ratio  $h$  in Algorithm 1 is always equal to 1, and the new state  $\boldsymbol{\theta}'$  is automatically accepted if  $\rho(S(\mathbf{y}'), S(\mathbf{y})) \leq \delta$ . However, this setting (uniform prior and uniform perturbation) is not equivalent to the three-step acceptance-rejection procedure described in Section 3 as the perturbation is made in the neighborhood  $P_{\boldsymbol{\theta}}$  of the current state  $\boldsymbol{\theta}$ . We denote  $l_1, \dots, l_4$  the four lengths of  $D$ , i.e.

$$\begin{aligned} l_1 &= \max_i(a'_i), \\ l_2 &= 3\hat{\sigma}(\mathbf{a}'), \\ l_3 &= \max_i(b'_i/a'_i) - 0.5, \\ l_4 &= 3\hat{\sigma}(\mathbf{b}'/\mathbf{a}'). \end{aligned}$$

Then, the set  $P$  which defines the support of the perturbation is here fixed as

$$P = \frac{1}{8} \prod_{i=1}^4 (-l_i, l_i). \quad (28)$$

Note that no sub-sampling procedure has been used for the Markov chain.

The  $\rho$  function which measures the deviation between the statistics of the sampled semi-axes and the observations is defined as the following

$$\begin{aligned} \rho : \mathbb{R}^s \times \mathbb{R}^s &\longrightarrow \mathbb{R}_+ \\ (x, y) &\longmapsto \max_{1 \leq i \leq s} \left( \left| \frac{x_i}{y_i} - 1 \right| \right), \end{aligned}$$

where  $s \in \mathbb{N}^*$  is the number of statistics. This function was initially proposed by Pritchard et al. (1999). Although the semi-major axis and the flattening coefficient of the random spheroid realizations are not observed, a natural choice of statistics  $S$  is their counterparts for the projected ellipses. Therefore, four statistics are first considered: the mean and standard deviation of the semi-major axis, and those of the flattening coefficient of the projected ellipses, as defined in Eq. (26).

## 5.2 Results

### 5.2.1 Influence of the tolerance $\delta$

The tolerance  $\delta$  has been shown to have an impact on the quality of the approximation of the posterior density approximation. Therefore, different values of  $\delta$  were tested: 0.8, 0.6, 0.4 and 0.2. In each case, we sampled 20 times from the approximate posterior distribution  $f(\cdot | \rho(S(\mathbf{y}'), S(\mathbf{y})) \leq \delta)$ , and computed the AMAP and APM estimators for each sample. The sample size is  $n = 50,000$  and the number of observed ellipses is  $k = 4,000$ . Fig. 2 gives the violin plots of both the AMAP

estimator (in blue) and the APM estimator (in green) for the four parameters. In each graph, the solid black line within the various violin plots corresponds to the median of the 20 estimates, while the 25th and 75th percentiles are highlighted with dashed black lines, and the target values with red lines.

