Approximate maximum likelihood estimation of the Bingham distribution

Marco Bee[a,*], Roberto Benedetti[b], Giuseppe Espa[c]

[a] *Department of Economics and Management, University of Trento*
[b] *Department of Economic Studies, University of Chieti-Pescara*
[c] *Department of Economics and Management, University of Trento*

Abstract

Maximum likelihood estimation of the Bingham distribution is difficult because the density function contains a normalization constant that cannot be computed in closed form. Given the availability of sufficient statistics, Approximate Maximum Likelihood Estimation (AMLE) is an appealing method that allows one to bypass the evaluation of the likelihood function. The impact of the input parameters of the AMLE algorithm is investigated and some methods for choosing their numerical values are suggested. Moreover, AMLE is compared to the standard approach which numerically maximizes the (approximate) likelihood obtained with the normalization constant estimated via the Holonomic Gradient Method (HGM). For the Bingham distribution on the sphere, simulation experiments and real-data applications produce similar outcomes for both methods. On the other hand, AMLE outperforms HGM when the dimension increases.

Keywords: Directional data, Simulation, Intractable Likelihood, Sufficient statistics

1. Introduction

The Bingham distribution is one of the most important models for directional data. In the three-dimensional case the distribution was introduced by Bingham (1974), who derived its main properties and found exact and asymptotic sampling distributions; see also Mardia and Jupp (2000). Recently, the properties of the large dimensional Bingham distribution have been studied by Kume and Walker (2014). The need of modeling such data arises in many scientific fields, such as geology (Peel et al., 2001), crystallography (Krieger Lassen et al., 1994) and bioinformatics (Kent and Hamelryck, 2005; Hamelryck et al.,

[*] Corresponding author

Email addresses: marco.bee@unitn.it (Marco Bee), roberto.benedetti@unich.it (Roberto Benedetti), giuseppe.espa@unitn.it (Giuseppe Espa)

2006; Boomsma et al., 2008); see also Mardia and Jupp (2000) or Fallaize and Kypraios (2016) and the references therein.

To outline the issue under investigation, we start with a general description of the problem. Consider a q-dimensional random vector $\boldsymbol{X}$ whose density contains a normalization constant depending on $\boldsymbol{\theta}$, where $\boldsymbol{\theta} \overset{\text{def}}{=} (\theta_1, \ldots, \theta_s)' \in \boldsymbol{\Theta} \subset \mathbb{R}^s$ is the parameter vector. Let

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{c(\boldsymbol{\theta})} \exp\{-h(\boldsymbol{x}; \boldsymbol{\theta})\}, \quad x \in \mathbb{R}^q, \tag{1}$$

be the functional form of such a density. If $c(\boldsymbol{\theta})$ cannot be computed in closed form, the most common strategy approximates $\tilde{c}(\boldsymbol{\theta})$ and maximizes the (approximate) likelihood obtained by plugging $\tilde{c}(\boldsymbol{\theta})$ into the likelihood. Distributions with densities that can be written as (1) are commonly encountered not only when working with directional data, but also in spatial statistics (Cressie, 1991, Section 7.2). In this field, MLE based on an approximation of the normalizing constant has been proposed, for example, by Friel and Pettitt (2004). MCMC methods for distributions with intractable normalization constants have been developed by Møeller et al. (2006) and Murray et al. (2006).

Let S_{q-1} be the sphere of unit radius in $\mathbb{R}^q$. The density of a q-dimensional Bingham random vector $\boldsymbol{X}$ with respect to the uniform measure over S_{q-1} is given by

$$f(\boldsymbol{x}; \boldsymbol{A}) = \frac{1}{c(\boldsymbol{A})} \exp\{-\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}\}, \quad \boldsymbol{x}'\boldsymbol{x} = 1, \ \boldsymbol{x} \in \mathbb{R}^q, \tag{2}$$

where $\boldsymbol{A}$ is a $q \times q$ symmetric matrix and $c(\boldsymbol{A})$ is the normalization constant. It is therefore clear that (2) is a special case of (1). The distribution can be derived from the intersection of a zero-mean multivariate normal distribution $\boldsymbol{W} \sim N_p(\boldsymbol{0}, \boldsymbol{\Psi})$ with the unit sphere in $\mathbb{R}^q$, a fact that clarifies the role of the matrix $\boldsymbol{A}$. In this case it turns out that $\boldsymbol{A} = \boldsymbol{\Psi}^{-1}$; in other words, the exponent of (2) is equal to the exponent of a zero-mean multivariate normal.

As $\boldsymbol{A}$ is symmetric, its singular value decomposition is given by $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}'$, where $\boldsymbol{V}$ is orthogonal and $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_q)$. It can be easily verified (Kume and Walker, 2006) that, if $\boldsymbol{X}$ follows a Bingham distribution with density $f(\boldsymbol{x}; \boldsymbol{A})$, the random vector $\boldsymbol{Y} = \boldsymbol{V}'\boldsymbol{X}$ follows a Bingham distribution with density $f(\boldsymbol{x}; \boldsymbol{\Lambda})$. Bingham (1974) has shown that the MLE of $\boldsymbol{V}$ is the matrix of eigenvectors of the sum of squares and products matrix $\sum_{j=1}^n \boldsymbol{x}_j \boldsymbol{x}_j'$, where n is the sample size, so that one can, without loss of generality, restrict attention to MLE of $\boldsymbol{\Lambda}$.

The distribution is antipodally symmetric but not circularly symmetric, and is not identifiable unless we introduce some constraint on $\boldsymbol{\Lambda}$, because (Bingham, 1974, Lemma 2.1) the density does not change if we add a positive constant to the λ_is. Thus, for the remainder of this paper, we will use the constraint $\lambda_q = 0$, and assume $\lambda_1 \geq \cdots \geq \lambda_q = 0$.

Exact evaluation of the likelihood corresponding to (2) is difficult because the normalization constant cannot be computed explicitly and depends on $\boldsymbol{\Lambda}$, so that it cannot be ignored. Although various methods have been proposed, numerical approximation of $c(\mathbf{\Lambda})$ is a computationally expensive problem. When $q = 3$, one can use power series and asymptotic series (Bingham, 1964). For a certain range of parameter values, the saddlepoint approximation works well (Kume and Wood, 2005). Finally, Sei and Kume (2015) show that the Holonomic Gradient Method (HGM) is quite accurate.

Having computed an approximate value of $c(\mathbf{\Lambda})$, MLE of the parameters can be performed by plugging it into the likelihood function and numerically maximizing the resulting (approximate) likelihood function. This is also known as approximate maximum likelihood estimation (Kume and Wood, 2005, Section 2.3), but is completely different from the Approximate Maximum Likelihood Estimation technique developed here.

In this paper we propose a simulation-based approach to MLE, called Approximate Maximum Likelihood Estimation (AMLE), whose main advantage is the avoidance of the evaluation of the normalization constant. Broadly speaking, the method is based on a frequentist reinterpretation of Approximate Bayesian Computation (ABC) techniques, and its properties have been derived by Rubio and Johansen (2013) in a general setup; AMLE-based estimation has been developed by Bee et al. (2015) for the autologistic model.

The underlying principle is to generate candidate parameter values from bounded distributions (they would be the prior distributions in a Bayesian framework), computing certain summary statistics using the simulated data and then accepting only the parameter values such that the corresponding summary statistic is “close” to its observed counterpart. Under regularity conditions, the mode of the empirical distribution of the accepted parameter values is an approximation of the MLE. The standard version of AMLE samples the candidate parameter values from uniform distributions, but it would be possible to use different priors (Rubio and Johansen, 2013, p. 1637).

The distinctive feature of AMLE with respect to more traditional approaches to MLE with intractable constants is that, instead of computing an approximation of the likelihood and maximizing it, one can directly approximate the MLE by simulating observations from the distribution of interest. It is worth noting that AMLE is a quite effective technique but cannot be applied in an automated way, even when the availability of sufficient statistics makes the choice of the summary statistics obvious. In particular, details such as the choice of the metric, the ABC sample size and the optimization of the approximated likelihood have to be selected on a case-by-case basis.

AMLE is particularly appealing when two conditions are satisfied. First, its theoretical foundations are more solid when the sufficient statistics of the model under investigation are known, because in this case the convergence of the estimator to the MLE is guaranteed. Second, exact simulation of the model must be possible, and it is highly desirable to have a computationally efficient sampling algorithm. In other words, the first condition is crucially important to make sure that the estimator has the same asymptotic behavior of the MLE, whereas the second condition is relevant to set up the algorithm and limit the computational burden. The Bingham distribution meets both requirements: the sufficient statistics are readily computed and random number generation can be accomplished via an accept-reject method developed by Kent et al. (2013). Hence, the present setup is very well suited to the use of AMLE.

The contribution of this article is twofold. First, we work out the details of a new approach to the estimation of the Bingham distribution based on the AMLE method of Rubio and Johansen (2013). Second, we carry out a numerical study aimed at comparing AMLE and the benchmark technique that uses the HGM approximation of the normalizing constant.

The rest of the paper is organized as follows. Section 2 outlines the AMLE approach in a general framework; Section 3 specializes it to the Bingham estimation problem; Section 4 gives the results of extensive simulation experiments and suggests some strategies for choosing the parameters of the algorithm; Section 5 analyzes two real datasets and Section 6 concludes.

