Bayesian Inference for directional data through ABC and homogeneous proper scoring rules

Monica Musio*
Dept. of Mathematics and Computer Science, University of Cagliari, Italy - email: mmusio@unica.it

Valentina Mameli
Dept. of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venezia, Italy, email: mameli.valentina@virgilio.it
Erlis Ruli
Dept. of Statistical Sciences, University of Padova, Italy, email: erlis.ruli@unipd.it
Laura Ventura
Dept. of Statistical Sciences, University of Padova, Italy, email: ventura@stat.unipd.it

Abstract

Both Bayesian and likelihood inference for models for directional data are difficult because typically the density function contains a normalization constant that cannot be computed in closed form. From a Bayesian perspective in such cases we resort to the Approximate Bayesian Computation (ABC) methodology, a computational tool that allows us to approximate posterior distributions when the likelihood function is intractable or unavailable, but it is possible to sample from the model. In this work we propose to use the scoring rule estimating equation as the summary statistic in ABC in order to obtain accurate approximations of the posterior distribution for the unknown parameters of a directional distribution. In particular, we focus on the Hyvärinen scoring rule, which has the property of homogeneity, i.e. the distribution need only be known up to normalisation. The method is illustrated through examples with simulated data.

Keywords: directional distributions; ABC; homogeneous scoring rules; estimating equations.

1. Introduction

In general directional statistics deal with observations on compact Riemannian manifolds. To introduce directional distributions, let's consider a random vector X whose density contains a normalization constant depending on a parameter θ, with $\theta \in \Theta \subseteq \mathbb{R}^m$, $m \geq 1$, given by

$$f(x;\theta) = \frac{1}{c(\theta)} \exp\{-h(x;\theta)\}, \tag{1}$$

where $h(x;\theta)$ is a known function and $c(\theta)$ is the normalizing constant. An example of (1) is the p-dimensional Bingham density, given by $f(x;\theta) = \frac{1}{c(\theta)} \exp\{-x^T\theta x\}\,, \quad x^Tx = 1\,,\ x \in \mathbb{R}^p$. One of the major problems for maximum likelihood estimation in directional models is that the normalizing constant is difficult to evaluate. Recently, Mardia et al. (2016) proposed, as an alternative, the method of the score matching estimation (Hyvärinen, 2005) on a compact oriented Riemannian manifold (see also Dawid and Lauritzen, 2005, and Parry *et al.*, 2012), that avoids the need to work with normalizing constants. Maximum likelihood and score matching estimations can be seen as special cases of a more general methodology based on *proper scoring rules*. A scoring rule (see, for instance, Dawid and Musio, 2014, and references therein) is a special kind of loss function designed to measure the quality of a probability distribution for a random variable, given its observed value. It is proper if it encourages honesty in the probability evaluation. Among the most important proper scoring rules are the logarithmic score, and the Hyvärinen score.

Each proper scoring rule determines a divergence and the minimization of the divergence leads to an estimator. When we use the Kullback-Leiber divergence (associated with the log score), the maximum likelihood estimator is obtained, while if the divergence is that associated with the Hyvärinen score we have the score matching estimator.

From a Bayesian perspective, when the computation of the likelihood is intractable, but it is easy to simulate from the

model, an approximation of the posterior distribution can be obtained by Approximate Bayesian Computation (ABC) techniques (see, e.g. Marin *et al.*, 2012, and references therein). The core idea of ABC is to simulate from the model for different parameter values, and keep those values that produces simulated datasets that approximately match the observed data. The most popular ABC approach is to consider an approximate matching of some summary statistics, evaluated at the observed and simulated data, by means of suitable distances. When the statistics are sufficient for the parameters of the model, this method leads to the exact posterior distribution as the distance tends to zero. However, in realistic applications sufficient statistics are not available and a careful selection of data summaries is necessary.
In this work we try to combine these two different methodologies. More precisely, extending results in Ruli *et al.* (2016), a scoring rule estimating function will be used as a summary statistic in ABC in order to obtain an accurate approximation to the posterior distribution for the unknown parameters of directional distributions, whose normalization constant is not known in closed form. In particular, because of its homogeneous property, we will consider the Hyvärinen scoring rule. The method is illustrated through examples with simulated data.

