A Stein Goodness-of-fit Test for Directional Distributions

Wenkai Xu
Gatsby Computational Neuroscience Unit

Takeru Matsuda
University of Tokyo
RIKEN Center for Brain Science

Abstract

In many fields, data appears in the form of direction (unit vector) and usual statistical procedures are not applicable to such directional data. In this study, we propose nonparametric goodness-of-fit testing procedures for general directional distributions based on kernel Stein discrepancy. Our method is based on Stein's operator on spheres, which is derived by using Stokes' theorem. Notably, the proposed method is applicable to distributions with an intractable normalization constant, which commonly appear in directional statistics. Experimental results demonstrate that the proposed methods control type-I error well and have larger power than existing tests, including the test based on the maximum mean discrepancy.

1 INTRODUCTION

In many applications, data is obtained in the form of directions and they are naturally identified with a vector on the unit hypersphere $\mathcal{S}^{d-1} = \{x \in \mathbb{R}^d \mid \|x\| = 1\} \subset \mathbb{R}^d$. For example, wind direction is represented by a vector on the unit circle $\mathcal{S}^1 \subset \mathbb{R}^2$ [Genton and Hering, 2007, Hering and Genton, 2010], while the protein structure is described by vectors on the unit sphere $\mathcal{S}^2 \subset \mathbb{R}^3$ [Hamelryck et al., 2006]. In addition, usual multivariate data in $\mathbb{R}^d$ is transformed to directional data by applying normalization, and such transformation is useful to analyze scale-invariant features. For example, [Banerjee et al., 2005] transformed text document and gene expression data into directional data and applied model-based clustering. Also, [Wang et al., 2017] showed that projecting face images to a unit hypersphere improves face recognition performance by convolutional neural networks. Statistical methods for such directional data have been widely studied in the field of directional statistics [Mardia and Jupp, 1999, Ley and Verdebout, 2017], and many statistical models of directional distributions have been proposed. One characteristic feature of directional distributions is that they often involve an intractable normalization constant. For example, the Fisher-Bingham distribution [Kent, 1982] is defined by an unnormalized density

$$p(x \mid A, b) \propto \exp(x^\top Ax + b^\top x), \quad x \in \mathcal{S}^{d-1},$$

and its normalization constant is not represented in closed form. Such intractable normalization constant makes statistical inferences for directional distributions computationally difficult. While directional data are becoming increasingly important in many applications such as bioinformatics, meteorology, chronobiology, and text/image analysis, to the best of our knowledge, goodness-of-fit testing for general directional distributions is not well established.

Several studies [Chwialkowski et al., 2016, Liu et al., 2016] have proposed kernel-based goodness-of-fit testing procedures for distributions on $\mathbb{R}^d$. These methods employ a model discrepancy measure called kernel Stein discrepancy (KSD), which is based on Stein's method [Barbour and Chen, 2005, Chen et al., 2010] and reproducing kernel Hibert space (RKHS) theory [Berlinet and Thomas, 2004, Muandet et al., 2017]. Notably, the KSD test is applicable to unnormalized models, because it utilizes only the derivative of the logarithm of the density like score matching [Hyvärinen, 2005]. This method is also applicable to model comparison [Jitkrittum et al., 2018, 2017, Kanagawa et al., 2019]. Recently, it has been extended to discrete distributions [Yang et al., 2018] and point processes [Yang et al., 2019]. On the other hand, applying Stein's method in the context of manifold structure is previously studied in [Barp et al., 2018] focusing on numerical integration problems for scalar functions and in [Liu and Zhu, 2018] dealing with Bayesian inference on density functions.

In this study, we develop goodness-of-fit testing procedures for general directional distributions by extending kernel Stein discrepancy. Our contributions are

Proceedings of the 23[rd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

as follows. We derive Stein's operator on the unit hypersphere $\mathcal{S}^{d-1}$ via Stokes' theorem and introduce directional kernel Stein discrepancy (dKSD). We propose dKSD-based goodness-of-fit testing procedures for general directional distributions including unnormalized ones, which do not require to sample from the null distribution. We show that the proposed methods control type-I error well and have larger power than existing tests in simulation.

2 BACKGROUND

2.1 Directional Distributions

Several distributions have been proposed for describing directional data on the unit hypersphere $\mathcal{S}^{d-1} = \{x \in \mathbb{R}^d \mid \|x\| = 1\}$. Here, we present two representative directional distributions: von Mises-Fisher and Fisher-Bingham. Figure 1 shows samples from these distributions on $\mathcal{S}^2$. See [Mardia and Jupp, 1999] for more detail.

In this paper, we define the probability density of directional distributions by taking the uniform distribution on $\mathcal{S}^{d-1}$ as base measure. Namely, the density of the uniform distribution is $p(x) \equiv 1$.

The von Mises-Fisher (or von Mises when $d = 2$) distribution is a directional counterpart of the isotropic Gaussian distribution on $\mathbb{R}^d$. Its density is given by

$$p(x \mid \mu, \kappa) = \frac{1}{C_d(\kappa)} \exp(\kappa \mu^\top x), \tag{1}$$

for $x \in \mathcal{S}^{d-1}$, where $\mu \in \mathcal{S}^{d-1}$, $\kappa > 0$,

$$C_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)},$$

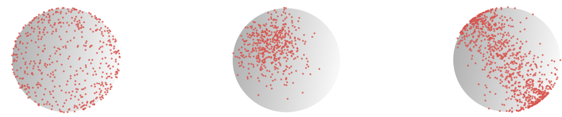
and I_v is the modified Bessel function of the first kind and order v. It is a unimodal distribution with peak at μ and degree of concentration specified by κ.

The Fisher-Bingham (or Kent) distribution is an extension of the von Mises-Fisher distribution [Kent, 1982]. Its density is given by

$$p(x \mid A, b) = \frac{1}{Z(A, b)} \exp(x^\top A x + b^\top x), \tag{2}$$

for $x \in \mathcal{S}^{d-1}$, where $A \in \mathbb{R}^{d \times d}$ is symmetric and $b \in \mathbb{R}^d$. The normalization constant $Z(A, b)$ is not represented in closed form in general.