2. Approximate Maximum Likelihood Estimation

The AMLE approach exploits the potential of ABC techniques in a frequentist setup. In the following we briefly describe the algorithm, referring to Rubio and Johansen (2013) for details.

Given a sample $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) \in \mathbb{R}^{q \times n}$ from a distribution with density function $f(\boldsymbol{y}; \boldsymbol{\theta})$, let $L(\boldsymbol{\theta}; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ be the likelihood function, where $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^s$ is a vector of parameters. If we temporarily assume a Bayesian setup and let $\pi(\boldsymbol{\theta})$ be the prior distribution of $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ is the posterior, given by

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} f(\boldsymbol{y}|\boldsymbol{t})\pi(\boldsymbol{t})d\boldsymbol{t}}. \tag{3}$$

Now consider the following approximation of the likelihood function:

$$\hat{f}_\epsilon(\boldsymbol{y}|\boldsymbol{\theta}) = \int_{\mathbb{R}^{q \times n}} K_\epsilon(\boldsymbol{y}|\boldsymbol{z}) f(\boldsymbol{z}|\boldsymbol{\theta}) d\boldsymbol{z}, \tag{4}$$

where $K_\epsilon(\boldsymbol{y}|\boldsymbol{z})$ is a normalized Markov kernel and ϵ is a scale parameter. Plugging (4) into (3) we can compute an approximation of the posterior:

$$\hat{\pi}_\epsilon(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\hat{f}_\epsilon(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} \hat{f}_\epsilon(\boldsymbol{y}|\boldsymbol{t})\pi(\boldsymbol{t})d\boldsymbol{t}}.$$

If we restrict the analysis to a uniform prior on a suitable set $\boldsymbol{D} \subset \mathbb{R}^s$, maximizing the likelihood and maximizing the posterior density is the same, provided that the posterior is written in the parameterization of interest.

Let $\boldsymbol{\eta} : \mathbb{R}^{q \times n} \to \mathbb{R}^l$ be a summary statistic. The kernel $K_\epsilon^\rho(\boldsymbol{s}|\boldsymbol{t})$ is defined on the space of these summary statistics as follows:

$$K_\epsilon^\rho(\boldsymbol{\eta}(\boldsymbol{y})|\boldsymbol{\eta}(\boldsymbol{z})) \propto \begin{cases} 1 & \rho(\boldsymbol{\eta}(\boldsymbol{y}), \boldsymbol{\eta}(\boldsymbol{z})) < \epsilon, \\ 0 & \text{otherwise}, \end{cases} \tag{5}$$

where $\rho : \mathbb{R}^l \times \mathbb{R}^l \to \mathbb{R}^+$ is a metric. If $\boldsymbol{\eta}(\boldsymbol{y}) = \boldsymbol{y}$, one obtains the Pritchard et al. (1999) ABC algorithm. Using a summary statistic $\boldsymbol{\eta}(\boldsymbol{y})$ instead of the original sample $\boldsymbol{y}$ implies no loss of information exactly if $\boldsymbol{\eta}$ is a jointly sufficient statistic for the unknown parameters of the model: in this case, $L(\boldsymbol{\theta};\boldsymbol{y}_1,\dots,\boldsymbol{y}_n) = L(\boldsymbol{\theta};\boldsymbol{\eta}(\boldsymbol{y}_1,\dots,\boldsymbol{y}_n))$, that is, conditioning upon the sufficient statistics is the same as conditioning upon the sample. Thus in the AMLE setup it is highly recommended to use the sufficient statistics of the model, if available.

The preceding discussion motivates the following algorithm:

Algorithm 1. *(AMLE)*

1. *Obtain a sample $\boldsymbol{\theta}^*_\epsilon = (\boldsymbol{\theta}^*_{\epsilon,1},\dots,\boldsymbol{\theta}^*_{\epsilon,m})'$ from the approximate posterior $\hat{\pi}_\epsilon(\boldsymbol{\theta}|\boldsymbol{y})$; m is commonly called ABC sample size;*
2. *Use this sample to construct a nonparametric estimator $\hat{\phi}$ of the density $\hat{\pi}_\epsilon(\boldsymbol{\theta}|\boldsymbol{y})$;*
3. *Compute the maximum of $\hat{\phi}$, $\tilde{\boldsymbol{\theta}}_{m,\epsilon}$. This is an approximation of the MLE $\hat{\boldsymbol{\theta}}$.*

The most common implementation of Step 1 is the ABC rejection algorithm described by the following pseudo-code.

Algorithm 2. *(ABC rejection algorithm)*

1. *Simulate $\boldsymbol{\theta}^*$ from the prior distribution $\boldsymbol{\pi}(\cdot)$;*
2. *Generate $\boldsymbol{y} = (y_1,\dots,y_n)'$ from $f(\cdot|\boldsymbol{\theta}^*)$;*
3. *Use $\boldsymbol{y}$ to compute summary statistics $\boldsymbol{\eta}(\boldsymbol{y})$; accept $\boldsymbol{\theta}^*$ with probability $\propto K^\rho_\epsilon(\boldsymbol{\eta}(\boldsymbol{y})|\boldsymbol{\eta}(\boldsymbol{z}))$, otherwise return to Step 1.*

In the basic AMLE setup, at Step 1 the prior $\boldsymbol{\pi}$ is the q-product of uniform distributions with supports on (generally different) intervals $[\theta_{iL},\theta_{iU}]$, $i=1,\dots,s$. The crucial result proved by Rubio and Johansen (2013) is that, under a mild condition about $K^\rho_\epsilon(\boldsymbol{y}|\boldsymbol{z})$, $\hat{\pi}_\epsilon(\boldsymbol{\theta}|\boldsymbol{y})$ converges pointwise to $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ as $\epsilon \to 0$, for any $\boldsymbol{\theta} \in \boldsymbol{D}$. As a corollary it can be shown that, if $\boldsymbol{\eta}$ is a sufficient statistic for $\boldsymbol{\theta}$, the ABC approximation converges pointwise to the posterior distribution.

Finally, under the additional condition of equicontinuity of $\hat{\pi}_\epsilon(\cdot|\boldsymbol{y})$ on $\boldsymbol{D}$, and provided $\pi(\cdot|\boldsymbol{y})$ has a unique maximizer $\tilde{\boldsymbol{\theta}}$, it is possible to show that $\lim_{\epsilon\to 0}\hat{\pi}_\epsilon(\tilde{\boldsymbol{\theta}}|\boldsymbol{y}) = \pi(\tilde{\boldsymbol{\theta}}|\boldsymbol{y})$.

Now suppose that a simple random sample $\boldsymbol{\theta}^*_\epsilon = (\boldsymbol{\theta}^*_{\epsilon,1},\dots,\boldsymbol{\theta}^*_{\epsilon,m})'$ from the approximate posterior $\hat{\pi}_\epsilon(\cdot|\boldsymbol{y})$ with mode $\tilde{\boldsymbol{\theta}}_\epsilon$ is available. Let $\tilde{\boldsymbol{\theta}}_{m,\epsilon}$ be an estimator of $\tilde{\boldsymbol{\theta}}_\epsilon$ obtained from $\boldsymbol{\theta}^*_\epsilon$ such that $\tilde{\boldsymbol{\theta}}_{m,\epsilon} \to \tilde{\boldsymbol{\theta}}_\epsilon$ almost surely when $m \to \infty$. Then, for any $\gamma > 0$, there exists $\epsilon > 0$ such that $\lim_{m\to\infty}|\hat{\pi}_\epsilon(\tilde{\boldsymbol{\theta}}_{m,\epsilon}|\boldsymbol{y}) - \pi(\tilde{\boldsymbol{\theta}}|\boldsymbol{y})| \leq \gamma$ almost surely.

Although non-sufficient summary statistics can be used and weaker asymptotic results can be obtained in this setup (Rubio and Johansen, 2013, Proposition 2), in this brief summary of the theory we have emphasized the role of sufficiency. The reason is not only that convergence to the MLE in the terms presented above depends on sufficiency, but also that sufficient statistics are available for the Bingham distribution, and this is a strong argument in favor of the use of AMLE for approximate MLE of its parameters.

3. AMLE of the Bingham distribution

Under the identifiability constraint $\lambda_q = 0$, the standard (i.e., with diagonal $\boldsymbol{\Lambda}$) q-dimensional Bingham density is given by

$$f(\boldsymbol{x}; \boldsymbol{\Lambda}) = \frac{\exp\left\{-\sum_{i=1}^{q-1} \lambda_i x_i^2\right\}}{c(\boldsymbol{\Lambda})}, \tag{6}$$

so that the joint density of a random sample $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ from (6) is

$$f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \boldsymbol{\Lambda}) = \frac{\exp\left\{-n \sum_{i=1}^{q-1} \lambda_i \eta_i\right\}}{c(\boldsymbol{\Lambda})},$$

where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_q)$ and $\eta_i = (1/n) \sum_{j=1}^{n} x_{j,i}^2$. Hence, by the factorization theorem, the statistic $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{q-1})'$ is jointly sufficient for $\lambda_1, \ldots, \lambda_{q-1}$.