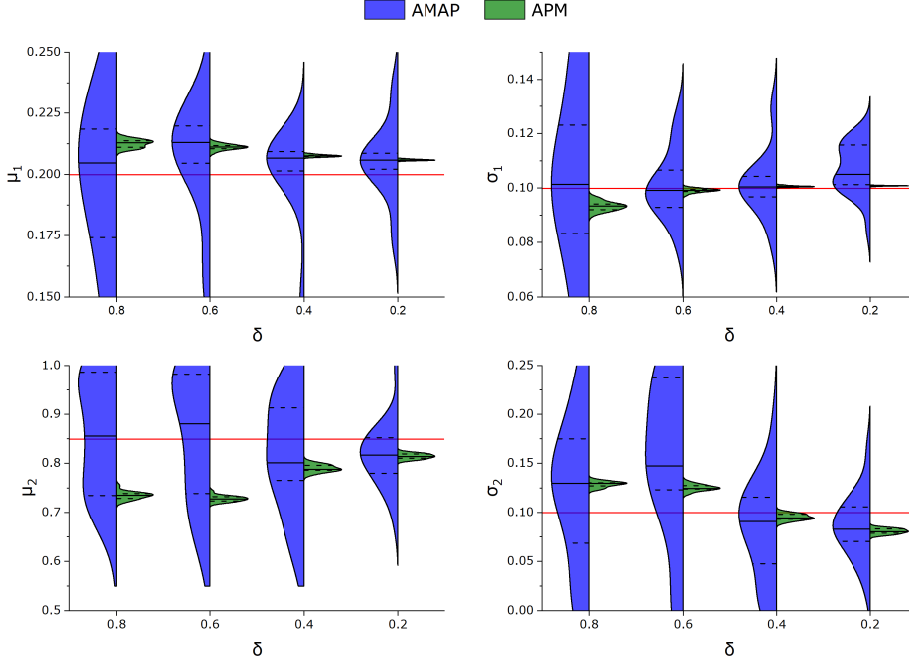


Fig. 2: Performances of the AMAP estimator (blue) and the APM estimator (green) with respect to the tolerance  $\delta$  for  $n=50,000$  and  $k=4,000$ . The median, the 25th and 75th percentiles are highlighted in solid black line and dashed black lines, respectively. The red lines correspond to the target values.

A decrease of the dispersion of the distribution is evidenced as  $\delta$  decreases for the two estimators. The distributions tend to be more concentrated around the corresponding theoretical values. This is not surprising as  $f(\cdot|\rho(S(\mathbf{y}'), S(\mathbf{y})))$  is a better approximation of  $f(\cdot|\mathbf{y})$  when  $\delta$  decreases, as indicated in Section 3. However, it can be noted that the 25th-75th percentile interval of the APM estimator does not systematically include the target value.

The mean absolute percentage error (mape) is given by

$$\text{mape}_j = \frac{1}{20} \sum_{i=1}^{20} \frac{|\widehat{\theta}_{i,j} - \theta_{\text{theo}}^{(j)}|}{\theta_{\text{theo}}^{(j)}} \times 100, \quad j \in \{1, \dots, 4\}, \quad (29)$$

where  $\widehat{\theta}_{i,j}$  is the estimator of parameter  $j$  for simulation  $i$ , and  $\theta_{\text{theo}}^{(j)}$  its target value. In the case  $\delta = 0.2$ , and for both estimators, the mape is very low ( $\leq 9\%$ ) for the parameters  $\mu_1$ ,  $\sigma_1$  and  $\mu_2$ , while it is higher (lower than 25% and 20% for the AMAP and APM respectively) for the standard deviation of the flattening

coefficient  $\sigma_2$ , see Table 1. One explanation is that the chosen summary statistics are not fully informative for the posterior distribution, see Section 5.2.4. Also, the tolerance  $\delta$  may still be too high, leading to a too coarse approximation of the posterior distribution to allow the parameter  $\sigma_2$  to be found precisely. The trend of the violin plots suggests that the bias and dispersion should decrease for smaller  $\delta$ . However, for such low  $\delta$  values, the computation time of the AMAP estimator rises significantly. One can expect that the AMAP estimator is even more efficient for lower value of  $\delta$ , but at the price of computation time.

The APM estimator has a slightly higher bias than the AMAP estimator for  $\mu_1$ ,  $\mu_2$  and  $\sigma_2$ , but its variance is much lower. This leads to a root mean square error (rmse) which is lower for the APM estimator, see Table 1. The higher variance of the AMAP is due to the implementation of a kernel density method for approximating the posterior density surface, which is known to be less robust for high dimension (typically  $d \geq 3$ ). Note that the computation time of the AMAP estimator is higher as it requires the estimation of the bandwidth matrix  $H$  with a  $k$ -fold cross-validation procedure and the use of a numerical optimization method.

Table 1: Bias, rmse and mape criteria of both AMAP and APM estimators for tolerance parameter  $\delta = 0.2$ .

		$\mu_1$	$\sigma_1$	$\mu_2$	$\sigma_2$
<b>Bias</b>	AMAP	0.0055	0.0070	-0.0320	-0.0141
	APM	0.0059	0.0009	-0.0355	-0.0196
<b>rmse</b>	AMAP	0.0158	0.0112	0.0713	0.0340
	APM	0.0059	0.0009	0.0358	0.0197
<b>mape</b>	AMAP	5.9	8.6	6.4	26.9
	APM	3	0.9	4.2	19.6

### 5.2.2 Influence of the number of samples $n$

To assess the influence of the sample size, five increasing values of  $n$  were tested: 10,000, 20,000, 30,000, 40,000 and 50,000. The tolerance  $\delta$  was fixed to 0.2, and the number of observations  $k$  to 4,000. Fig. 3 gives the violin plots of the estimations for each parameter and number of samples  $n$ .