The goodness-of-fit test for general directional distributions is not well established, to the best of our knowledge. Tests for specific distributions such as uniform [Figueiredo, 2007, García-Portugués and Verdebout, 2018, Mardia and Jupp, 1999] and von Mises-Fisher



(a) Uniform (b) von Mises-Fisher (c) Fisher-Bingham

Figure 1: Samples from directional distributions on $\mathcal{S}^2$

[Figueiredo, 2012, Mardia et al., 1984] cannot be readily extended to general directional distributions. Although [Boente et al., 2014] proposed testing procedures based on the kernel density estimator, they are difficult to apply to unnormalized models such as the Fisher-Bingham distribution (2), because they require the normalization constant of the null model to calculate the L^p test statistics.

2.2 Kernel Stein Discrepancy on $\mathbb{R}^d$

Here, we briefly review the goodness-of-fit testing with kernel Stein discrepancy on $\mathbb{R}^d$ by [Chwialkowski et al., 2016, Liu et al., 2016], which is inspired from [Gorham and Mackey, 2015, Ley et al., 2017]. See [Chwialkowski et al., 2016, Liu et al., 2016] for more detail.

Let q be a smooth probability density on $\mathbb{R}^d$. For a smooth function $f = (f_1, \dots, f_d) : \mathbb{R}^d \to \mathbb{R}^d$, Stein's operator $\mathcal{T}_q$ is defined by

$$\mathcal{T}_q f(x) = \sum_{i=1}^{d} \left(f_i(x) \frac{\partial}{\partial x^i} \log q(x) + \frac{\partial}{\partial x^i} f_i(x) \right). \tag{3}$$

From integration by parts on $\mathbb{R}^d$, we obtain the equality

$$\mathrm{E}_q[\mathcal{T}_q f] = 0$$

under mild regularity conditions. Since Stein's operator $\mathcal{T}_q$ depends on the density q only through the derivatives of $\log q$, it does not involve the normalization constant of q, which is a useful property for dealing with unnormalized models [Hyvärinen, 2005].

Let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) on $\mathbb{R}^d$ and $\mathcal{H}^d$ be its product. By using Stein's operator, kernel Stein discrepancy (KSD) [Chwialkowski et al., 2016, Liu et al., 2016] between two densities p and q is defined as $\mathrm{KSD}(p, q) = \sup_{\|f\|_{\mathcal{H}^d} \leq 1} \mathrm{E}_p[\mathcal{T}_q f]$.

It is shown that $\mathrm{KSD}(p, q) \geq 0$ and $\mathrm{KSD}(p, q) = 0$ if and only if $p = q$ under mild regularity conditions [Chwialkowski et al., 2016]. Thus, KSD is a proper discrepancy measure between densities. After some calculation, $\mathrm{KSD}(p, q)$ is rewritten as

$$\mathrm{KSD}^2(p, q) = \mathrm{E}_{x, \tilde{x} \sim p}[h_q(x, \tilde{x})], \tag{4}$$

where h_q does not involve p.

Now, suppose we have samples $x_1, \dots, x_n$ from unknown density p on $\mathbb{R}^d$. Based on (4), estimates of $\mathrm{KSD}^2(p,q)$ are obtained by using U-statistics or V-statistics. These estimates can be used to test the hypothesis $H_0 : p = q$. The critical value is determined by bootstrap based on the theory of U-statistics or V-statistics. In this way, a general method of goodness-of-fit test on $\mathbb{R}^d$ is obtained, which is applicable to unnormalized models as well.

3 STEIN'S OPERATOR ON $\mathcal{S}^{d-1}$

In this section, we derive Stein's operator for distributions on spheres. The derivation is based on Stokes' theorem, which is a fundamental theorem in differential geometry.

3.1 Differential Forms and Stokes' Theorem

The original derivation of Stein's operator for distributions on $\mathbb{R}^d$ was based on integration by parts, in which the boundary term vanishes due to the decaying property of the probability density. We need a different argument for spheres because its topology is different from $\mathbb{R}^d$. Specifically, differential forms and Stokes' theorem are essential to discuss integration by parts on spheres. Here, we briefly review these concepts. See [Flanders, 1963, Spivak, 2018] for more detail and rigorous treatments.

Let M be a d-dimensional closed manifold and take its local coordinate system $x^1, \dots, x^d$. We introduce symbols $\mathrm{d}x^1, \dots, \mathrm{d}x^d$ and an associative and anti-symmetric operation $\wedge$ between them called the wedge product: $\mathrm{d}x^i \wedge \mathrm{d}x^j = -\mathrm{d}x^j \wedge \mathrm{d}x^i$. Note that $\mathrm{d}x^i \wedge \mathrm{d}x^i = 0$. Then, a p-form ω on M $(0 \leq p \leq d)$ is defined as

$$\omega = \sum_{i_1 \cdots i_p} f_{i_1 \cdots i_p} \mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_p},$$

where the sum is taken over all p-tuples $\{i_1, \cdots, i_p\} \subset \{1, \dots, d\}$ and each $f_{i_1 \cdots i_p}$ is a smooth function on M. The exterior derivative $\mathrm{d}\omega$ of ω is defined as the $(p+1)$-form given by

$$\mathrm{d}\omega = \sum_{i_1 \cdots i_p} \sum_{i=1}^{d} \frac{\partial f_{i_1 \cdots i_p}}{\partial x^i} \mathrm{d}x^i \wedge \mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_p}.$$

For another coordinate system $y^1, \dots, y^d$ on M, the differential form is transformed by

$$\mathrm{d}y^j = \sum_{i=1}^{d} \frac{\partial y^j}{\partial x^i} \mathrm{d}x^i.$$

The integration of a d-form on a d-dimensional manifold is naturally defined like the usual integration on $\mathbb{R}^d$ and invariant with respect to the coordinate selection. Correspondingly, the integration by parts formula on $\mathbb{R}^d$ is generalized in the form of Stokes' theorem.

Theorem 1 (Stokes' theorem)**.** *Let ∂M be the boundary of M and ω be a $(d-1)$-form on M. Then,*

$$\int_M \mathrm{d}\omega = \int_{\partial M} \omega.$$

In particular, since $\partial \mathcal{S}^{d-1}$ is empty, we obtain the following.

Corollary 1. *Let ω be a $(d-2)$-form on $\mathcal{S}^{d-1}$. Then,*

$$\int_{S^{d-1}} \mathrm{d}\omega = 0. \tag{5}$$

Corollary 1 plays an important role in the derivation of Stein's operator on $\mathcal{S}^{d-1}$.