The Bingham distribution can be simulated by means of an accept-reject algorithm (Kent et al., 2013; see also Fallaize and Kypraios, 2016) that uses the Angular Central Gaussian distribution (ACG; Tyler, 1987) as an envelope. As pointed out by Kent et al. (2013), evaluating the acceptance probability is difficult because it depends on the normalizing constant; however, it has been verified empirically that the efficiency is never lower than 52% when $q = 3$ (Kent et al., 2013). For larger q, the efficiency deteriorates rather quickly; although the actual acceptance rate depends on the numerical values of the parameters, when $q = 7$ some simulations whose results are not reported here give an average acceptance probability close to the 10% found by Fallaize and Kypraios (2016). Hence, AMLE becomes computationally more demanding for large-dimensional problems; see Section 4.3 for further details.

According to algorithms 1 and 2, a pseudo-code of AMLE for a q-dimensional standard Bingham random vector $\boldsymbol{X}$ is as follows.

Algorithm 3. *(AMLE of the Bingham distribution)*

1. *Simulate* $\boldsymbol{\lambda}^*$ *from the prior distribution* $\pi(\boldsymbol{\lambda}) = \prod_{i=1}^{q-1} \pi(\lambda_i)$*, where* $\pi(\lambda_i)$ *is* $U(\lambda_{iL}, \lambda_{iU})$*;*
2. *Generate* $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)'$ *from* $f(\cdot|\boldsymbol{\lambda}^*)$*, where* f *is the Bingham density;*
3. *Use* $\boldsymbol{y}$ *to compute sufficient statistics* $\boldsymbol{\eta}^{sim}$*; accept* $\boldsymbol{\lambda}^*$ *with probability* $\propto K_\epsilon^\rho(\boldsymbol{\eta}^{obs}|\boldsymbol{\eta}^{sim})$*, otherwise return to Step 1. Here,* $\boldsymbol{\eta}^{obs} = (\eta_1, \ldots, \eta_{q-1})' = (\sum_{j=1}^{n} x_{j,1}^2, \ldots, \sum_{j=1}^{n} x_{j,q-1}^2)'/n$ *are the observed sufficient statistics.*

4. *Repeat steps 1-3 until m vectors of simulated parameter values $\boldsymbol{\lambda}^*_\epsilon = (\boldsymbol{\lambda}^*_{\epsilon,1}, \ldots, \boldsymbol{\lambda}^*_{\epsilon,m})'$ from the approximate posterior $\hat{\pi}_\epsilon(\boldsymbol{\lambda}|\boldsymbol{y})$ are accepted; $\boldsymbol{\lambda}^*_\epsilon$ is the ABC sample.*
5. *Use $\boldsymbol{\lambda}^*_\epsilon$ to find a nonparametric estimator $\hat{\phi}$ of the density $\hat{\pi}_\epsilon(\boldsymbol{\lambda}|\boldsymbol{y})$;*
6. *Compute the maximum of $\hat{\phi}$, $\tilde{\boldsymbol{\lambda}}_{m,\epsilon}$. This is an approximation of the MLE $\hat{\boldsymbol{\lambda}}$.*

Two additional comments are in order regarding Algorithm 3. As mentioned in Section 1, to ensure identifiability we use the constraint $\lambda_1 \geq \cdots \geq \lambda_q = 0$, and the case of general (non-diagonal) $\boldsymbol{A}$ is dealt with by considering a rotation implied by the empirical sum of squares and products matrix $\sum_{j=1}^n \boldsymbol{x}_j \boldsymbol{x}_j'$. The spectral decomposition of $\boldsymbol{A}$ is $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}'$ and its MLE is $\boldsymbol{V}\hat{\boldsymbol{\Lambda}}\boldsymbol{V}'$ (Bingham, 1974, Theorem 6.1(c)), where $\hat{\boldsymbol{\Lambda}} = \text{diag}(\hat{\lambda}_i)$ is the diagonal matrix of the MLEs of λ_i $(i = 1, \ldots, q-1)$ and $\boldsymbol{V}$ is the matrix of the eigenvectors of $\sum_{j=1}^n \boldsymbol{x}_j \boldsymbol{x}_j'$. When working with data, we assume that the observed sufficient statistics determine the ranking of the parameters, i.e. $\eta_i < \eta_j \Rightarrow \lambda_i > \lambda_j$. At Step 1, to take care of this issue, we first check that $\lambda_i^* > \cdots > \lambda_{q-1}^*$. If this condition is satisfied, the algorithm proceeds; otherwise, Step 1 is repeated. Clearly, the number of candidate parameter values rejected for this reason increases when (i) two or more sufficient statistics are close to each other, so that the supports of the uniform distributions are characterized by more overlap, and (ii) the dimension q gets large. To overcome this difficulty, it would be possible to generate candidate parameter values from conditional uniform distributions: $\lambda_i^* \sim U[\lambda_{iL}, \lambda_{i-1}^*]$, $i = 2, \ldots, q-1$. However, we have verified via simulation that the computational gain associated with this conditional sampling approach is negligible compared to the total computational burden of the algorithm. Thus in the following we adhere to the procedure described in Step 1 of Algorithm 3.

Second, the mode of the joint posterior is typically approximated by using the maximum of the multivariate kernel density fitted to the data using the `kde` command of the `ks R` package (Duong, 2014). However, this issue requires special attention when the dimension of the problem gets larger, as kernel density approximations quickly become less reliable. In particular, the `kde` command does not work for dimensions larger than 6 (and even if it worked, a very large ABC sample size would be necessary for good results). For these reasons, when $q > 3$, we investigate some alternative techniques. Specifically, we approximate the mode of the joint distribution via:

1. the maximum of the multivariate kernel density ("K"; only when $q \leq 6$);
2. the sample mean ("M");
3. the maximum of the product of the univariate kernel densities estimated using the marginal data ("P");
4. the mean shift algorithm ("MS");

In case 3 the algorithm sequentially uses the marginal data, so that only univariate estimation is required. The remaining methods are truly multivariate. For MS we use the `bmsClustering` command of the `MeanShift R` package

(Ciollaro and Wang, 2016), which is based on the so-called blurring mean shift algorithm. We have also used the standard version of the algorithm, and the results are identical to the third decimal place. As we know that the distribution is unimodal, we specify that there is only one cluster. The cases 2 to 4 have the advantage that there is no need to construct a computationally expensive non-parametric approximation of the multidimensional density at all and allow one to implement AMLE in dimensions where the `kde` limitations preclude its use, provided that the ABC samples are large enough. It is worth noting that a further approach, not implemented here, may be convenient in large-dimensional setups: grouping the parameters into sets of tractable small dimensions (e.g., by using scatterplots to visually identify the most correlated parameters), fitting a multivariate kernel density to each block of parameters, and maximizing this product.

Finally, Algorithm 3 can in principle be simplified by exploiting the MCMC approach, based on the Metropolis algorithm, developed by Fallaize and Kypraios (2016) to obtain an exact sample from $\hat{\pi}_\epsilon(\boldsymbol{\lambda}|\boldsymbol{y})$. In particular, it would be possible to replace steps 2 and 3 of Algorithm 3 by steps 2 and 3 of the algorithm presented on p. 352 of Fallaize and Kypraios (2016), using a uniform prior. This may result in a faster algorithm, but it would not produce truly independent samples and, as Rubio and Johansen (2013, Sect. 3.1) point out, dependence between samples produced via MCMC techniques can make density estimation more complicated. Although dependence can be reduced by thinning the sample, the implementation of an MCMC-based approach is likely to be non-trivial, and is certainly non-automatic, as it requires that all classical MCMC inputs be set (proposal distribution, burn-in period, stopping criterion, etc.).

Accordingly, it is not clear a priori whether it is preferable to rely on an MCMC-based approach or on the ABC-based method used in the current paper. While this issue is certainly interesting, not only for the estimation of the Bingham distribution, but also in more general setups, we do not pursue it here.

3.1. The standard MLE approach

The log-likelihood function of the standard Bingham distribution is given by

$$l(\boldsymbol{\Lambda}; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = -n\left(\sum_{i=1}^{q-1} \lambda_i \eta_i + n \log(c(\boldsymbol{\Lambda}))\right). \tag{7}$$

The benchmark method for computing MLEs is based on the maximization of the approximate likelihood function obtained by plugging an estimate $\hat{c}(\boldsymbol{\Lambda})$ of $c(\boldsymbol{\Lambda})$ into (7). The first-order conditions are given by

$$-n\left(\eta_i + \frac{\hat{c}_i'(\boldsymbol{\Lambda})}{\hat{c}(\boldsymbol{\Lambda})}\right) = 0, \qquad i = 1, \ldots, q-1,$$

where $\hat{c}_i'(\boldsymbol{\Lambda})$ is the estimate of the partial derivative of $c(\boldsymbol{\Lambda})$ with respect to the i-th parameter. Sei and Kume (2015) propose to estimate the constant by means of the holonomic gradient method, which is implemented in the `R` package `hgm`

(Takayama et al., 2015). In the following, we will call HGMs the MLEs obtained with this approach.

4. Simulation experiments

The choice of the parameters of the AMLE algorithm is a delicate issue that deserves a detailed investigation, because inappropriate values can have a dramatic impact on the results.