2. Proper scoring rules

A scoring rule is a way of assessing how well a distribution can do on predicting the value of a random variable X. It assigns a number to every combination of a distribution Q, You quote to represent the uncertantly about X, and the value x of X that Nature will reveal. So if You quote Q and Nature chooses x, You will obtain the unsatisfactoriness score $S(x, Q)$. The score is said to be *proper* (PSR) if, for any distribution P for X, the expected score $S(P, Q) = \mathrm{E}_{X\sim P} S(X, Q)$ is minimised by quoting $Q = P$. Starting from a PSR S, we can define the *discrepancy* or *divergence* of $Q \in \mathcal{P}$ from $P \in \mathcal{P}$ as $D(P, Q) = S(P, Q) - S(P, P)$. For a proper scoring rule $D(P, Q) \geq 0$. Examples of PSRs include the *log score*, $S(x, Q) = -\log q(x)$, where $q(\cdot)$ is the density of Q with respect to some underlying measure μ. In this case $D(P, Q) = \int d\mu(y) \cdot p(y) \ln\{p(y)/q(y)\}$ is the *Kullback-Leibler Discrepancy* $K(P, Q)$. Another important example is the *Hyvärinen score*. Let X be a variable taking values in $\mathcal{X} = \mathbb{R}^p$. It is defined by

$$S(x, Q) = \Delta \ln q(x) + \frac{1}{2} |\nabla \ln q(x)|^2 = \frac{\Delta\sqrt{q(x)}}{\sqrt{q(x)}},$$

where ∇ denotes gradient, and Δ the Laplacian operator, $\sum_{i=1}^{q} \partial^2/(\partial x_i)^2$ on $\mathcal{X}$; $|u|^2 = \langle u, u\rangle$, with $\langle u, v\rangle$ being the inner product between vectors u and v, is the squared length of vector u. An important property of this scoring rule is *homogeneity*: it is unchanged if $q(\cdot)$ is scaled by a positive constant. In particular, $S(x, Q)$ can be computed without knowledge of the normalising constant of the distribution Q.
We can generalize the definition of the Hyvärinen scoring rule on a $p-$dimensional Riemannian manifold $M \subset \mathbb{R}^q$ (Dawid and Lauritzen, 2005). For $u = u(x)$ and $v = v(x)$ being real-valued functions on M, the inner product between u and v is defined by $< u, v >= (\nabla u)^T (G(x))^{-1} (\nabla v)$, where $G(\theta) = (g_{ij}(\theta))$, $i, j = 1, \cdots, p$ is the metric tensor on M and $G(x))^{-1} = (g^{ij}(x))$ is the inverse matrix. A uniform measure on M can be defined in local coordinates by $d\mu(x) = (detG(x))^{1/2} dx$. Moreover, the Laplace-Beltrami operator is given by

$$\Delta u = \sum_{i,j=1}^{p} (det(G(x)))^{-1/2} \partial/\partial x_i \left[det(G(x))^{1/2} g^{ij} \partial u/\partial x_j \right],$$

while the Hyvärinen score on M, has the form

$$S(x, q) = \Delta \log q + \frac{1}{2} ||\nabla \log q||^2. \tag{2}$$

3. Estimation

Let $\{P_\theta : \theta \in \Theta\}$, where Θ is an open subset of $\boldsymbol{R}^m$, be a parametric family of distributions for $X \in \mathcal{X}$. We suppose given a PSR S on X. Let $(x_1, \ldots, x_n)$ be a random sample from P_θ, and denote by $\widehat{P}$ the empirical distribution of the sample. We might estimate θ by that value minimising $D(\widehat{P}, P_\theta)$, where D is the discrepancy associated with S. Equivalently, since $D(\widehat{P}, P_\theta) = S(\widehat{P}, P_\theta) - S(\widehat{P}, \widehat{P})$, we minimise $nS(\widehat{P}, P_\theta)$, which is the total *empirical score*, $\sum_{i=1}^{n} S(x_i, P_\theta)$. This is equivalent to solve the estimating equation

$$\sum_{i=1}^{n} s(x_i, \theta) = 0, \tag{3}$$

where $s(x,\theta) = \partial S(x,P_\theta)/\partial\theta$. This supplies an unbiased estimating equation (see Dawid and Lauritzen, 2005), thereby leading to an M-estimator. Note that when we use the log score, (3) is just the likelihood equation, and we obtain the maximum likelihood estimator. We can thus apply standard results on unbiased estimating equations to describe the properties of the minimum score estimator $\widehat{\theta}_S$. In particular, in repeated *i.i.d.* sampling, under broad regularity conditions on the model, the minimum score estimator $\widehat{\theta}_S$ is asymptotically consistent and normally distributed with mean θ and variance $V(\theta)$, where