3.2 Spherical Coordinate System

In this paper, we use the spherical coordinate system $\theta = (\theta^1, \dots, \theta^{d-1})$ on $\mathcal{S}^{d-1}$ defined by

$$\begin{pmatrix} \theta^1 \\ \theta^2 \\ \theta^3 \\ \vdots \\ \theta^{d-1} \end{pmatrix} \mapsto \begin{pmatrix} \cos\theta^1 \\ \sin\theta^1 \cos\theta^2 \\ \sin\theta^1 \sin\theta^2 \cos\theta^3 \\ \vdots \\ \sin\theta^1 \cdots \sin\theta^{d-1} \end{pmatrix} \in \mathcal{S}^{d-1}, \tag{6}$$

where $(\theta^1, \dots, \theta^{d-2}) \in [0, \pi)^{d-2}$ and $\theta^{d-1} \in [0, 2\pi)$. In this coordinate system, the volume element [Flanders, 1963] is given by

$$\mathrm{d}S = J(\theta^1, \dots, \theta^{d-1}) \mathrm{d}\theta^1 \wedge \cdots \wedge \mathrm{d}\theta^{d-1},$$

where

$$J(\theta^1, \dots, \theta^{d-1}) = \sin^{d-2}(\theta^1) \sin^{d-3}(\theta^2) \cdots \sin(\theta^{d-2}).$$

Note that $J(\theta^1) = 1$ when $d = 2$. Since the surface area of $\mathcal{S}^{d-1}$ is $S_{d-1} = 2\pi^{d/2}/\Gamma(d/2)$, the uniform distribution on $\mathcal{S}^{d-1}$ corresponds to the $(d-1)$-form η on $\mathcal{S}^{d-1}$ given by

$$\eta = \frac{1}{S_{d-1}} J(\theta^1, \dots, \theta^{d-1}) \mathrm{d}\theta^1 \wedge \cdots \wedge \mathrm{d}\theta^{d-1}.$$

By using this, the directional distribution on $\mathcal{S}^{d-1}$ with density p is represented by the $(d-1)$-form ω given by

$$\omega = p\eta.$$

Thus, expectation of a function g with respect to p is obtained by

$$\begin{aligned} \mathrm{E}_p[g] &= \int_{\mathcal{S}^{d-1}} g\omega \\ &= \frac{1}{S_{d-1}} \int_0^{2\pi} \int_0^{\pi} \cdots \int_0^{\pi} g(\theta) p(\theta) J(\theta) \mathrm{d}\theta^1 \cdots \mathrm{d}\theta^{d-1}. \end{aligned}$$

3.3 Stein's Operator on $\mathcal{S}^{d-1}$

Now, we derive Stein's operator on $\mathcal{S}^{d-1}$ in the spherical coordinate.

Theorem 2 (Stein's operator on $\mathcal{S}^{d-1}$)**.** *Let p be a smooth probability density on $\mathcal{S}^{d-1}$. For smooth functions $f_1,\dots,f_{d-1}:\mathcal{S}^{d-1}\to\mathbb{R}$, define a function $\mathcal{A}_p f:\mathcal{S}^{d-1}\to\mathbb{R}$ by*

$$\mathcal{A}_p f=\sum_{i=1}^{d-1}\left(\frac{\partial f_i}{\partial\theta^i}+f_i\frac{\partial}{\partial\theta^i}\log(pJ)\right). \tag{7}$$

Then,

$$\mathrm{E}_p[\mathcal{A}_p f]=0.$$

Proof. Let $\mathrm{d}\theta^{(-i)}=\mathrm{d}\theta^{i+1}\wedge\cdots\wedge\mathrm{d}\theta^{d-1}\wedge\mathrm{d}\theta^1\cdots\wedge\mathrm{d}\theta^{i-1}$ be a $(d-2)$-form on $\mathcal{S}^{d-1}$ for $i=1,\dots,d-1$. Consider a $(d-2)$-form ω on $\mathcal{S}^{d-1}$ defined by

$$\omega=\sum_{i=1}^{d-1}f_i\mathrm{d}\theta^{(-i)}.$$

Then,

$$\begin{aligned}\mathrm{d}(pJ\omega)&=\sum_{i=1}^{d-1}\left(f_i\frac{\partial}{\partial\theta^i}(pJ)+pJ\frac{\partial f_i}{\partial\theta^i}\right)\mathrm{d}\theta^1\wedge\cdots\wedge\mathrm{d}\theta^{d-1}\\&=(pJ\mathcal{A}_p f)\mathrm{d}\theta^1\wedge\cdots\wedge\mathrm{d}\theta^{d-1}.\end{aligned}$$

From Corollary 1, $\mathrm{E}_p[\mathcal{A}_p f]=\int_{\mathcal{S}^{d-1}}\mathrm{d}(pJ\omega)=0$. □

Although Stein's operator on $\mathcal{S}^{d-1}$ has a similar form to the original Stein's operator on $\mathbb{R}^d$ in (3), its derivation is different from the original one due to the topology of spheres. Whereas the original derivation on $\mathbb{R}^d$ required vanishing density at the boundary, our derivation on $\mathcal{S}^{d-1}$ is free from such assumption. Also note that, although we use the spherical coordinate system in this paper, we can derive Stein's operator in other coordinate systems as well.

4 KERNEL STEIN DISCREPANCY ON $\mathcal{S}^{d-1}$

Based on Stein's operator on $\mathcal{S}^{d-1}$ in (7), we define the Stein discrepancy and its kernelized counterpart between two directional distributions via kernel mean embeddings, similar to [Chwialkowski et al., 2016, Liu et al., 2016].

Let $\mathcal{H}$ be an RKHS on $\mathcal{S}^{d-1}$ with reproducing kernel k and let $\mathcal{H}^{d-1}$ be its product. We define the directional kernel Stein discrepancy (dKSD) by

$$\mathrm{dKSD}(p,q)=\sup_{\|f\|_{\mathcal{H}^{d-1}}\le 1}\mathbb{E}_p[\mathcal{A}_q f] \tag{8}$$

Let x and $\tilde{x}$ be points on $\mathcal{S}^{d-1}$ with spherical coordinates θ and $\tilde{\theta}$, respectively. We identify the kernel function $k(x,\tilde{x})$ with a function of θ and $\tilde{\theta}$ through (6) and take its derivatives. For example, when $d=2$ and $k(x,\tilde{x})=\exp(\kappa x^\top\tilde{x})=\exp(\kappa\cos(\theta-\tilde{\theta}))$, we have

$$\frac{\partial^2}{\partial\theta\partial\tilde{\theta}}k(x,\tilde{x})=\kappa(\cos(\theta-\tilde{\theta})-\kappa\sin^2(\theta-\tilde{\theta}))\exp(\kappa\cos(\theta-\tilde{\theta})).$$