No clear improvement of the performance of the estimators is highlighted when  $n$  increases. For the given tolerance value ( $\delta = 0.2$ ), 10,000 samples may be enough to reach such levels of bias. For both estimators, the mape, see Eq. (29), is similar to that of the previous numerical test for all parameters, regardless of the number of samples.

### 5.2.3 Influence of the number of observations $k$

The number of observations influences the quality of the information contained in the summary statistics. In practice, to assess this effects, a database of  $k_{max} = 4,000$  observations has been created, and subdivided into six shuffled parts of 100,

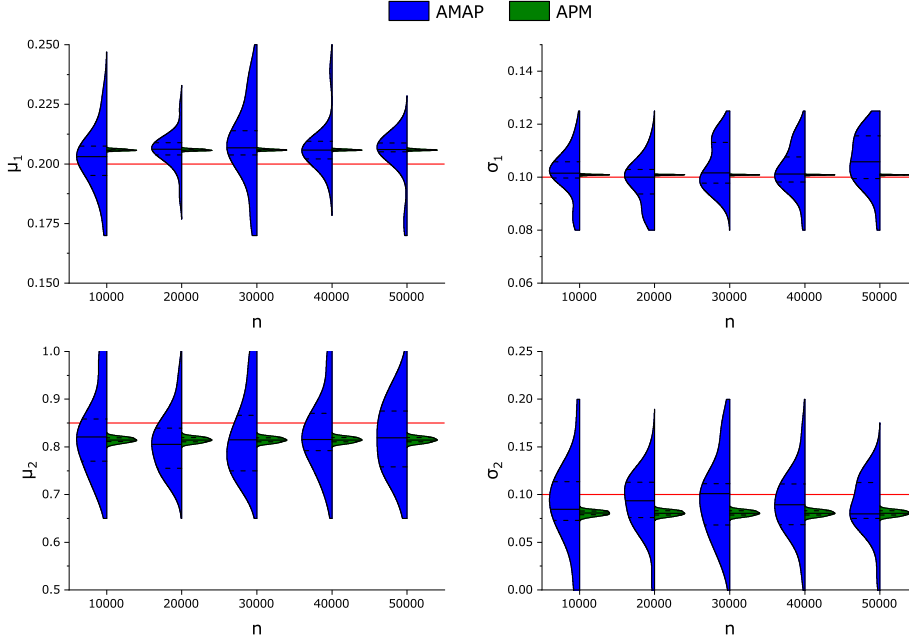


Fig. 3: Performances of the AMAP estimator (blue) and the APM estimator (green) with respect to the number of samples  $n$  for  $\delta = 0.2$  and  $k = 4,000$ . The median, the 25th and 75th percentiles are highlighted in solid black line and dashed black lines, respectively. The red lines correspond to the target values.

500, 1000, 2,000, 3,000 and 4,000 observations. The tolerance  $\delta$  is fixed to 0.2 and  $n = 50,000$ .

As evidenced by Fig. 4, there is a clear reduction of the bias between  $k = 100$  and  $k = 500$ , especially for the mean and standard deviation of the semi-major axis  $\mu_1$  and  $\sigma_1$ . For a higher number of observations, the gain in reducing the bias is less obvious. For the AMAP estimator, a decrease of the dispersion of  $\mu_1$ ,  $\mu_2$  and  $\sigma_2$  can be observed in Fig. 4, as suggested by the relative interquartile ranges reported in Table 2. The relative interquartile ranges  $IQR_j$ , for the parameter  $\theta_{\text{theo}}^{(j)}$  and depending on the numbers of samples, is defined as

$$IQR_j = \frac{Q_3 - Q_1}{m_j} \times 100, \quad j \in \{1, \dots, 4\}, \quad (30)$$

where  $Q_3$  and  $Q_1$  are respectively the 75- and 25-percentile, and  $m_i$  is the median of the estimations for the parameter  $\theta_{\text{theo}}^{(j)}$ .

The effect of an increasing number of observations is not significant beyond  $k > 500$ . This is not really surprising, given the considered summary statistics (means and standard deviations of the ellipses semi-axes), several hundreds of observations are sufficient to summarize the information.