$$V(\theta) = K(\theta)^{-1}J(\theta)(K(\theta)^{-1})^T,$$

with $J(\theta) = E\left[s(X,\theta)s(X,\theta)^T\right]$ the *variability matrix*, and $K(\theta) = E\left[\nabla s(X,\theta)^T\right]$ the *sensitivity matrix*. In contrast to the case for full likelihood, J and K are different in general. The matrix $G(\theta) = V(\theta)^{-1}$ is known as the Godambe information matrix (see Dawid *at al.* , 2016). The sandwich form of $V(\theta)$ is due to the failure of the information identity since, in general, $K(\theta) \neq J(\theta)$. In the following we will denote by $\widehat{\theta}_H$ the *Hyvärinen estimator* (i.e. the minimum score estimator based on the Hyvärinen scoring rule).
In the case in which the density belongs to the canonical exponential family, $f(x) = f(x;\theta) \propto \exp\left\{\theta^T t(x)\right\}$, with θ the p-vector of natural parameters, $t(x) = (t_1(x),\cdots,t_p(x))$ is a vector of sufficient statistics and $t_l(x)$ $(l = 1,\cdots,p)$ are twice continuously differentiable w.r.t. x, Mardia et al. (2016) showed that the unbiased estimating equation based on the Hyvärinen score in the manifold M reduces to

$$s(\theta) = W\theta - d, \tag{4}$$

with W and d having components

$$w_{l_1 l_2} = \mathrm{E}\left(< t_{l_1}(x), t_{l_2}(x) >\right), \quad d_l = -\mathrm{E}\left(\Delta t_l(x)\right) \quad (l, l_1, l_2 = 1,\cdots,p). \tag{5}$$

It can be shown that $\hat{\theta}_H \dot{\sim} N(\theta, V(\theta))$, where

$$V(\theta) = K(\theta)^{-1}J(\theta)(K(\theta)^{-1})^T,$$

with

$$J(\theta) = (W\theta - d)(W\theta - d)^T \quad \text{and} \quad K(\theta) = W$$

(see Dawid *at al.* (2016)).

4. ABC methodology

In a Bayesian framework let $\pi(\theta)$ be a prior distribution for the parameter $\theta \in \Theta \subseteq \mathbb{R}^m$, let $L(\theta) = L(\theta;x) = f(x;\theta)$ be the likelihood function based on the data x and let $\pi(\theta|x) \propto \pi(\theta)L(\theta)$ be the posterior distribution of θ. Suppose that the likelihood function is intractable, for example for computational reasons. The main objective of ABC algorithms is to approximate the posterior distribution when it is easy to simulate from the density function $f(x;\theta)$. Let $\eta(x) \in \mathcal{S} \subset \mathbb{R}^p$ be a vector of summary statistics (not necessary sufficient), x^{obs} be the observed data, $\rho(\eta(x), \eta(x^{obs}))$ a metric distance between $\eta(x)$ and its observed value $\eta(x^{obs})$, and ϵ a tolerance threshold. The ABC methods approximates $\pi(\theta|x)$ by

$$\pi_\epsilon(\theta|\eta(x^{obs})) = \int_{\mathcal{S}} \pi_\epsilon(\theta,\eta(x)|\eta(x^{obs}))d\eta$$

and

$$\pi_\epsilon(\theta,\eta(x)|\eta(x^{obs})) \propto \pi(\theta)f(x;\theta)I_{A_{\epsilon,x^{obs}}}(x),$$

where I is the indicator function of the set $A_{\epsilon,x^{obs}}(x) = \left\{x : \rho(\eta(x),\eta(x^{obs})) \leq \epsilon\right\}$. The basic idea of the ABC algorithm relies on simulation by the mixture representation method consisting in generating, say, N values of θ from $\pi(\theta)$ and using them to generate the corresponding N values of η from $\pi(x|\theta)$ at the simulated θ. We accept all values of θ such that $\rho(\eta(x),\eta(x^{obs})) < \epsilon$. For $\epsilon \to 0$ the ABC method has been proven to return a consistent estimator of the posterior $\pi(\theta|\eta(x^{obs}))$. Moreover, if η is sufficient and $\epsilon \to 0$, then $\pi^\epsilon(\theta|\eta(x^{obs}) \to \pi(\theta|x)$. Ruli et al. (2016) propose a variant of the ABC in which a scaled composite likelihood score function, evaluated at the observed composite likelihood estimate, is used a summary statistic (ABC-cs). Generalizing the idea developed in Ruli et al. (2016) in this

paper we propose to use as a summary statistic in ABC a rescaled version of the Hyvärinen score estimator (ABC-H) (see Algorithm 1).