We first apply the algorithm to two synthetic datasets from the Bingham distribution on the unit sphere, so that $q = 3$. As one of the aims consists of comparing AMLE to existing estimation methods, we consider the two samples analyzed by Mardia and Zemroch (1977) and Fallaize and Kypraios (2016). They are called Dataset 1 and Dataset 2, with sufficient statistics respectively equal to $\boldsymbol{\eta}^{obs} = (0.30, 0.32)'$ and $\boldsymbol{\eta}^{obs} = (0.02, 0.40)'$. The sample size is $n = 100$ in both cases.

The very first step consists in determining the ranges D_i of the uniform priors, i.e. the intervals such that $\lambda_i \in D_i$ $(i = 1, \dots, q-1)$. The relationship between the λ_is and the eigenvalues of the sample covariance matrix is quite complicated (Love, 2007), so that no simple moment-based procedure can be used to find initial values of the parameters. However, the concentration of the i-th marginal distribution of $\boldsymbol{X}$ is a monotone function of λ_i: as λ_i gets larger, the distribution is more peaked along the i-th direction. This feature may be used as a guideline to come up with an interval. In absence of any optimal procedure, a small pilot simulation is usually enough to obtain reasonably precise information about the ranges D_i. In any case, it is not worth spending much time on the fine tuning of the D_is, because the supports of the uniform distributions have a rather limited effect on the computational burden (see Section 4.1 for details).

Besides ϵ, the other crucial parameter for the properties of the estimators is the metric ρ. When the absolute values of the parameters differ considerably from each other, the normalized version of the Euclidean distance $\bar{d}(\boldsymbol{x}, \boldsymbol{y}) \overset{\text{def}}{=} \sqrt{\sum_{i=1}^{p}((x_i - y_i)/x_i)^2}$ is usually preferable to the Euclidean distance $d(\boldsymbol{x}, \boldsymbol{y}) \overset{\text{def}}{=} \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$; see Sousa et al. (2009) and Beaumont (2010) for the use of $\bar{d}$ in the ABC setup (note that $\bar{d}$ is not a distance, but this is not crucial here).

Before turning to the analysis of the impact of the input parameters of the algorithm, it is worth considering the role of the sample size n. Consider Dataset 2, with $\lambda_1 = 25.31$, $\lambda_2 = 0.762$ and $\lambda_3 = 0$. For $n \in \{500, 1000, 10\,000, 30\,000\}$ we simulate n observations from the Bingham distribution with these parameters and compute $\boldsymbol{\eta}^{obs}$. Then we sample 500 pairs of parameter values $\lambda_{i,1} \sim U(15, 40)$ and $\lambda_{i,2} \sim U(0, 2)$, for each pair of parameter values we simulate n observations from the Bingham distribution, compute the corresponding simulated sufficient statistic $\boldsymbol{\eta}_i^{sim} = (\eta_{i,1}^{sim}, \eta_{i,2}^{sim})$ and finally the numerical values of $\bar{d}_i(\boldsymbol{\eta}^{obs}, \boldsymbol{\eta}_i^{sim})$ $(i = 1, \dots, 500)$.

Figure A.1 (all the numbers prefixed by "A." refer to figures reported in the supplementary material) shows the scatterplot of $\lambda_{i,1}$ and $\bar{d}_i(\boldsymbol{\eta}^{obs}, \boldsymbol{\eta}_i^{sim})$ for $n \in \{500,\ 1000,\ 10\,000,\ 30\,000\}$. The horizontal line, arbitrarily drawn at

$\bar{d} = 0.04$, helps to identify the values of λ_1 corresponding to small values of $\bar{d}$, i.e. the values of λ_1 that would be included in an hypothetical ABC sample determined by the condition $\bar{d} < 0.04$ at step 4 of Algorithm 3. Whereas the shape of the cloud is approximately the same in the four panels, as n gets larger the borders are smoother and the distribution is more peaked near the true value so that the smallest values of $\bar{d}$ correspond to values of λ_1 closer to the true value. This is particularly evident for $n = 30\,000$.

4.1. Choosing the parameters of the algorithm

In practice, it is often difficult to have an idea of the values of the normalized Euclidean distance between the observed and simulated values of the summary statistics, unless one has some information about their distributions. Thus, it is more common to choose, instead of ϵ, the fraction of accepted values f (Sousa et al., 2009). In this case one simulates a large number of candidate parameters from the uniform distributions, uses them to sample the distribution and compute the summary statistics, and then includes in the ABC sample only the parameters corresponding to some predefined fraction f with the smallest values of the distance between $\boldsymbol{\eta}^{obs}$ and $\boldsymbol{\eta}^{sim}$. In the following we adhere to this way of proceeding and study how the properties of the estimators depend on f. Typical values of f used in the ABC literature range from 10^{-2} to 10^{-5} (see, e.g., Sousa et al., 2009, and the references therein).

The ranges D_i $(i = 1, 2)$ in datasets 1 and 2 are determined by means of the following simulation, whose details are explained focusing on Dataset 1. The value of $\boldsymbol{\eta}^{obs}$ suggests marginal distributions with rather high dispersion, i.e. small values of λ_1 and λ_2. Simulating $n_p = 10\,000$ candidate values of the parameters λ_1 and λ_2 respectively from the $U(0, 3)$ and $U(0, 2)$ distributions and using $f = 10\%$, we obtain empirical ranges $[\min \lambda_i, \max \lambda_i]$ equal to $[0.012, 2.218]$ for λ_1 and $[0.001, 1.604]$ for λ_2. According to these outcomes, all the analyses can be safely carried out with $D_1 = (0, 3)$ and $D_2 = (0, 2)$. A similar analysis for Dataset 2 gives ranges $[13, 45]$ for λ_1 and $[0, 2]$ for λ_2.

We now analyze the effect of f on the estimators. Note that $f = m/n_p$, where m is the ABC sample size (always equal to 1000 in this experiment) and n_p is the number of candidate parameter values simulated from the uniform distributions. Various fractions f are obtained keeping $m = 1000$ and using different values of n_p. Specifically, we simulate samples ranging from $n_p = 10^5$ to $n_p = 25 \cdot 10^7$ from the uniforms, use them for sampling the Bingham distribution and compute the sufficient statistics. From each sample, we determine the ABC sample by taking the $m = 1000$ observations with the smallest normalized distance between $\boldsymbol{\eta}^{obs}$ and $\boldsymbol{\eta}^{sim}$, and computing AMLE by taking the mode of the kernel density estimated on those observations. The values of f are between 10^{-2} and $4 \cdot 10^{-6}$, so that they cover a range larger than the one typically used in the ABC literature (Sousa et al., 2009).

Panels (a) and (c) of Figure A.2 show the AMLEs obtained, whereas (b) and (d) display the corresponding standard errors, given by the empirical standard deviations of the simulated distributions of the estimators. The performance of AMLE clearly deteriorates only for the last two values of f, respectively equal to $8 \cdot 10^{-3}$ and 10^{-2}. If we omit them, the graphs seem to be characterized mostly by sampling variability: if, in each of the graphs, we omit the last two values and fit a simple linear regression, we never obtain a slope significantly different from zero, and the correlogram does not suggest the presence of any autocorrelation. The same analysis for Dataset 2 gives similar results.

To disentangle the effects of m and n_p, we carry out two further numerical investigations. In the first experiment we use $m \in \{30,\ 100,\ 500,\ 1000,\ 2000,\ 3000,\ 4000\}$ with $n_p = 10^6$; in the second one, we use the same values of m, but in each case we choose n_p so as to keep $f = 0.5\%$. For Dataset 1, the simulated distributions of $\hat{\lambda}_1$ and $\hat{\lambda}_2$ in the first experiment are displayed in panels (a) and (b) of Figure A.3, whereas panels (c) and (d) show the boxplots of the parameters in the second experiment. The same graphs for Dataset 2 are in Figure A.4. In both cases the boxplots suggest that the performance of the estimators is worse for $m \leq 500$. On the other hand, the distributions are similar for $m > 500$. Finally, Figure A.5 shows the AMLE estimators of $\hat{\lambda}_1$ and $\hat{\lambda}_2$ in Dataset 1 for an ABC sample size $m \in \{1, \ldots, 2000\}$ with $n_p = 2 \cdot 10^5$. Figure A.6 displays the same results for Dataset 2; the AMLE estimators are computed using the sample mean. Both graphs suggest that the estimators become approximately stable for m between 500 and 1000.