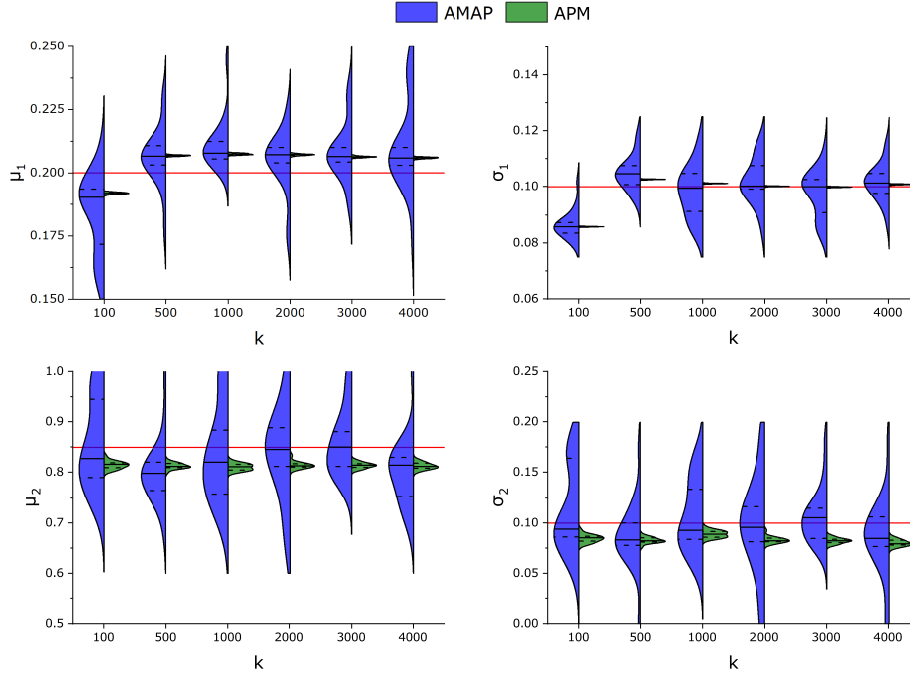


Fig. 4: Performances of the AMAP estimator (blue) and the APM estimator (green) with respect to the number of observations  $k$  for  $\delta = 0.2$  and  $n = 50,000$ . The median, the 25th and 75th percentiles are highlighted in solid black line and dashed black lines, respectively. The red lines correspond to the target values.

Table 2: Interquartile range of the estimations with respect to the number of observations  $k$  for the AMAP estimator.

	100	500	1000	2000	3000	4000
$\hat{\mu}_1$	11	4	3	3	3	3
$\hat{\sigma}_1$	4	6	14	8	11	7
$\hat{\mu}_2$	18	7	15	9	8	9
$\hat{\sigma}_2$	83	27	54	37	29	34

#### 5.2.4 Influence of the summary statistics

As mentioned in Section 3, the choice of summary statistics is crucial for the approximation of the posterior distribution. However, there is no general rule for determining the number or even the form of these statistics. The choice is often suggested by the problem itself and its underlying application. On the one hand, an insufficient number of statistics may not contain enough information to reliably infer on the random spheroid. The problem is said to be *sub-determined*. On the other hand, a large number of statistics with respect to the number of parameters to be determined, can interfere with important information and misconduct the sampler's convergence in Algorithm 1. The problem is in this case *over-determined*.

In order to highlight the effect of the number of statistics, two additional tests were carried out:

1. *test 1*: three statistics, instead of the four previously used, are selected:

$$S = (\hat{\mu}(a'), \hat{\sigma}(a'), \hat{\mu}(b'/a')) ;$$

2. *test 2*: five statistics are taken into account:

$$S = (\hat{\mu}(a'), \hat{\sigma}(a'), \hat{\mu}(b'/a'), \hat{\sigma}(b'/a'), \hat{\mu}_{aire})$$

where  $\hat{\mu}_{aire}$  is the average area of the projected ellipses.

The AMAP estimators were computed for  $k = 4,000$ ,  $\delta = 0.2$  and  $n = 50,000$  for the two tests. The violin plots, obtained for 20 optimizations for each test, are compared in Fig. 5. The *test 0* is the results of the "reference" case study (i.e. with the 4 statistics considered so far).