Algorithm 1 A sample $(\theta_1, \cdots, \theta_N)$ from $\pi_\epsilon(\theta|\eta_H(x^{obs}))$

Given the observed data x^{obs} compute the Hyvärinen estimate $\hat{\theta}_H^{obs}$.
Compute $B(\hat{\theta}^{obs})$.
for $i = 1$ to N **do**
  repeat
    draw $\theta^* \sim \pi(\theta)$
    draw $x \sim f(x; \theta^*)$
  until $\rho(\eta_H(\hat{\theta}^{obs}, x), \eta_H(\hat{\theta}^{obs}, x^{obs})) \leq \epsilon$
  set $\theta_i = \theta^*$
end for

We denote by $H(\theta, x)$ the estimating equation based on the Hyärinen score for θ , $\hat{\theta}_H^{obs}$ is the Hyvärinen estimate when $x = x^{obs}$, and we define the rescaled score statistic $\eta_H(\hat{\theta}^{obs}, x) = B(\hat{\theta}^{obs})^{-1} H(\hat{\theta}^{obs}, x)$ where the matrix $B(\theta)$ is such that $J(\theta) = B(\theta)B(\theta)^T$. We have applied this algorithm to simulate from the posterior distribution when data are generate from a directional model using as ρ the Euclidean distance. The following theorem holds:

Theorem 0.1 *The ABC-H algorithm with the rescaled score statistic $\eta(\hat{\theta}^{obs}, x)$, as $\epsilon \longrightarrow 0$, leads to an approximate posterior distribution with the correct curvature and it is invariant to reparametrisations.*

The proof follows the same line as that of Theorem 3.2 of Ruli *et al.* (2016). The adjustment scale of the score statistic is necessary in order to recover the information identity, i.e. the correct curvature, as well as invariance to reparametrisations.

5. Examples

In this section we provide simulation results to assess the performance of this new methodology. We discuss two examples of directional distributions that belong to the exponential family and for which it is then possible compute the sufficient statistics and apply also the "classical" version of the ABC algorithm, that we have used for comparison.

5.1 The von Mises-Fisher distribution

This example focus on inference for the parameters of the von Mises-Fisher density in dimension 3. In general, the von Mises-Fisher density on S^{p-1} is defined as

$$f(z; \theta) \propto \exp(-\theta^T z),$$

where $t(z) = (z_1, \cdots, z_p)^T$ is a vector of sufficient statistics and $\theta = k\mu$ is a vector of parameters where $k \in \mathbb{R}$ and $\mu \in \mathbb{R}^p$ is such that $\|\mu\| = 1$. In the circle it is convenient to write $\mu = (\cos(\tau), \sin(\tau))$ in polar coordinates. Note that, if $p = 2$, and the data are represented in polar coordinates $(z_{1h}, z_{2h}) = (\cos(\phi_h), \sin(\phi_h))$, $(h = 1, \cdots, n)$, the sufficient statistics are $t_1(\phi) = \cos(\phi)$ and $t_2(\phi) = \sin(\phi)$. W and d (defined in (4) and (5)) are equal to

$$w_{11} = \frac{1}{2}(1 - \bar{C}_2),\ w_{12} = -\frac{1}{2}\bar{S}_2,\ w_{22} = \frac{1}{2}(1 + \bar{C}_2),$$