Let

$$\begin{aligned}h_q(x,\tilde{x})=&\,k(x,\tilde{x})\sum_{i=1}^{d-1}\frac{\partial}{\partial\theta^i}\log(q(\theta)J(\theta))\frac{\partial}{\partial\tilde{\theta}^i}\log(q(\tilde{\theta})J(\tilde{\theta}))\\&+\sum_{i=1}^{d-1}\frac{\partial}{\partial\theta^i}\log(q(\theta)J(\theta))\frac{\partial}{\partial\tilde{\theta}^i}k(x,\tilde{x})\\&+\sum_{i=1}^{d-1}\frac{\partial}{\partial\tilde{\theta}^i}\log(q(\tilde{\theta})J(\tilde{\theta}))\frac{\partial}{\partial\theta^i}k(x,\tilde{x})\\&+\sum_{i=1}^{d-1}\frac{\partial^2}{\partial\theta^i\partial\tilde{\theta}^i}k(x,\tilde{x}).\end{aligned}$$

Similarly to the original KSD (4), dKSD is rewritten as follows.

Theorem 3. *Assume p and q are smooth densities on $\mathcal{S}^{d-1}$ and the reproducing kernel k of $\mathcal{H}$ is a smooth function on $\mathcal{S}^{d-1}\times\mathcal{S}^{d-1}$. Then,*

$$\mathrm{dKSD}^2(p,q)=\mathbb{E}_{x,\tilde{x}\sim p}[h_q(x,\tilde{x})]. \tag{9}$$

Proof. Since Stein's operator $\mathcal{A}_q$ is linear from (7), $\mathrm{E}_p[\mathcal{A}_q f]$ is a linear functional of $f\in\mathcal{H}^{d-1}$. Then, from Riesz representation theorem, there uniquely exists $g=(g_1,\dots,g_{d-1})\in\mathcal{H}^{d-1}$ such that $\mathrm{E}_p[\mathcal{A}_q f]=(f,g)_{\mathcal{H}^{d-1}}$. By using the reproducing property of $\mathcal{H}$, we obtain

$$g_i(x)=\mathrm{E}_{\tilde{x}\sim p}\left[k(x,\tilde{x})\frac{\partial}{\partial\tilde{\theta}^i}\log(q(\tilde{\theta})J(\tilde{\theta}))+\frac{\partial}{\partial\tilde{\theta}^i}k(x,\tilde{x})\right], \tag{10}$$

for $i=1,\dots,d-1$. Thus, the maximization in (8) is attained by $f=g/\|g\|_{\mathcal{H}^{d-1}}$ and $\mathrm{dKSD}(p,q)=\|g\|_{\mathcal{H}^{d-1}}$. Therefore, after straightforward calculations, we obtain (9). □

Importantly, the function h_q in (9) does not involve p. Therefore, we can estimate $\mathrm{dKSD}^2(p,q)$ based on samples from p and apply it to goodness-of-fit testing.

From the following theorem, $\mathrm{dKSD}^2(p,q)$ provides a proper discrepancy measure between directional distributions. Let

$$L_i(x)=\frac{\partial}{\partial\theta^i}\log\frac{q(\theta)}{p(\theta)},\quad i=1,\dots,d-1.$$

Theorem 4. *Let p and q be smooth densities on $\mathcal{S}^{d-1}$. Assume the following: 1) The kernel k is C_0-universal* [Carmeli et al., 2010, Definition 4.1]*; 2) $\mathrm{E}_{x,\tilde{x}\sim p} h_p(x,\tilde{x}) < \infty$; 3) $\mathrm{E}_p \|L(x)\|^2 < \infty$.*

Then, $\mathrm{dKSD}^2(p,q) \geq 0$ and $\mathrm{dKSD}^2(p,q) = 0$ if and only if $p = q$.

Proof. From the proof of Theorem 3, we have $\mathrm{dKSD}^2(p,q) = \|g\|^2_{\mathcal{H}^{d-1}} \geq 0$, where $g = (g_1, \dots, g_{d-1})$ is defined as (10). If $p = q$, then $\mathrm{dKSD}^2(p,q) = 0$ from the definition (8) and Theorem 2. Conversely, if $\mathrm{dKSD}^2(p,q) = 0$, then $g = 0$, namely $g_i = 0$ for $i = 1, \dots, d-1$. Then, from $\log(q/p) = \log(qJ) - \log(pJ)$, we obtain

$$\mathrm{E}_{\tilde{x}\sim p}\left[L_i(\tilde{x})k(x,\tilde{x})\right] = g_i(x) - \mathrm{E}_{\tilde{x}\sim p}\left[\mathcal{A}_p k(x,\tilde{x})\right] = 0,$$

for every x. Since k is C_0-universal, it implies $L_i = 0$ [Carmeli et al., 2010, Theorem 4.2b]. Therefore, $\log(q/p)$ is constant on $\mathcal{S}^{d-1}$. Since both p and q are densities on $\mathcal{S}^{d-1}$ that integrate to one, we obtain $p = q$.

□

RKHS on $\mathcal{S}^{d-1}$ To apply dKSD for goodness-of-fit testing, we need to choose an RKHS on $\mathcal{S}^{d-1}$ that satisfies the conditions in Theorem 4. In this paper, we use the RKHS generated by the von-Mises Fisher kernel:

$$k(x,\tilde{x}) = \exp(\kappa x^\top \tilde{x}), \quad x, \tilde{x} \in \mathcal{S}^{d-1},$$

where $\kappa > 0$ is a concentration parameter that has a similar role to the band-width parameter in the Gaussian kernel. Since both x and $\tilde{x}$ have unit norm, their inner product $x^\top \tilde{x}$ is equal to the cosine of their angular separation. We discuss the method to choose κ in Section 5.3. See [Gneiting et al., 2013] for general discussion on RKHS on $\mathcal{S}^{d-1}$.

5 GOODNESS-OF-FIT TESTING VIA dKSD

In this section, we develop goodness-of-fit testing procedures based on dKSD. Suppose $x_1, \cdots, x_n \sim p$ and we test $H_0 : p = q$ with significance level α.

5.1 Test with U-statistics

From (9), an unbiased estimate of $\mathrm{dKSD}^2(p,q)$ is obtained in the form of U-statistics [Lee, 1990]:

$$\mathrm{dKSD}_u^2(p,q) = \frac{1}{n(n-1)} \sum_{i \neq j} h_q(x_i, x_j). \tag{11}$$

Algorithm 1 dKSD test via U-statistics (dKSDu)

Input:
samples $x_1, \dots, x_n \sim p$
null density q
kernel function k
test size α
bootstrap sample size B

Objective: Test $H_0 : p = q$ versus $H_1 : p \neq q$.