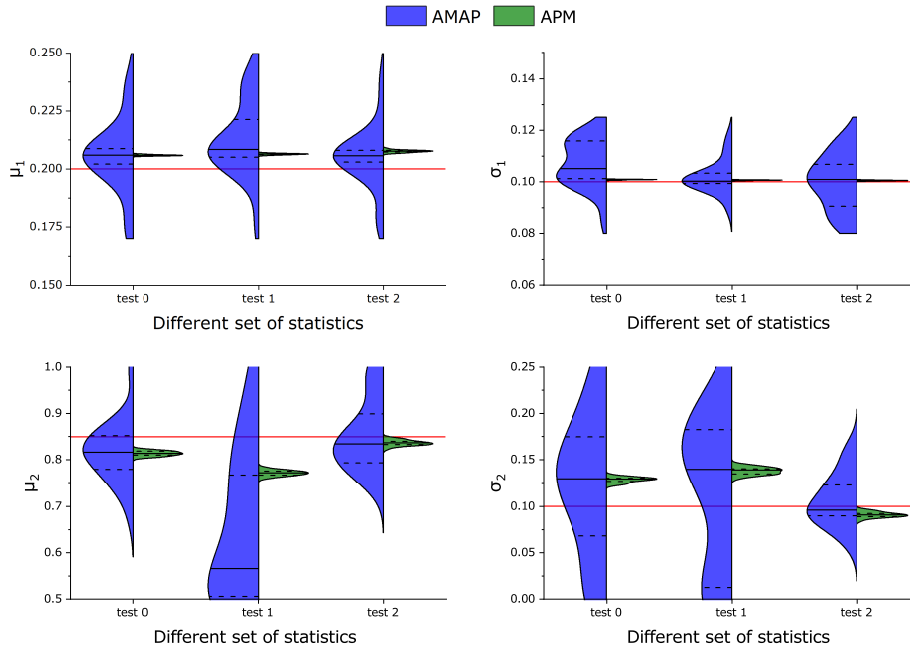


Fig. 5: Performances of the AMAP estimator (blue) and the APM estimator (green) with respect to the choice of statistics for  $\delta = 0.2$ ,  $k = 4,000$  and  $n = 50,000$ . The median, the 25- and 75-percentiles are highlighted in solid black line and dashed black lines, respectively. The red lines correspond to the target values.

It can be seen that both the AMAP and the APM estimators exhibit higher performances for *test 2*. The addition of the information on the average area of the projected ellipses enables to reduce the bias for all estimated parameters compared to the "reference" *test 0*, without impact on the dispersion as highlighted by the

rmse reported in Table 3. The mean absolute percentage error is also decreased with less than 10% for all parameters in the case of the APM estimator. For the AMAP estimator, the mape of  $\sigma_2$  has also decreased compared to *test 0*, but is still higher (around 20%) than the mape of the other parameters, see Table 3. Note also that the bias of the AMAP estimator is lower than the APM estimator, and that the 25th and 75th percentile interval of the APM estimator does not systematically include its target value as previously mentioned.

On the other hand, we can clearly observe from Fig. 5 that the reduction in the number of statistics (*test 1*) leads to a higher bias and dispersion, especially for the parameters  $\hat{\mu}_2$  and  $\hat{\sigma}_2$  of the AMAP estimator. This can be explained by a lack of information, as the standard deviation of the flattening coefficient of the projected ellipses has been removed.

Table 3: Bias, rmse and mape criteria of both AMAP and APM estimators for *test 2*.

		$\mu_1$	$\sigma_1$	$\mu_2$	$\sigma_2$
<b>Bias</b>	AMAP	0.0051	-0.0003	0.0062	0.0069
	APM	0.0076	0.0005	-0.0129	-0.0089
<b>rmse</b>	AMAP	0.0137	0.0102	0.0743	0.0281
	APM	0.0076	0.0005	0.0140	0.0093
<b>mape</b>	AMAP	5	8.1	6.8	22
	APM	3.8	0.5	1.5	9

## 6 Conclusion

In this study, a Bayesian approach is proposed to estimate the parameters of the semi-axis probability laws of a random spheroid, using its observed orthogonal projections. This method enables to infer 3D geometrical characteristics of ellipsoid-like particle systems, from 2D projection image, for which the particles are independent and identical distributed realizations of the considered random spheroid. The main assumption is on the probability laws of the semi-axes, which requires some *a priori* information in practical application. Two estimators are compared, an approximation of the maximum a posteriori and an approximation of the posterior mean, called respectively the AMAP and APM. Both estimators require a Markov chain Monte Carlo procedure to sample from an approximation of the posterior distribution (the ABC algorithm). Besides, an optimization algorithm is implemented to maximize a kernel density estimator of the approximate posterior density in the case of the AMAP estimator.