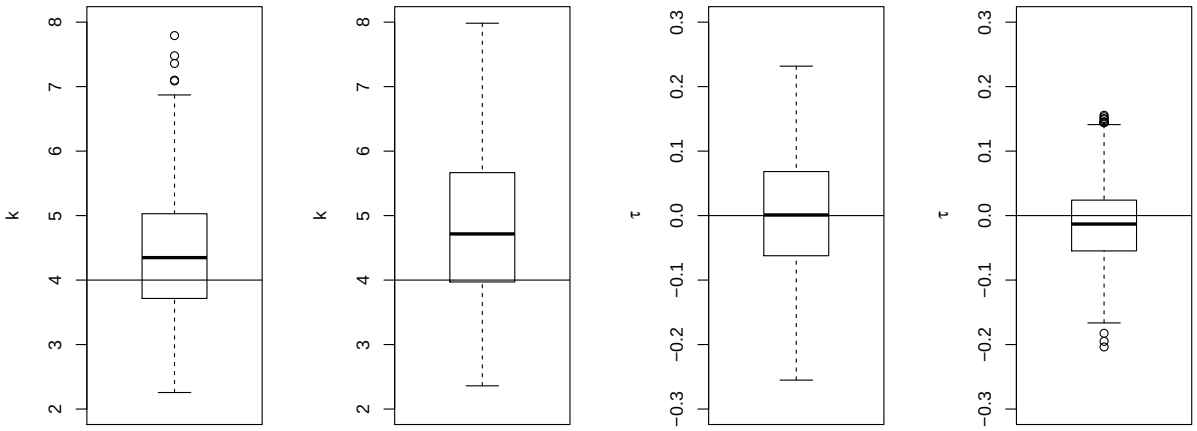
and $d_1 = \bar{C}$, $d_2 = \bar{S}$ where $\bar{C} = \frac{1}{n}\sum_{h=1}^n \cos(\phi_h)$, $\bar{S} = \frac{1}{n}\sum_{h=1}^n \sin(\phi_h)$,

$$\bar{C}_2 = \frac{1}{n}\sum_{h=1}^n \cos(2\phi_h) \quad \bar{S}_2 = \frac{1}{n}\sum_{h=1}^n \sin(2\phi_h) \quad \bar{R}_2 = \sqrt{\bar{C}_2^2 + \bar{S}_2^2}.$$

The score estimator becomes

$$\hat{\tau}_H = atan2\{\bar{C}\bar{S}_2 + \bar{S}(1 - \bar{C}_2), \bar{C}(1 + \bar{C}_2) + \bar{S}\bar{S}_2\} \tag{6}$$

Figure 1: Samples drawn from the parameters k and τ using the $ABC-H$ posterior (left panels) and the "classical" ABC posterior (right panels) with $\epsilon = 0.01$. The horizontal lines represent the true parameter values.



$$\hat{k}_H = \frac{2\sqrt{\bar{R}^2(1+\bar{R}_2^2) + 2(\bar{C}^2 - \bar{S}^2)\bar{C}_2 + 4\bar{C}\bar{S}\bar{S}_2}}{(1-\bar{R}_2^2)}. \tag{7}$$

As an illustration of our methodology we simulate a sample of $n = 100$ drawn from the von Mises-Fisher density in S^1 with $k = 4$, $\tau = 0$. We consider $N = 10^5$ final samples obtained with ϵ fixed at 0.01% quantile of the observed euclidean distance. The boxplots of marginal posterior approximations, computed both with the "classical" ABC and with the ABC-H, are shown in Figure 1.

5.2 The Bingham distribution

In general, the Bingham density on S^{p-1} is defined as

$$f(x;A) = \frac{1}{C(A)} \exp\left\{-x^T A x\right\}, \quad \text{s.t.} \quad x^T x = 1, \quad \text{with} \quad x \in \mathbb{R}^p.$$