Test procedure:
1: Compute the U-statistics $\mathrm{dKSD}_u^2(p,q)$ via (11).
2: Compute $n \times n$ matrix H with $H_{ij} = h_q(x_i, x_j)$ and its eigenvalues $\hat{c}_1, \dots, \hat{c}_n$.
3: **for** $t = 1 : B$ **do**
4: Sample $Z_1, \dots, Z_n \sim \mathcal{N}(0,1)$ independently.
5: Compute $S_t = \sum_{j=1}^n \hat{c}_j (Z_j^2 - 1)$.
6: **end for**
7: Determine the $(1-\alpha)$-quantile $\gamma_{1-\alpha}$ of $S_1, \dots, S_B$.

Output:
Reject H_0 if $n \cdot \mathrm{dKSD}_u^2(p,q) > \gamma_{1-\alpha}$; otherwise do not reject.

From the U-statistics theory [Lee, 1990], the asymptotic distribution of $\mathrm{dKSD}_u^2(p,q)$ is explicitly obtained as follows. Here, $\xrightarrow{d}$ denotes the convergence in distribution.

Theorem 5. *Under the conditions in Theorem 4, the following statements hold.*

1. *Under $H_0 : p = q$, the asymptotic distribution of $\mathrm{dKSD}_u^2(p,q)$ is*

$$n \cdot \mathrm{dKSD}_u^2(p,q) \xrightarrow{d} \sum_{j=1}^{\infty} c_j (Z_j^2 - 1), \tag{12}$$

where Z_j are i.i.d. standard Gaussian random variables and c_j are the eigenvalues of the kernel $h_q(x,\tilde{x})$ under $p(\tilde{x})$:

$$\int h_q(x,\tilde{x}) \phi_j(\tilde{x}) p(\tilde{x}) \mathrm{d}\tilde{x} = c_j \phi_j(x), \quad \phi_j(x) \neq 0.$$

2. *Under $H_1 : p \neq q$, the asymptotic distribution of $\mathrm{dKSD}_u^2(p,q)$ is*

$$\sqrt{n}(\mathrm{dKSD}_u^2(p,q) - \mathrm{dKSD}^2(p,q)) \xrightarrow{d} \mathcal{N}(0, \sigma_u^2),$$

where $\sigma_u^2 = \mathrm{Var}_{x\sim p}[\mathrm{E}_{\tilde{x}\sim p}[h_q(x,\tilde{x})]] \neq 0$.

The proof is essentially the same with Theorem 4.1 of [Liu et al., 2016]. We employ Theorem 5 for goodness-of-fit. Namely, we generate bootstrap samples from an approximation of the null distribution (12) of $n \cdot \mathrm{dKSD}_u^2(p,q)$ and compare their $(1-\alpha)$ quantile with the realized value of $n \cdot \mathrm{dKSD}_u^2(p,q)$. To approximate the null, we truncate the infinite sum in (12) following [Gretton et al., 2009]: $\sum_{j=1}^{n} \hat{c}_j(Z_j^2 - 1)$, where $\hat{c}_j$ are eigenvalues of the $n \times n$ matrix H with $H_{ij} = h(x_i, x_j)$ and $Z_1, \ldots, Z_n$ are independent standard Gaussian random variables. The testing procedure is outlined in Algorithm 1.

5.2 Wild Bootstrap Test with V-statistics

Here, we propose another testing procedure with wild bootstrap adapted from [Chwialkowski et al., 2016, Section 2.2], which is applicable even when observations $x_1, \ldots, x_n \sim p$ are not independent. It is based on the V-statistics

$$\mathrm{dKSD}_b^2(p, q) = \frac{1}{n^2} \sum_{i,j} h_q(x_i, x_j). \tag{13}$$

For each $t = 1, \ldots, B$, we sample uniform i.i.d. variables $U_1, \ldots, U_n \sim \mathrm{U}[0, 1]$, let $W_{0,t} = 1$ and define

$$W_{i,t} = \mathbf{1}_{\{U_i > a_t\}} W_{i-1,t} - \mathbf{1}_{\{U_i < a_t\}} W_{i-1,t}, \tag{14}$$

for $i = 1, \ldots, n$, where $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function and a_t is the probability of sign change which is set to 0.5 when $x_1, \ldots, x_n$ are independent.

Then, wild bootstrap samples are given by

$$S_t = \frac{1}{n^2} \sum_{i,j} W_{i,t} W_{j,t} h(x_i, x_j), \quad t = 1, \ldots, n. \tag{15}$$

We reject the null if the test statistic $\mathrm{dKSD}_b^2(p, q)$ in (13) exceeds the $(1 - \alpha)$ quantile of $S_1, \ldots, S_B$. The testing procedure is outlined in Algorithm 2.

Algorithm 2 dKSD test via wild bootstrap (dKSDv)

Input:
- samples $x_1, \ldots, x_n \sim p$
- null density q
- kernel function k
- test size α
- bootstrap sample size B

Objective: Test $H_0 : p = q$ versus $H_1 : p \neq q$.

Test procedure:
1. Compute the V-statistics $\mathrm{dKSD}_b^2(p, q)$ via (13).
2. **for** $t = 1 : B$ **do**
3. Sample $W_{1,t}, \ldots, W_{n,t}$ via (14).
4. Compute S_t by (15).
5. **end for**
6. Determine the $(1 - \alpha)$-quantile $\gamma_{1-\alpha}$ of $S_1, \ldots, S_B$.

Output:
Reject H_0 if $\mathrm{dKSD}_b^2(p, q) > \gamma_{1-\alpha}$; otherwise do not reject.

5.3 Kernel Choice

In kernel-based testing, the performance is sensitive to the choice of kernel parameters such as the bandwidth parameter in Gaussian kernels [Gretton et al., 2012]. For the proposed dKSD tests with the von Mises-Fisher kernel $k(x, x') = \exp(\kappa x^\top x')$, the choice of concentration parameter κ is crucial. Namely, if κ is too small, the test magnifies any small difference between observed samples, and gives high type-I error. On the other hand, if κ is too large, the test fails to detect the discrepancy between two different distributions. Previous works [Chwialkowski et al., 2016, Gretton et al., 2012, Jitkrittum et al., 2018, 2016, 2017] proposed to choose the kernel parameter by maximizing the test power, which is defined as the probability of rejecting H_0 when it is false. Here, we provide a method for choosing the kernel parameter by maximizing the test power of dKSDu.