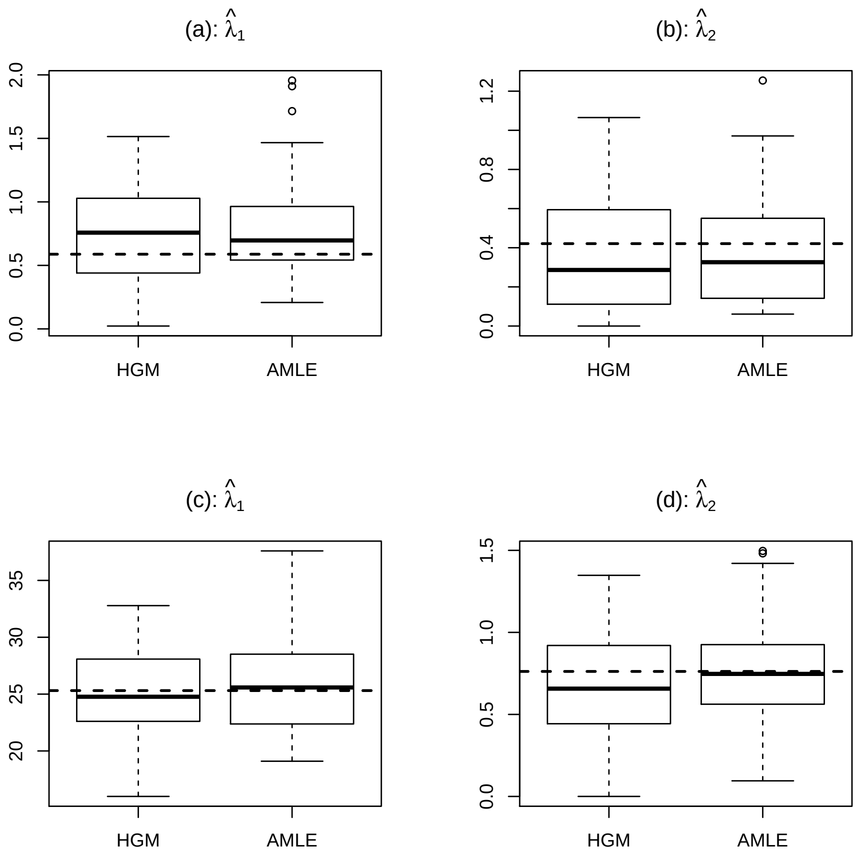
According to the outcomes just presented, we carry out all the computations for both datasets with $m = 1000$ and $n_p = 2 \cdot 10^5$. The computational burden associated to a fraction $f = 5 \cdot 10^{-3}$, obtained with $m = 1000$ and $n_p = 2 \cdot 10^5$, is relatively small (approximately 11 minutes for Dataset 1 on a CORE i7 processor with the R programming language and 8Gb of RAM memory). The larger range of the first uniform distribution increases the computing time in Dataset 2 to approximately 14 minutes. The modest difference between the two computational costs suggests that the ranges of the uniform distributions are not critical for the total time taken by the procedure.

4.2. Three-dimensional experiments

Figure 1 shows the empirical distributions of the HGM and AMLE parameter estimates obtained in 50 replications for Dataset 1 (panels (a) and (b)) and 2 (panels ((c) and (d)). Panels (a) and (c) are the boxplots of HGMs, panels (b) and (d) refer to AMLEs. Each replication of the experiment consists of simulating 100 observations from the Bingham distribution with parameters $\boldsymbol{\lambda} = (0.588, 0.421)'$ for Dataset 1 and $\boldsymbol{\lambda} = (25.31, 0.762)'$ for Dataset 2 and computing the HGM and AMLE estimators.

There are little differences between the two estimators. The HGM boxplots show somewhat more regular distributions, but it is worth noting that, in Dataset 1, out of 50 replications, HGM performed via constrained optimization over the rectangle $(0, 2) \times (0, 2)$ produced 2 estimates of λ_1 and 9 estimates of λ_2 equal to 0. Table 1 shows, for both methods, the point estimates, the standard error, the coefficient of variation of the RMSE and the relative performance. The coefficient of variation of the RMSE, given by $CV(RMSE)_{\hat{\lambda}_i} = RMSE(\hat{\lambda}_i)/\hat{\lambda}_i$, $i = 1, 2$, was preferred to the RMSE because of the large value

Figure 1: Distributions of $\hat{\lambda}_1$ and $\hat{\lambda}_2$ for Dataset 1 (panels (a) and (b)) and Dataset 2 (panels (c) and (d)) over 50 replications of HGM and AMLE.



	Dataset #	parameter	p. e.	s. e.	CV(RMSE)	rel. perf.
AMLE	1	$\hat{\lambda}_1$	0.797	0.396	0.762	1.135
		$\hat{\lambda}_2$	0.382	0.280	0.671	0.930
	2	$\hat{\lambda}_1$	25.798	4.415	0.175	1.176
		$\hat{\lambda}_2$	0.739	0.328	0.432	1.060
HGM	1	$\hat{\lambda}_1$	0.730	0.369	0.672	
		$\hat{\lambda}_2$	0.357	0.297	0.722	
	2	$\hat{\lambda}_1$	24.939	3.760	0.149	
		$\hat{\lambda}_2$	0.683	0.300	0.407	

Table 1: Point estimates (p. e.), standard errors (s. e.), CV(RMSE)s and relative performances (rel. perf.) of the estimators of the parameters of the Bingham distribution in Dataset 1 and 2. All the measures are computed using 50 replications. The AMLE estimation procedure is implemented with $m = 1000$ and $n_p = 2 \cdot 10^5$. The true values of the parameters are $\boldsymbol{\lambda} = (0.588, 0.421)'$ in Dataset 1 and $\boldsymbol{\lambda} = (25.31, 0.762)'$ in Dataset 2.

of the first parameter in Dataset 2; relative performance is defined as the ratio $CV(RMSE)^{AMLE}/CV(RMSE)^{HGM}$. HGMs show a slightly better performance in Dataset 2, whereas in Dataset 1 the CVRMSEs are approximately the same. In the latter case AMLE is preferred because AMLEs are strictly positive with probability 1.

Focusing on the AMLE approach, Figure A.7 shows the simulated distribution of the $m = 1000$ accepted values of the parameters for Dataset 1 (panels (a) and (b)) and 2 (panels (c) and (d)). Figure A.8 shows an ABC sample of size 1000 from the joint distribution of the parameters for Dataset 1 (panel (a)) and Dataset 2 (panel (b)). In both cases, the results are very similar to those obtained by Fallaize and Kypraios (2016).

4.3. Large-dimensional experiments

In large-dimensional frameworks, the performance of both estimators is expected to deteriorate. As for AMLE, a larger m is likely to be necessary, because multivariate kernel density estimation suffers from the curse of dimensionality and, in practice, more observations are required when the dimension of the problem gets larger. For different reasons, standard numerical optimization techniques become quickly less reliable as the number of parameters increases. Thus, in this section we investigate the performance of the estimators for the Bingham distribution with $q > 3$, approximating the mode of the joint distribution implementing the methods introduced in Section 3.

4.3.1. A 5-dimensional example

Consider first a sample of size $n = 100$ from the standard Bingham distribution with $q = 5$. We borrow the setup used by Sei and Kume (2015, p. 329), simulating 100 observations from the Bingham distribution with parameters $\boldsymbol{\lambda} = (7.188333, 3.120184, 1.543555, 0.628081, 0)'$. Some pilot simulations

	parameter	p. e.	s. e.	CV(RMSE)	rel. perf.
AMLE^M	$\hat{\lambda}_1$	7.090	0.230	0.035	0.224
	$\hat{\lambda}_2$	3.032	0.321	0.107	0.488
	$\hat{\lambda}_3$	1.531	0.115	0.075	0.255
	$\hat{\lambda}_4$	0.519	0.129	0.270	0.385
HGM	$\hat{\lambda}_1$	6.997	1.097	0.155	
	$\hat{\lambda}_2$	3.218	0.676	0.219	
	$\hat{\lambda}_3$	1.610	0.446	0.292	
	$\hat{\lambda}_4$	0.512	0.440	0.701	

Table 2: Point estimates (p. e.), standard errors (s. e.), CV(RMSE)s and relative performances of the HGM and AMLE^M estimators of the parameters of the Bingham distribution in the 5-dimensional example. AMLE^M estimators are computed with $m = 1000$ and $n_p = 6 \cdot 10^5$ and the mode of the posterior is approximated by the sample mean. The true values of the parameters are $\boldsymbol{\lambda} = (7.188333, 3.120184, 1.543555, 0.628081, 0)'$.

similar to those carried out in Section 4.1 suggest that, given the larger number of parameters, $n_p = 2 \cdot 10^5$ may be too small, and the variance becomes approximately stable only for $n_p \gtrapprox 5 \cdot 10^5$; accordingly, we use $n_p = 6 \cdot 10^5$. Even though the simulation becomes heavier when the dimension increases, because, as pointed out in Section 3, the acceptance rate of the algorithm sharply decreases, this value of n_p still guarantees a reasonable computational burden (approximately 35 minutes on a CORE i7 processor with the R programming language and 8Gb of RAM memory).

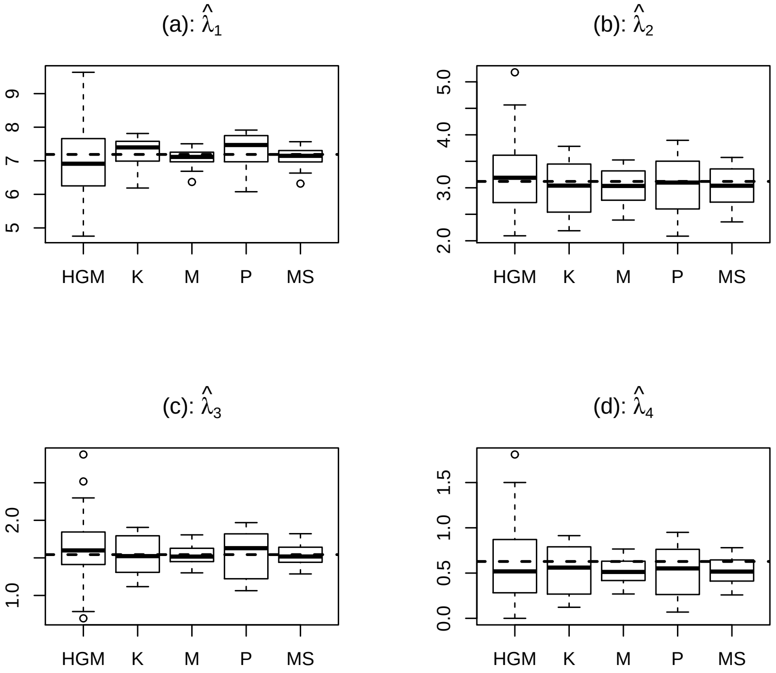
The results obtained for $\hat{\lambda}_1, \ldots, \hat{\lambda}_4$ (λ_5 is equal to zero in order to ensure identifiability) via HGM and the best AMLE approach (i.e., the one using sample means, AMLE^M from now on) are reported in Table 2, whereas Figure 2 shows the boxplots and Figure A.9 displays the bias and the CV(RMSE).