One important result is Eq. (16), which establishes the relationship between the observed projected ellipses and the random spheroid parameters. This new stereological result can be implemented (for example, with a maximum likelihood estimation technique) if the joint probability of  $\alpha$  and  $\beta$  are available in analytical form, or can be approximated. Moreover, the algorithms 1 and 2 provide an approximation of the spheroid parameters with the APM and AMAP estimators, respectively. Furthermore, the AMAP estimator can provide an approximation of the maximum likelihood if the prior is chosen to be uniform.

This algorithm has been numerically validated in the case of a gamma distribution and a truncated normal distribution for the semi-major axis and the flattening coefficient, respectively. Satisfying results were achieved with less than 10% of maximum absolute percentage error for most of the parameters in the case of the five statistics for both estimators (except for  $\sigma_2$  of the AMAP estimator). It has been shown that reducing the tolerance  $\delta$  results in a much more statistically efficient estimators. Simulations have also shown that, for  $\delta = 0.2$ , a number of 10,000 samples seems to be enough to achieve relatively low bias. Besides, the number of observations does not impact significantly the performances of both estimators for the chosen summary statistics. The numerical experiments suggest that between 500-1,000 observed ellipses may be enough for the proposed "reference" summary statistics. Furthermore, the comparison between the AMAP and the APM estimators highlights the good performances of the relatively simple APM estimator for all numerical tests. It has been highlighted that the APM estimator generally has a much lower variance and a slightly higher bias than the AMAP estimator. In addition, the 25-75th percentile interval of the APM estimator does not systematically include its target value. This point must be taken into account in a metrological use of this approach, especially in the context of multi-phase flows characterization, notably if the systematic bias is not directly known. Note that other performances of the estimators may be expected for very different choice of distribution type of the semi-major axis and flattening coefficient, like for heavy-tail distributions for example.

One important consideration, and possible limitation, of the ABC algorithm is the selection of sufficient statistics. It is in general difficult to ensure that the chosen statistics are Bayes sufficient. However, the simulation results show that relatively low bias and dispersion can be achieved for both estimators when choos-

ing the means and the standard deviations of the projected semi-axes as set of statistics, and even lower root mean squared error is achieved if the average area of the projected ellipses is added. Besides, four parameters were recovered in these simulations. For higher dimension  $d$ , a representative sample set of the posterior distribution can be more difficult to obtain as it would require a significant number of samples, hence increasing the computation time.