We employ an approximation formula for the test power of dKSDu under $H_1 : p \neq q$. Since

$$D := \sqrt{n} \frac{\mathrm{dKSD}_u^2(p, q) - \mathrm{dKSD}^2(p, q)}{\sigma_u} \xrightarrow{d} \mathcal{N}(0, 1)$$

from Theorem 5, we have

$$\begin{aligned}
&\Pr_{H_1}(n \cdot \mathrm{dKSD}_u^2(p, q) > r) \\
=&\Pr_{H_1}\left(D > \frac{r}{\sqrt{n}\sigma_{H_1}} - \sqrt{n}\frac{\mathrm{dKSD}^2(p, q)}{\sigma_u}\right) \\
\approx&1 - \Phi\left(\frac{r}{\sqrt{n}\sigma_u} - \sqrt{n}\frac{\mathrm{dKSD}^2(p, q)}{\sigma_u}\right),
\end{aligned}$$

for large n and fixed r, where Φ denotes the cumulative distribution function of the standard Gaussian distribution and σ_u^2 is defined in Theorem 5. Following the argument in [Sutherland et al., 2016], we use the approximation

$$\frac{r}{\sqrt{n}\sigma_u} - \sqrt{n}\frac{\mathrm{dKSD}^2(p, q)}{\sigma_u} \approx -\sqrt{n}\frac{\mathrm{dKSD}^2(p, q)}{\sigma_u}.$$

Thus, to maximize the test power, we choose κ by

$$\kappa^* = \arg\max_{\kappa} \frac{\mathrm{dKSD}^2(p, q)}{\sigma_u}.$$

In practice, we use part of the data to calculate $\mathrm{dKSD}_u^2(p, q)/(\hat{\sigma}_u + \lambda)$, where $\hat{\sigma}_u$ is an unbiased estimate of σ_u and a regularization parameter $\lambda > 0$ is added for numerical stability. Then, we select κ^* by grid search and apply the dKSD tests to the rest of the data. In our experiments, this method had better testing performance than selecting the kernel parameter by the methods proposed in density estimation literature [García-Portugués et al., 2013b,a, Taylor, 2008].

5.4 Test with Maximum Mean Discrepancy

A proxy way to tackle the goodness-of-fit test on $\mathcal{S}^{d-1}$ is via the two-sample test with maximum mean discrepancy (MMD) [Gretton et al., 2007]. Namely, to test whether $x_1, \dots, x_n$ is from density q, we draw samples $y_1, \dots, y_m$ from q and determine whether $x_1, \dots, x_n$ and $y_1, \dots, y_m$ are from the same distribution. See [Gretton et al., 2007] for details. We compare the performance of the proposed dKSD tests with the MMD two-sample test in Section 6. Note that the MMD two-sample test requires to sample from the null distribution q, which can be computationally intensive for directional distributions especially in high dimension. On the other hand, the proposed dKSD tests do not need samples from the null.

6 EXPERIMENTAL RESULTS

Here, we validate the proposed dKSD tests by simulation. We employ the von Mises-Fisher kernel for both the dKSD tests and MMD two-sample test in Section 5.4. The bootstrap sample size is set to $B = 1000$. The significance level is set to $\alpha = 0.01$. In MMD two-sample test, we set $m = n$.

6.1 Circular Uniform Distribution

First, we consider the circular ($d = 2$) uniform distribution, for which several goodness-of-fit tests have been proposed such as Rayleigh test and Kuiper test [Mardia and Jupp, 1999]. See Supplementary Material for details of Rayleigh test and Kuiper test. We compare the proposed dKSD tests with these existing tests and MMD two-sample test. We repeated 600 trials to calculate rejection rates.

n	Rayleigh	Kuiper	dKSDu	dKSDv	MMD
30	0.006	0.010	0.011	0.007	0.013
50	0.015	0.011	0.015	0.015	0.016
100	0.010	0.011	0.008	0.011	0.030
200	0.015	0.018	0.010	0.015	0.013

Table 1: Type-I error of tests for the circular uniform distribution

n	Rayleigh	Kuiper	dKSDu	dKSDv	MMD
30	0.138	0.128	0.560	0.338	0.133
50	0.308	0.267	0.750	0.898	0.317
100	0.712	0.667	0.820	1.0	0.583
200	0.980	0.962	0.900	1.0	0.900

Table 2: Rejection rates for the circular uniform distribution under the von Mises distribution with $\kappa = 0.5$

n	Rayleigh	Kuiper	dKSDu	dKSDv	MMD
30	0.757	0.731	0.650	0.831	0.600
50	0.957	0.940	0.750	1.0	0.833
100	1.0	1.0	0.833	1.0	0.983
200	1.0	1.0	0.96	1.0	1.0

Table 3: Rejection rates for the circular uniform distribution under the von Mises distribution with $\kappa = 1$

Table 1 presents the rejection rate at the null. The type-I errors of all tests are well controlled to the significance level $\alpha = 0.01$. Tables 2 and 3 present the rejection rate under the von Mises distribution with concentration $\kappa = 0.5$ and $\kappa = 1$, respectively. The power of all tests increases with increasing n or κ and converges to one. The dKSDv has the largest power.

6.2 von Mises-Fisher Distribution

Next, we consider the von Mises-Fisher distribution vMF(μ, κ) in (1). We compare the proposed dKSD tests with MMD two-sample test. We repeated 200 trials to calculate rejection rates.

We set the null and alternative distributions to vMF$(\mu_0, 1)$ and vMF$(\mu, 1 + \sigma)$, respectively, where $\mu_0 = (1, 0, \dots, 0) \in \mathcal{S}^{d-1}$, $\mu \in \mathcal{S}^{d-1}$ and $\sigma \geq 0$. We generated samples from the von Mises-Fisher distribution by using the methods proposed in [Jakob, 2012].