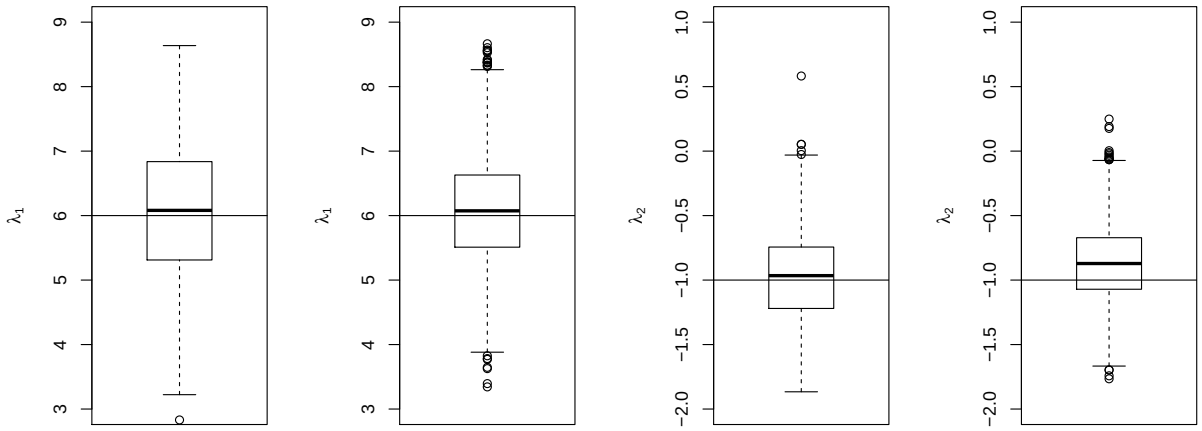
Here A is a symmetric $p \times p$ matrix. Let $A = \Gamma\Lambda\Gamma^T$ be the spectral decomposition of A. The maximum likelihood estimator of Γ is the matrix of eigenvectors of the sum and product matrix $\sum_j x_j x_j^T$. If $X \sim Bingham(A)$ then $Z = \Gamma^T X \sim Bingham(\Lambda)$, where Γ represents the orientation parameters and $\Lambda = diag(\lambda_j)$ represent the concentration parameters, with $\sum_{j=1}^n \lambda_j = tr(A) = 0$. Taking $A = \Lambda$, then $x^T\Lambda x = \sum_{j=1}^{p-1} \lambda_j (x_j^2 - x_p^2)$, with the constraint $\sum_{j=1}^n \lambda_j = tr(A) = 0$. $C(\Lambda) = \int_{x \in S^{p-1}} \exp\left\{-\sum_{i=1}^{p-1} \lambda_i x_i^2\right\} dS^{p-1}(x)$. W and d for $i, j = 1, \cdots, p-1$, are defined as

$$w_{ij} = \begin{cases} (4/n)\sum_k \left(x_{ik}^2 + x_{pk}^2 - \left(x_{ik}^2 - x_{pk}^2\right)^2\right) & \text{if} \quad i = j \\ (4/n)\sum_k \left(x_{pk}^2 - \left(x_{ik}^2 - x_{pk}^2\right)\left(x_{jk}^2 - x_{pk}^2\right)\right) & \text{otherwise,} \end{cases}$$

$$d_i = -(2p/n)\sum_k \left(x_{ik}^2 - x_{pk}^2\right)$$

(see Mardia et al., 2016)). As an illustration of our methodology we have applied the "classical" ABC and the ABC-H methods in the case in which p=3. We simulate a sample of $n = 350$ drawn for the Bingham density with $\lambda_1 = 6$,

Figure 2: Samples drawn from the parameters λ_1 and λ_2 using the $ABC - H$ posterior (left panels) and the "classical" ABC posterior (right panels) with $\epsilon = 0.0025$. The horizontal lines represent the true parameter values.



$\lambda_2 = -1$. We consider $N = 3 \times 10^5$ final samples obtained with ϵ fixed at 0.0025% quantile of the observed euclidean distance. The boxplots of marginal posterior approximations are shown in Figure 2.

6. Conclusions

Our results suggest that combining the ABC approach with the Hyvärinen scoring rule estimating equation is a promising approach which avoids the calculation of the normalizing constant. Further works will be devoted to investigate the method in a real case study and for other directional models.

References

Dawid, A. P., Lauritzen, S. L. (2005). The geometry of decision theory. In *Proceedings of the Second International Symposium on Information Geometry and its Applications*, 22–28. University of Tokyo.

Dawid, A.P., & Musio, M. (2014). Theory and Applications of Proper Scoring Rules, *Metron*, **72**, 169–183.

Dawid, A. P., Musio, M., Ventura, L. (2016). Minimum scoring rule inference. *Scand J Stat*, **43**, 1, 123–138.

Hyvärinen, A. (2005). Estimation of Non-Normalized Statistical Models by Score Matching. *Mach Learn Res*, **6**, 695–709.

Mardia, K.V., & Jupp, P.E. (1999). Directional Statistics (2nd ed.). Chichester, UK: Wiley.

Mardia, K.V., Kent, J.T., Laha, A.K. (2016). Score matching estimators for directional distributions, *arXiv*:1604.08470v1.

Marin, J.M., Pudlo, P., Robert, C.P., & Ryder, R.J. (2012). Approximate Bayesian computational methods. *Stat Comput*, **22**, 1167–1180.

Parry, M.F., Dawid, A.P., & Lauritzen, S.L. (2012). Proper Local Scoring Rules, *Ann Statist*, **561**, 40–92.

Ruli, E., Sartori, N., & Ventura, L. (2016). Approximate Bayesian computation with composite score functions. *Stat Comput*, **26**, 3, 679–692.