## References

- Abramson M (2002) Pattern search algorithms for mixed variable general constrained optimization problems. PhD thesis, Department of Computational and Applied Mathematics, Rice University
- Beaumont M, Zhang W, Balding D (2002) Approximate Bayesian computation in population genetics. *Genetics* 162(4):2025–2035
- Beaumont MA (2010) Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics* 41:379–406
- Beaumont MA, Cornuet JM, Marin JM, Robert CP (2009) Adaptive approximate Bayesian computation. *Biometrika* 96(4):983–990
- Becquet C, Przeworski M (2007) A new approach to estimate parameters of speciation models with application to apes. *Genome Research* 17(10):1505–1519
- Bresler Y, Macovski A (1987) Three-dimensional reconstruction from projections with incomplete and noisy data by object estimation. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 35(8):1139–1152
- Brewer MJ (2000) A Bayesian model for local smoothing in kernel density estimation. *Statistics and Computing* 10(4):299–309
- Cramér H (1946) *Mathematical methods of statistics (PMS-9)*, vol 9. Princeton University Press
- Davidon WC (1991) Variable metric method for minimization. *SIAM Journal on Optimization* 1(1):1–17
- Duong T, Hazelton M (2003) Plug-in bandwidth matrices for bivariate kernel density estimation. *Journal of Nonparametric Statistics* 15(1):17–30
- Fessler JA, Macovski A (1991) Object-based 3-D reconstruction of arterial trees from magnetic resonance angiograms. *IEEE Transactions on Medical Imaging* 10(1):25–39
- Goodfellow I, Bengio Y, Courville A (2016) *Deep learning*. MIT press
- Jaggi S, Karl WC, Willsky AS (1995) Estimation of dynamically evolving ellipsoids with applications to medical imaging. *IEEE Transactions on Medical Imaging* 14(2):249–258
- Karl WC, Verghese GC, Willsky AS (1994) Reconstructing ellipsoids from projections. *CVGIP: Graphical Models and Image Processing* 56(2):124–139
- Kracht W, Emery X, Paredes C (2013) A stochastic approach for measuring bubble size distribution via image analysis. *International Journal of Mineral Processing* 121:6–11
- de Langlard M, Al-Saddik H, Charton S, Debayle J, Lamadie F (2018a) An efficiency improved recognition algorithm for highly overlapping ellipses: application to dense bubbly flows. *Pattern Recognition Letters* 101:88–95
- de Langlard M, Lamadie F, Charton S, Debayle J (2018b) A 3D stochastic model for geometrical characterization of particles in two-phase flow applications. *Image Analysis & Stereology* 37:233–247
- Launeau P, Robin PYF (2005) Determination of fabric and strain ellipsoids from measured sectional ellipses - implementation and applications. *Journal of Structural Geology* 27(12):2223–2233
- Liu R, Ruiter N, Schwarzenberg G, Zapf M, Stotzka R, Gemmeke H (2006) 2J-2 3D ultrasound computer tomography: Results with a clinical breast phantom. In: 2006 IEEE Ultrasonics Symposium, IEEE, pp 989–992
- Marjoram P, Molitor J, Plagnol V, Tavaré S (2003) Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences* 100(26):15324–15328
- Merola F, Miccio L, Memmolo P, Di Caprio G, Galli A, Puglisi R, Balduzzi D, Coppola G, Netti P, Ferraro P (2013) Digital holography as a method for 3D imaging and estimating the biovolume of motile cells. *Lab on a Chip* 13(23):4512–4516

- Noumeir R (1999) Detecting three-dimensional rotation of an ellipsoid from its orthographic projections. *Pattern Recognition Letters* 20(6):585–590
- Ozturk-Isik E, Chen AP, Crane JC, Bian W, Xu D, Han ET, Chang SM, Vigneron DB, Nelson SJ (2009) 3D sensitivity encoded ellipsoidal MR spectroscopic imaging of gliomas at 3T. *Magnetic Resonance Imaging* 27(9):1249–1257
- Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999) Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular Biology and Evolution* 16(12):1791–1798
- Ratmann O, Jørgensen O, Hinkley T, Stumpf M, Richardson S, Wiuf C (2007) Using likelihood-free inference to compare evolutionary dynamics of the protein networks of *H. pylori* and *P. falciparum*. *PLoS Computational Biology* 3(11):2266–2278
- Rubio F, Johansen AM, et al. (2013) A simple approach to maximum intractable likelihood estimation. *Electronic Journal of Statistics* 7:1632–1654
- Ruppert D, Wand MP (1994) Multivariate locally weighted least squares regression. *The Annals of Statistics* 22(3):1346–1370
- Sahagian DL, Proussevitch AA (1998) 3D particle size distributions from 2D observations: stereology for natural applications. *Journal of Volcanology and Geothermal Research* 84(3–4):173–196
- Silverman BW (2018) *Density estimation for statistics and data analysis*. Routledge
- Sisson SA, Fan Y, Tanaka MM (2007) Sequential Monte-Carlo without likelihoods. *Proceedings of the National Academy of Sciences* 104(6):1760–1765
- Sisson SA, Fan Y, Beaumont M (2018) *Handbook of approximate Bayesian computation*. Chapman and Hall/CRC
- Sköld M, Roberts GO (2003) Density estimation for the Metropolis–Hastings algorithm. *Scandinavian journal of statistics* 30(4):699–718
- Sugiyama M (2015) *Introduction to statistical machine learning*. Morgan Kaufmann
- Tavaré S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. *Genetics* 145(2):505–518
- Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MP (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface* 6(31):187–202
- Vickers G (1996) The projected areas of ellipsoids and cylinders. *Powder Technology* 86(2):195–200
- Zhang X, King ML, Hyndman RJ (2006) A Bayesian approach to bandwidth selection for multivariate kernel density estimation. *Computational Statistics & Data Analysis* 50(11):3009–3031
- Zougab N, Adjabi S, Kokonendji CC (2014) Bayesian estimation of adaptive bandwidth matrices in multivariate kernel density estimation. *Computational Statistics & Data Analysis* 75:28–38