Both the Table and the figures suggest that, in terms of CV(RMSE), AMLE is significantly more efficient than HGM in this case. In addition, the latter method has the same problem noted in the simulation experiment for Dataset 1: in 3 out of 50 cases, the HGM estimator of λ_4 is equal to 0. There is a non-negligible difference among the various versions of AMLE: overall, "M" and "MS", whose bias and CV(RMSE) are almost indistinguishable, give the best results; (see Bee and Trapin, 2016, for a similar result). AMLE^M has a CV(RMSE) between 2 and 5 times smaller than the CV(RMSE) of HGM (see Table 2). There is little difference between HGM and AMLE in terms of bias (Figure A.9; although the bias of HGM is the largest one for all parameters), whereas AMLE outperforms HGM more markedly in terms of CV(RMSE). This implies that the variance of AMLE is smaller, as can be noted from Figure 2.

4.3.2. A 10-dimensional example

To conclude the simulation experiments, we tackle a challenging 10-dimensional example. Sampling 100 observations from the Bingham distribution with

Figure 2: Distributions of the HGM and AMLE estimators of λ_1, λ_2, λ_3 and λ_4 in the 5-dimensional example with 50 replications. The dashed horizontal lines denote the true values of the parameters. The AMLE is obtained as the maximum of the multivariate kernel density ("K"), the sample mean ("M"), the maximum of the product of the univariate kernel densities ("P") and via the mean shift algorithm ("MS"). The number of replications is equal to 50.



(a): $\hat{\lambda}_1$

(b): $\hat{\lambda}_2$

(c): $\hat{\lambda}_3$

(d): $\hat{\lambda}_4$

parameter vector

$$\boldsymbol{\lambda} = (25.3,\ 10,\ 6,\ 5.5,\ 3.7,\ 2.5,\ 2,\ 1.35,\ 0.6)',$$

we obtain the joint observed sufficient statistics

$$\boldsymbol{\eta}^{obs} = (0.01875, 0.0431, 0.0667, 0.0831, 0.0884, 0.1073, 0.1204, 0.1358, 0.1538, 0.1812)'.$$

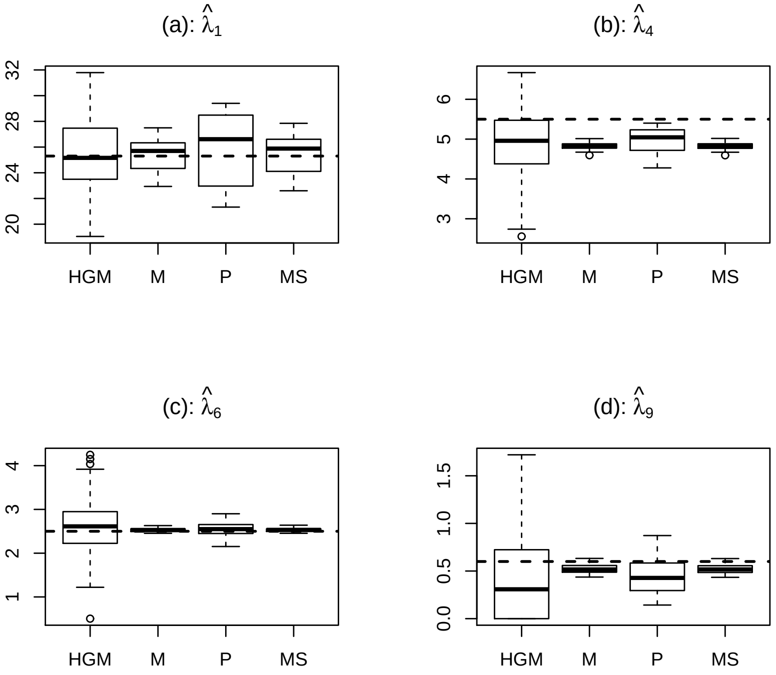
We perform AMLE with $n_p = 10^5$: this is a small number in this setup, but, given the high rejection rate, it corresponds to an approximately 11-hour long simulation experiment, which is still viable and in line with the spirit of setting up a procedure characterized by an acceptable balance between statistical precision and computational cost. Moreover, the HGM approach has a non-negligible computational cost as well, as it takes approximately 45 minutes in this setup. To find the mode of the approximated posterior we use the same algorithms of the 5-dimensional case, except multivariate kernel density estimation which is not implemented in `kde` for dimensions larger than 6.

Figure 3 shows the simulated distribution of the estimators of λ_1, λ_4, λ_6 and λ_9 (the boxplots of the remaining parameters are qualitatively similar), Table 3 compares HGM to AMLE^M, which is again the best version of AMLE, and Figure A.10 shows the bias (panel (a)) and the CV(RMSE) (panel (b)) of all the estimators.

Overall, the performance of the estimators deteriorates considerably in this case, but this is unsurprising if we consider that estimating nine parameters with 100 observations is a difficult task. Turning to the two methods of estimation, HGM is again the worst performer in terms of CV(RMSE), and this is mainly due to a larger variability. Moreover, panel (d) of Figure 3 shows that the HGM estimator of λ_9 has a very skewed distribution. The reason is indeed the same as noted above: many values of $\hat{\lambda}_9$ are equal to zero. Here this drawback is more widespread, as 15 values of $\hat{\lambda}_9$, 4 values of $\hat{\lambda}_6$ and 1 value of $\hat{\lambda}_4$ are equal to zero. AMLEs are not very precise as well; however, despite the relatively small value of n_p, they are considerably more stable than the HGM estimators, so that the CV(RMSE) is always smaller (see Table 3). Analogously to the 5-dimensional case, "M" and "MS" are nearly indistinguishable and give better results with respect to the remaining versions of AMLE. Table 3 suggests that the advantage of AMLE^M with respect to HGM is substantial, as the CV(RMSE) of AMLE^M is from approximately 2 to more than 10 times smaller than the CV(RMSE) of HGM. Finally, Figure A.10 shows that the bias is approximately the same across all the estimators, but, for all versions, AMLE outperforms HGM in terms of CV(RMSE).

As for AMLE, figures A.11 and A.12 show the marginal distributions of the accepted values of two randomly chosen marginals. The histograms confirm that the distributions of λ_1 and λ_5 still have the desirable properties obtained in the preceding experiments. With respect to the three- and five-dimensional cases analyzed above, the main difference is an increased variability, which is explained by the larger dimension of the problem and by the smaller n_p used for the 10-dimensional case. It should be noted that, for the purposes of this

Figure 3: Distributions of the HGM and AMLE estimators $\hat{\lambda}_1$, $\hat{\lambda}_4$, $\hat{\lambda}_6$ and $\hat{\lambda}_9$ in the 10-dimensional example with 50 replications. The dashed horizontal lines denote the true values of the parameters. The AMLE is obtained as the maximum of the multivariate kernel density ("K"), the sample mean ("M"), the maximum of the product of the univariate kernel densities ("P") and via the mean shift algorithm ("MS"). The number of replications is equal to 50.



	Estimator	Point estimate	Standard error	CV(RMSE)	Rel. Perf.
AMLE	$\hat{\lambda}_1$	25.33	1.14	0.05	0.40
	$\hat{\lambda}_2$	9.96	0.84	0.08	0.38
	$\hat{\lambda}_3$	6.11	0.11	0.02	0.14
	$\hat{\lambda}_4$	4.82	0.09	0.12	0.63
	$\hat{\lambda}_5$	3.59	0.12	0.04	0.19
	$\hat{\lambda}_6$	2.53	0.04	0.02	0.07
	$\hat{\lambda}_7$	1.96	0.04	0.03	0.08
	$\hat{\lambda}_8$	1.31	0.04	0.04	0.08
	$\hat{\lambda}_9$	0.52	0.04	0.15	0.17
HGM	$\hat{\lambda}_1$	25.49	2.86	0.11	
	$\hat{\lambda}_2$	10.62	2.12	0.22	
	$\hat{\lambda}_3$	6.35	1.01	0.18	
	$\hat{\lambda}_4$	4.94	0.93	0.20	
	$\hat{\lambda}_5$	3.55	0.84	0.23	
	$\hat{\lambda}_6$	2.55	0.76	0.30	
	$\hat{\lambda}_7$	1.92	0.73	0.37	
	$\hat{\lambda}_8$	1.19	0.66	0.50	
	$\hat{\lambda}_9$	0.47	0.50	0.86	

Table 3: Point estimates, standard errors, CV(RMSE)s and relative performances (rel. perf.) of the HGM and AMLE estimators of the parameters of the Bingham distribution in the 10-dimensional example. AMLE estimators are computed with $m = 1000$ and $n_p = 6 \cdot 10^5$ and the mode of the posterior is approximated by the sample mean. The true values of the parameters are $\boldsymbol{\lambda} = (25.3, 10, 6, 5.5, 3.7, 2.5, 2, 1.35, 0.6)'$.

simulation exercise, we have been forced to employ a rather large f, but, in a single estimation step, one may accept a higher computing time and thus choose a larger n_p, which would result in more accurate results.