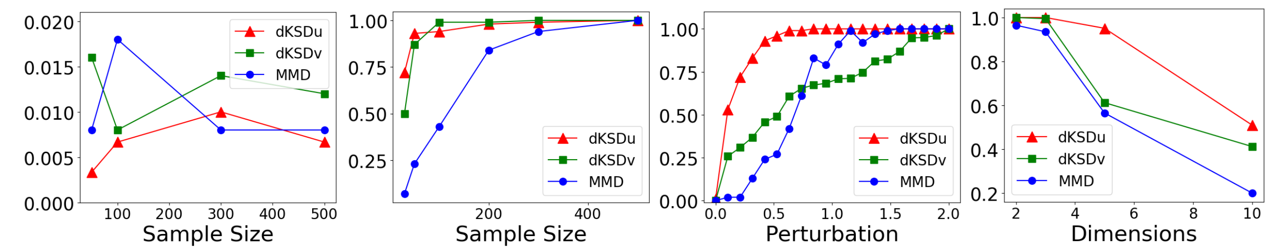
Figure 2(a) plots the rejection rate under the null ($\mu = \mu_0, \sigma = 0$) with respect to n for $d = 3$. The type-I errors of dKSD tests are well controlled to the significance level $\alpha = 0.01$. Figure 2(b) plots the rejection rate with respect to n for $d = 3$, $\mu = \mu_0$ and $\sigma = 1$. Both dKSDu and dKSDv have larger power than MMD two-sample test. Figure 2(c) plots the rejection rate with respect to σ for $d = 3$, $n = 200$ and $\mu = \mu_0$. The dKSDu has the largest power and achieves almost 100% power around $\kappa = 0.3$. Figure 2(d) plots the rejection rate with respect to d for $n = 200$, $\mu = (1/\sqrt{d})\mathbf{1_d}$ and $\sigma = 0.5$, where $\mathbf{1_d}$ denotes the all one vector. Although the test power decreases for higher dimension, dKSD tests have larger power than MMD two-sample test in all dimensions.

6.3 Fisher-Bingham Distribution

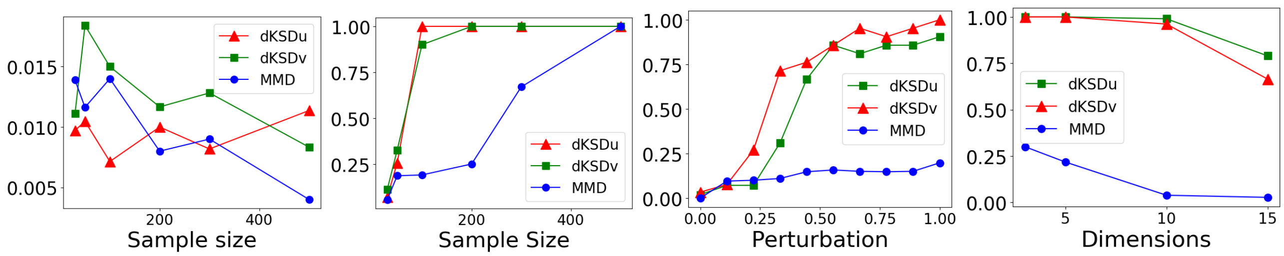
Finally, we consider the Fisher-Bingham distribution (2). Here, we focus on the *Bingham distribution*, FB(A): the distribution in the Fisher-Bingham family that only includes second order terms:

$$p(x \mid A) \propto \exp(x^\top Ax), \quad x \in \mathcal{S}^{d-1},$$

where $A \in \mathbb{R}^{d \times d}$ is symmetric. The normalization constant does not have closed form in general. We compare



(a) $d = 3$, $\mu = \mu_0$, $\sigma = 0$ (b) $d = 3$, $\mu = \mu_0$, $\sigma = 1$ (c) $d = 3$, $n = 200$, $\mu = \mu_0$ (d) $n = 200$, $\mu = \frac{\mathbf{1_d}}{\sqrt{d}}$, $\sigma = \frac{1}{2}$



(e) $d = 3$, $\sigma = 0$ (f) $d = 3$, $\sigma = 1$ (g) $d = 3$, $n = 200$ (h) $n = 200$, $\sigma = 1$

Figure 2: Rejection rates for (a)-(d) von Mises-Fisher and (e)-(h) Bingham distributions

the proposed dKSD tests with MMD two-sample test. We repeated 200 trials to calculate rejection rates.

We set the null distribution to FB(A) with

$$A_{ij} = \begin{cases} 2 & (i = j) \\ 1 & (i \neq j) \end{cases},$$

and the alternative distribution to FB(A') with $A' = A + \sigma \mathbf{1}_{d,d}$, where $\sigma \geq 0$ and $\mathbf{1}_{d,d}$ denotes the $d \times d$ matrix with all entries one. We generated samples from the Fisher-Bingham distribution via rejection sampling with angular central Gaussian proposals [Kent et al., 2013, Fallaize and Kypraios, 2016].

Figure 2(e) plots the rejection rate under the null ($\sigma = 0$) with respect to n for $d = 3$. The type-I errors of dKSD tests are approximately controlled to the significance level $\alpha = 0.01$. Figure 2(f) plots the rejection rate with respect to n for $d = 3$ and $\sigma = 1$. The dKSD tests have larger power and achieve almost 100% power around $n = 100$. Figure 2(g) plots the rejection rate with respect to σ for $n = 200$ and $d = 3$. Again, the dKSD tests have larger power and capture small perturbation. Figure 2(h) plots the rejection rate with respect to d for $n = 200$ and $\sigma = 1$. The dKSD tests attain almost 80% power even when the dimension is as large as 15, whereas the power of the MMD two-sample test is smaller than 20% for all dimensions.

Table 4 presents the computational time for $d = 3$. The dKSD tests are more computationally efficient than MMD test. The computational time of MMD test grows rapidly with the sample size n, because it requires to sample from the Bingham distribution.

n	dKSDu	dKSDv	MMD
30	0.005	0.009	0.091
50	0.011	0.015	0.120
100	0.027	0.030	0.180
200	0.096	0.105	0.379
300	0.227	0.238	0.704
500	0.588	0.574	2.614

Table 4: Computational time (in seconds) for Bingham.

7 CONCLUSION

In this study, we developed goodness-of-fit testing procedures for general directional distributions including unnormalized ones, based on an extension of Stein's operator and kernel Stein discrepancy. Experimental results demonstrated that the proposed methods control type-I errors well and attain larger power than existing tests, without sampling from the null distribution.

Although we focused on the unit hypersphere $\mathcal{S}^{d-1}$, our derivation of Stein's operator and kernel Stein discrepancy is applicable to general manifolds as well. It is an interesting future work to extend the proposed methods to general manifolds such as Stiefel manifolds and Grassmann manifolds

Acknowledgement

TM was supported by JSPS KAKENHI Grant Number 19K20220. WX was supported by Gatsby Charitable Foundation.