5. Real-data applications

5.1. Calcite grains data

This example has first been used by Bingham (1974) to illustrate the MLE approach. Fallaize and Kypraios (2016) provide a Bayesian analysis of the same dataset. The data consist of $n = 150$ measurements on the c-axis of calcite grains from the Taconic Mountains of New York state.

Given the sum of squares and products matrix $\boldsymbol{SS} \overset{\text{def}}{=} \sum_{i=1}^{150} \boldsymbol{x}_i \boldsymbol{x}_i'$, the sufficient statistics $\eta_i = \sum_{j=1}^{n} x_{j,i}^2/n$ ($i = 1, \ldots, q-1$) are given by $\lambda_{(i)}^{SS}/n$, where $\lambda_{(i)}^{SS}$ are the $q-1$ smallest eigenvalues of $\boldsymbol{SS}$ in ascending order, i.e. $\lambda_{(1)}^{SS} \leq \cdots \leq \lambda_{(q-1)}^{SS}$.

		λ_1^A	λ_2^A	λ_1^B	λ_2^B	λ_1^S	λ_2^S
AMLE	Point estimate	4.812	3.732	5.069	2.916	1.846	1.067
	Standard error	0.267	0.205	0.281	0.160	0.132	0.161
HGM	Point estimate	5.059	3.804	5.094	2.941	1.809	1.025
	Standard error	0.344	0.265	0.333	0.238	0.173	0.234

Table 4: Point estimates and standard errors for the earthquake data. AMLE uses $m = 1000$ and $n_p = 2 \cdot 10^5$. Standard errors are computed with 100 non-parametric bootstrap replications.

From Bingham (1974) we know that

$$\boldsymbol{SS} = \begin{pmatrix} 76.5575 & 18.2147 & 12.2406 \\ 18.2147 & 46.7740 & 6.8589 \\ 12.2406 & 6.8589 & 26.667 \end{pmatrix}.$$

By means of the usual pilot simulation we choose $D_1 = [1, 6]$ and $D_2 = [0.5, 4]$ in the implementation of the algorithm. The AMLEs obtained with $m = 1000$ and $n_p = 10^5$ are $\hat{\lambda}_1 = 3.567$ and $\hat{\lambda}_2 = 1.963$. The HGMs are identical to those found by Bingham (1974), i.e. $\hat{\lambda}_1 = 3.518$ and $\hat{\lambda}_2 = 1.956$. Both results are very close to the estimates reported by Fallaize and Kypraios (2016).

5.2. Earthquake data

The earthquake example is the second real-data application proposed by Fallaize and Kypraios (2016), and is based on data first analyzed by Arnold and Jupp (2013). These two references also give a full description of the data and of their interpretation, which is therefore omitted here. Three clusters of three-dimensional observations, called respectively A, B and S (i.e., $q = 3$), are available, and the corresponding sample sizes and sufficient statistics are $n_A = n_B = 50$, $n_S = 32$, $\boldsymbol{\eta}_A = (0.1152360, 0.1571938)'$, $\boldsymbol{\eta}_B = (0.1127693, 0.1987671)'$ and $\boldsymbol{\eta}_S = (0.2288201, 0.3035098)'$. For each dataset, we fit a Bingham distribution, and the results are displayed in Table 4. The AMLE parameters are $m = 1000$ and $n_p = 2 \cdot 10^5$. Throughout this section, standard errors are computed via non-parametric bootstrap with 100 replications.

To evaluate whether there is no difference between the clusters A and B, we compute an approximate 95% confidence region for $\boldsymbol{\lambda}^A - \boldsymbol{\lambda}^B$. Two methods are used. The first is parametric, based on the assumption of bivariate normality; the second is non-parametric and uses bivariate kernel density estimation. We only show the AMLE outcomes here, as with HGM similar results are obtained.

A graphical representation of the results is given in figures 4 and 5. The bivariate confidence regions computed with the two methods are quite similar, and the results are in line with those obtained by Fallaize and Kypraios (2016). Even though the number of bootstrap replications is rather small for computing a 95% confidence level, and therefore the curves are not very smooth, the outcome is clear. The origin is contained in the confidence interval of panel (a), suggesting that $\boldsymbol{\lambda}^A$ is not significantly different from $\boldsymbol{\lambda}^B$. On the other hand, the origin is well outside the confidence interval for $\boldsymbol{\lambda}^S - \boldsymbol{\lambda}^B$ (panel (b)), so that we reject the hypothesis $\boldsymbol{\lambda}^S = \boldsymbol{\lambda}^B$ at the 5% level.

6. Conclusion

This paper studies approximate maximum likelihood estimation of the Bingham distribution. We develop a method exploiting Approximate Bayesian Computation techniques to approximate the MLEs. This approach, based on Rubio and Johansen (2013), is particularly well-suited for the Bingham distribution. First, it bypasses the problem of evaluating the normalizing constant. Second, the sufficient statistics are readily computed. Third, an efficient random number generator is available. While the importance of the first feature is immediately apparent, the second can be shown to play a key role for the theoretical properties of the estimators, and the third is needed for an efficient implementation of the algorithm.

Besides assessing the merits of AMLE, we carry out a comparison with the likelihood approach based on the approximation of the normalizing constant and the numerical maximization of the approximated likelihood (Bingham, 1974; Kume and Wood, 2005; Sei and Kume, 2015). Overall, the two approaches have a similar performance in the three-dimensional case; as the dimension increases, AMLE outperforms HGM. This is not surprising in light of the fact that deterministic numerical methods suffer more than simulation-based methods from the "curse of dimensionality" (see, for example, Glasserman, 2003, pp. 2-3). In general, AMLE has an heavier computational burden with respect to HGM, but in large dimension HGM computing times are non-negligible as well.

There is a striking resemblance between our outcomes and the output of the Bayesian analysis, not based on ABC but rather on Markov Chain Monte Carlo methods, carried out by Fallaize and Kypraios (2016). This is in line with the modest impact of the prior distribution found by Fallaize and Kypraios (2016) by means of a prior sensitivity analysis.

To conclude, we mention two issues that deserve further investigation. First, when AMLE is used for estimating the parameters of the three-dimensional Bingham distribution, computing times are acceptable; however, when the dimension of the problem increases, it may be important to devise more efficient implementations of the ABC rejection algorithm, possibly incorporating recent developments of the ABC literature into AMLE. Second, the possible modification of AMLE as outlined at the end of Section 3 requires a thorough analysis. An algorithm that uses exact MCMC instead of ABC to obtain the posterior samples may be computationally more efficient, but the relative performance of the two approaches need to be carefully studied.

Acknowledgments. An earlier version of this paper was presented at the 8th International Conference of the ERCIM WG on Computational and Methodological Statistics (London, December 12-14, 2015). We thank the participants

Figure 4: Scatterplots of λ_1 and λ_2 in the three samples of the earthquake example. The ABC sample of size is $m = 1000$, and the total number of candidate pairs (λ_1, λ_2) is $n_p = 2 \cdot 10^5$ in each case.