References

A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6:1345–1382, 2005.

Andrew D Barbour and Louis Hsiao Yun Chen. *An introduction to Stein's method*, volume 4. World Scientific, 2005.

Alessandro Barp, Chris Oates, Emilio Porcu, and Mark Girolami. A riemannian-stein kernel method. *arXiv preprint arXiv:1810.04946*, 2018.

Alain Berlinet and Christine Thomas. *Reproducing kernel Hilbert spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004.

Graciela Boente, Daniela Rodriguez, and Wenceslao González Manteiga. Goodness-of-fit test for directional data. *Scandinavian Journal of Statistics*, 41(1):259–275, 2014.

Claudio Carmeli, Ernesto De Vito, Alessandro Toigo, and Veronica Umanitá. Vector valued reproducing kernel hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61, 2010.

L. H. Y. Chen, L. Goldstein, and Q. M. Shao. *Normal approximation by Stein's method*. Springer, 2010.

Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *International Conference on Machine Learning*, pages 2606–2615, 2016.

Christopher J Fallaize and Theodore Kypraios. Exact bayesian inference for the bingham distribution. *Statistics and Computing*, 26(1-2):349–360, 2016.

Adelaide Figueiredo. Comparison of tests of uniformity defined on the hypersphere. *Statistics & probability letters*, 77(3):329–334, 2007.

Adelaide Maria Sousa Figueiredo. Goodness-of-fit for a concentrated von mises-fisher distribution. *Computational Statistics*, 27(1):69–82, 2012.

H. Flanders. *Differential Forms with Applications to the Physical Sciences*. Dover, 1963.

Eduardo García-Portugués and Thomas Verdebout. An overview of uniformity tests on the hypersphere. *arXiv preprint arXiv:1804.00286*, 2018.

Eduardo García-Portugués, Rosa M Crujeiras, and Wenceslao González-Manteiga. Kernel density estimation for directional–linear data. *Journal of Multivariate Analysis*, 121:152–175, 2013a.

Eduardo García-Portugués et al. Exact risk improvement of bandwidth selectors for kernel density estimation with directional data. *Electronic Journal of Statistics*, 7:1655–1685, 2013b.

Marc Genton and Amanda Hering. Blowing in the wind. *Significance*, 4(1):11–14, 2007.

Tilmann Gneiting et al. Strictly and non-strictly positive definite functions on spheres. *Bernoulli*, 19(4): 1327–1349, 2013.

Jackson Gorham and Lester Mackey. Measuring sample quality with stein's method. In *Advances in Neural Information Processing Systems*, pages 226–234, 2015.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pages 513–520, 2007.

Arthur Gretton, Kenji Fukumizu, Zaid Harchaoui, and Bharath K Sriperumbudur. A fast, consistent kernel two-sample test. In *Advances in neural information processing systems*, pages 673–681, 2009.

Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *Advances in neural information processing systems*, pages 1205–1213, 2012.

T. Hamelryck, J. T. Kent, and A. Krogh. Sampling realistic protein conformations using local structural bias. *PLoS Comput. Biol.*, 2:e131, 2006.

Amanda S Hering and Marc G Genton. Powering up with space-time wind forecasting. *Journal of the American Statistical Association*, 105(489):92–104, 2010.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(Apr):695–709, 2005.

Wenzel Jakob. Numerically stable sampling of the von mises-fisher distribution on s^2 (and other tricks). *Interactive Geometry Lab, ETH Zürich, Tech. Rep*, 2012.

Wittawat Jitkrittum, Zoltán Szabó, Kacper P Chwialkowski, and Arthur Gretton. Interpretable distribution features with maximum testing power. In *Advances in Neural Information Processing Systems*, pages 181–189, 2016.

Wittawat Jitkrittum, Wenkai Xu, Zoltán Szabó, Kenji Fukumizu, and Arthur Gretton. A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems*, pages 262–271, 2017.

Wittawat Jitkrittum, Heishiro Kanagawa, Patsorn Sangkloy, James Hays, Bernhard Schölkopf, and Arthur Gretton. Informative features for model comparison. In *Advances in Neural Information Processing Systems*, pages 808–819, 2018.

Heishiro Kanagawa, Wittawat Jitkrittum, Lester Mackey, Kenji Fukumizu, and Arthur Gretton. A kernel stein test for comparing latent variable models. *arXiv preprint arXiv:1907.00586*, 2019.

J. T. Kent. The fisher–bingham distribution on the sphere. *J. Royal. Stat. Soc. B*, 44:71–80, 1982.

John T Kent, Asaad M Ganeiber, and Kanti V Mardia. A new method to simulate the bingham and related distributions in directional data analysis with applications. *arXiv preprint arXiv:1310.8110*, 2013.

A. J. Lee. *U-Statistics: Theory and Practice*. CRC Press, 1990.

Christophe Ley and Thomas Verdebout. *Modern directional statistics*. Chapman and Hall/CRC, 2017.

Christophe Ley, Gesine Reinert, Yvik Swan, et al. Stein's method for comparison of univariate distributions. *Probability Surveys*, 14:1–52, 2017.

Chang Liu and Jun Zhu. Riemannian stein variational gradient descent for bayesian inference. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Qiang Liu, Jason Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning*, pages 276–284, 2016.

K. V. Mardia and P. E. Jupp. *Directional Statistics*. Wiley, New York, NY, 1999.

K. V. Mardia, D. Holmes, and J. Kent. A goodness-of-fit test for the von mises-fisher distribution. *J. Royal. Stat. Soc. B*, 46:72–78, 1984.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10 (1-2):1–141, 2017.

Michael Spivak. *Calculus on manifolds: a modern approach to classical theorems of advanced calculus*. CRC press, 2018.

Dougal J Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, and Arthur Gretton. Generative models and model criticism via optimized maximum mean discrepancy. *arXiv preprint arXiv:1611.04488*, 2016.

Charles C Taylor. Automatic bandwidth selection for circular density estimation. *Computational Statistics & Data Analysis*, 52(7):3493–3500, 2008.

Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: l_2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049. ACM, 2017.

Jiasen Yang, Qiang Liu, Vinayak Rao, and Jennifer Neville. Goodness-of-fit testing for discrete distributions via stein discrepancy. In *International Conference on Machine Learning*, pages 5557–5566, 2018.

Jiasen Yang, Vinayak Rao, and Jennifer Neville. A stein–papangelou goodness-of-fit test for point processes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 226–235, 2019.