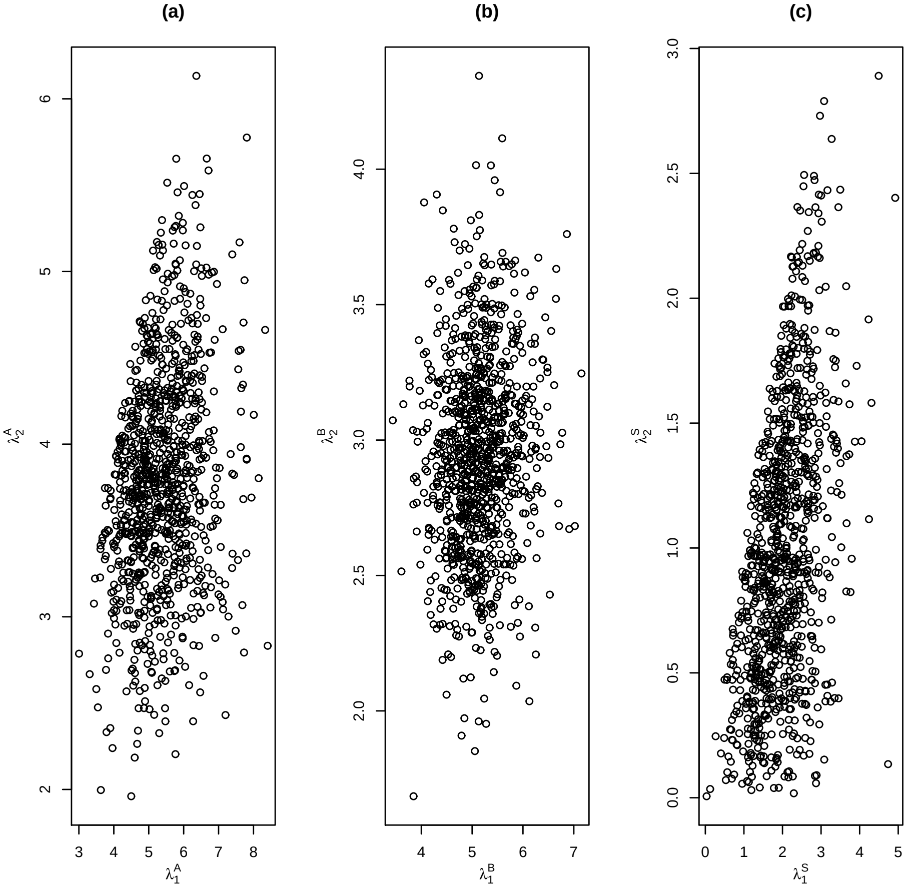
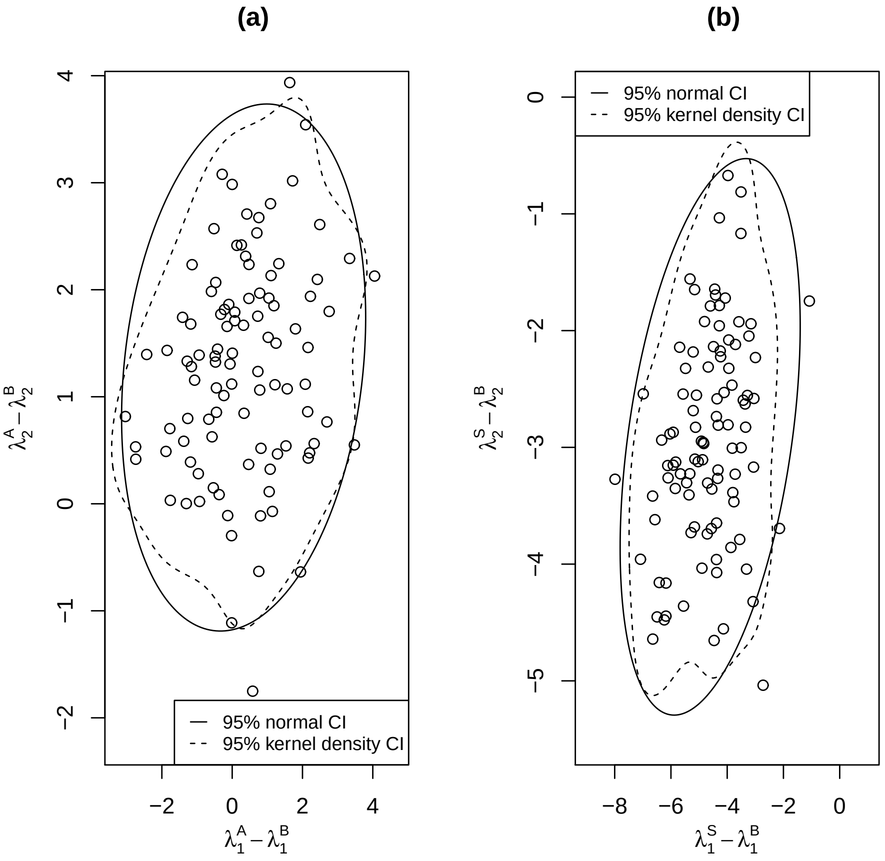


Figure 5: Scatterplots of $\hat{\lambda}_{1i}^{A} - \hat{\lambda}_{1i}^{B}$ vs $\hat{\lambda}_{2i}^{A} - \hat{\lambda}_{2i}^{B}$ (panel (a)) and of $\hat{\lambda}_{1i}^{S} - \hat{\lambda}_{1i}^{B}$ vs $\hat{\lambda}_{2i}^{S} - \hat{\lambda}_{2i}^{B}$ (panel (b)), where $\hat{\lambda}_{ji}^{K}$ is the estimate of the j-th parameter ($j = 1, 2$) in the K-th dataset ($K = A, B, S$) at the i-th replication ($i = 1, \ldots, 100$) of the non-parametric bootstrap procedure discussed in the text. The ABC sample size is $m = 1000$ and the total number of candidate pairs (λ_1, λ_2) is $n_p = 10^5$.



for helpful comments and suggestions. We also thank Richard Arnold (Victoria University of Wellington) for providing the earthquake data and two anonymous reviewers for valuable comments that considerably improved an earlier version of the paper. Finally, we thank Leah Wetherill for useful suggestions. The `R` codes used for simulating and estimating the Bingham distribution are available at `http://marcobee.weebly.com/software.html`.

References

Arnold R, Jupp PE. Statistics of orthogonal axial frames. Biometrika 2013;100:571–86.

Beaumont MA. Approximate Bayesian Computation in evolution and ecology. Annual Review of Ecology, Evolution, and Systematics 2010;41:379–406.

Bee M, Espa G, Giuliani D. Approximate maximum likelihood estimation of the autologistic model. Computational Statistics and Data Analysis 2015;84:14–26.

Bee M, Trapin L. A simple approach to the estimation of Tukey's gh distribution. Journal of Statistical Computation and Simulation 2016;86:3287–302.

Bingham C. Distributions on the sphere and on the projective plane. Ph.D. thesis; Yale University; 1964.

Bingham C. An antipodally symmetric distribution on the sphere. Annals of Statistics 1974;2:1201–25.

Boomsma W, Mardia KV, Taylor CC, Ferkinghoff-Borg J, Krogh A, Hamelryck T. A generative, probabilistic model of local protein structure. Proceedings of the National Academy of Sciences 2008;105:8932–7.

Ciollaro M, Wang D. MeanShift; 2016. URL: `http://CRAN.R-project.org/package=MeanShift`; R package version 1.1-1.

Cressie NAC. Statistics for Spatial Data. Wiley, 1991.

Duong T. Kde: Kernel smoothing; 2014. URL: `http://CRAN.R-project.org/package=kde`; R package version 1.9.2.

Fallaize CJ, Kypraios T. Exact Bayesian inference for the Bingham distribution. Statistics and Computing 2016;26:349–60.

Friel N, Pettitt AN. Likelihood estimation and inference for the autologistic model. Journal of Computational and Graphical Statistics 2004;13:232–46.

Glasserman P. Monte Carlo Methods in Financial Engineering. Springer, 2003.

Hamelryck T, Kent J, Krogh A. Sampling realistic protein conformations using local structural bias. PLoS Computational Biology 2006;e131.

Kent JT, Ganeiber AM, Mardia KV. A new method to simulate the Bingham and related distributions in directional data analysis with applications. http://arxivorg/abs/13108110 2013;.

Kent JT, Hamelryck T. Using the Fisher-Bingham distribution in stochastic models for protein structure. In: Barber S, Baxter PD, Mardia KV, Walls RE, editors. Quantitative Biology, Shape Analysis, and Wavelets. Leeds University Press; 2005. p. 57–60.

Krieger Lassen NC, Juul Jensen D, Conradsen K. On the statistical analysis of orientation data. Acta Crystallographica 1994;A50:741–8.

Kume A, Walker SG. Sampling from compositional and directional distributions. Statistics 2006;16:261–5.

Kume A, Walker SG. On the Bingham distribution with large dimension. Journal of Multivariate Analysis 2014;124:345–52.

Kume A, Wood ATA. Saddlepoint approximations for the Bingham and Fisher-Bingham normalising constants. Biometrika 2005;92:465–76.

Love J. Bingham statistics. In: Gubbins D, Bervera E, editors. Encyclopedia of geomagnetism and paleomagnetism. Springer; 2007. p. 45–7.

Mardia KV, Jupp PE. Directional Statistics. Wiley, 2000.

Mardia KV, Zemroch PJ. Table of maximum likelihood estimates for the Bingham distribution. Journal of Statistical Computation and Simulation 1977;6:29–34.

Møeller J, Pettitt AN, Reeves R, Berthelsen KK. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. Biometrika 2006;93(2):451–8.

Murray I, Ghahramani Z, MacKay D. MCMC for doubly-intractable distributions. In: Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence UAI06. 2006. p. 359–66.

Peel D, Whiten WJ, McLachlan GJ. Fitting mixtures of Kent distributions to aid in joint set identification. Journal of the American Statistical Association 2001;96:56–63.

Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MT. Population growth of human Y chromosomes: A study of Y chromosome microsatellites. Molecular Biology and Evolution 1999;16:1791–1798.

Rubio FJ, Johansen AM. A simple approach to maximum intractable likelihood estimation. Electronic Journal of Statistics 2013;7:1632–54.

Sei T, Kume A. Calculating the normalising constant of the Bingham distribution on the sphere using the holonomic gradient method. Statistics and Computing 2015;25:321–32.

Sousa VC, Fritz M, Beaumont MA, Chikhi L. Approximate Bayesian Computation without summary statistics: The case of admixture. Genetics 2009;181:1507–19.

Takayama N, Koyama T, Sei T, Nakayama H, Nishiyama K. Hgm: Holonomic Gradient Method and Gradient Descent; 2015. R package version 1.11.

Tyler DE. Statistical analysis for the angular central Gaussian distribution on the sphere. Biometrika 1987;74:579–89.