Interpretable Stein Goodness-of-fit Tests on Riemannian Manifolds

Wenkai Xu [1] **Takeru Matsuda** [2]

Abstract

In many applications, we encounter data on Riemannian manifolds such as torus and rotation groups. Standard statistical procedures for multivariate data are not applicable to such data. In this study, we develop goodness-of-fit testing and interpretable model criticism methods for general distributions on Riemannian manifolds, including those with an intractable normalization constant. The proposed methods are based on extensions of kernel Stein discrepancy, which are derived from Stein operators on Riemannian manifolds. We discuss the connections between the proposed tests with existing ones and provide a theoretical analysis of their asymptotic Bahadur efficiency. Simulation results and real data applications show validity and usefulness of the proposed methods.

1. Introduction

In many scientific and machine learning applications, data appear in the domains described by Riemannian manifolds. For example, structures of proteins and molecules are described by a pair of angular variables, which is identified with a point on the torus (Singh et al., 2002). In computer vision and related studies, the orientation of a camera is represented by a 3×3 rotation matrix, which gives rise to data on the rotation group (Song et al., 2009). Other examples include the orbit of a comet (Jupp et al., 1979) and the vectorcardiogram data (Downs, 1972). In addition, shape analysis (Dryden & Mardia, 2016) and compositional data analysis (Pawlowsky-Glahn & Buccianti, 2011) also deal with complex data defined on Riemannian manifolds. Recently, Klein et al. (2020) developed a graphical model on torus to analyze phase coupling between neuronal activities. Since the usual statistical procedures for Euclidean data are not applicable in such scenarios, many studies have developed statistical models and methods tailored for data on Riemannian manifolds (Chikuse, 2012; Mardia & Jupp, 1999; Ley & Verdebout, 2017).

Statistical models on Riemannian manifolds are often given in the form of unnormalized densities with a computationally intractable normalization constant. For example, the Fisher distribution on the rotation group (Chikuse, 2012; Sei et al., 2013) is defined by

$$p(X \mid \Theta) \propto \exp(\mathrm{tr}(\Theta^\top X)), \tag{1}$$

and its normalization constant is not given in closed form. Statistical inference with such models can become computationally intensive due to the intractable normalization constant. Thus, statistical methods on Riemannian manifolds that do not require computation of the normalization constant have been developed for several tasks such as parameter estimation (Mardia et al., 2016) and sampling (Girolami et al., 2009; Ma et al., 2015). However, goodness-of-fit testing or model criticism procedures for general distributions on Riemannian manifolds is not established, to the best of our knowledge.

Kernel Stein discrepancy (KSD) (Gorham & Mackey, 2015; Ley et al., 2017) is a discrepancy measure between distributions based on Stein's method (Barbour & Chen, 2005; Chen et al., 2010) and reproducing kernel Hilbert space (RKHS) theory (Berlinet & Thomas, 2004). KSD provides a general procedure for goodness-of-fit testing that does not require computation of the normalization constant, and it has shown state-of-the-art performance in various scenarios including Euclidean data (Chwialkowski et al., 2016; Liu et al., 2016), discrete data (Yang et al., 2018), point processes (Yang et al., 2019), directional data (Xu & Matsuda, 2020), censored data (Fernandez et al., 2020) and random graphs (Xu & Reinert, 2021). In addition, by using the technique of optimizing test power (Gretton et al., 2012; Sutherland et al., 2016), KSD-based testing procedures also enable extraction of distributional features to perform model criticism (Jitkrittum et al., 2017; 2018; Kanagawa et al., 2019; Jitkrittum et al., 2020). We note that Stein's method has recently been extended to Riemannian manifolds and studied for numerical integration (Barp et al., 2018) and Bayesian inference (Liu & Zhu, 2018).

In this paper, we develop goodness-of-fit testing and interpretable model criticism methods for general distributions on Riemannian manifolds. After briefly reviewing

[1]Gatsby Computational Neuroscience Unit, London, United Kingdom [2]RIKEN Center for Brain Science, Tokyo, Japan. Correspondence to: Wenkai Xu <xwk4813@gmail.com>, Takeru Matsuda <takeru.matsuda@riken.jp>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

background topics, we first introduce several types of Stein operators on Riemannian manifolds by using Stokes' theorem. Then, we define manifold kernel Stein discrepancies (mKSD) based on them and propose goodness-of-fit testing procedures, which do not require computation of the normalization constant. We also develop mKSD-based interpretable model criticism procedures. Theoretical comparisons of test performance in terms of Bahadur efficiency are provided, and simulation results validate the claims. Finally, we provide real data applications to demonstrate the usefulness of the proposed methods.

2. Background

2.1. Distributions on Riemannian Manifolds

In this paper, we focus on distributions on a smooth Riemannian manifold $(\mathcal{M}, g)$, where g is a Riemannian metric on $\mathcal{M}$[1]. See Kobayashi & Nomizu (1963) for details on Riemannian geometry. Here, we give several examples that will be used in experiments. Note that we define the probability density of each distribution by its Radon–Nikodym derivative with respect to the volume element of $(\mathcal{M}, g)$.

Torus Bivariate circular data $(x_1, x_2) \in [0, 2\pi)^2$ can be viewed as data on the torus $\mathcal{S}_1 \times \mathcal{S}_1$, where we identify $(\cos x, \sin x) \in \mathcal{S}_1$ with $x \in [0, 2\pi)$. To describe dependence between circular variables, Singh et al. (2002) proposed the bivariate von-Mises distribution:

$$p(x_1, x_2 \mid \eta) \propto \exp(\kappa_1 \cos(x_1 - \mu_1) + \kappa_2 \cos(x_2 - \mu_2) + \lambda_{12} \sin(x_1 - \mu_1) \sin(x_2 - \mu_2)), \quad (2)$$

where $\eta = (\kappa_1, \kappa_2, \mu_1, \mu_2, \lambda_{12})$, $\kappa_1 \geq 0$, $\kappa_2 \geq 0$, $0 \leq \mu_1 < 2\pi$ and $0 \leq \mu_2 < 2\pi$. Its normalization constant is not represented in closed form. We will apply this model to wind direction data in Section 8.

Rotation group The rotation group $\mathrm{SO(m)}$ is defined as

$$\mathrm{SO(m)} = \{X \in \mathbb{R}^{m \times m} \mid X^\top X = I_m, \det X = 1\},$$

where I_m is the m-dimensional identity matrix. The Fisher distribution (Chikuse, 2012; Sei et al., 2013) on $\mathrm{SO(m)}$ is defined as

$$p(X \mid \Theta) \propto \exp(\mathrm{tr}(\Theta^\top X)),$$

for which the normalization constant is not given in closed form. We will apply this model to vectorcardiogram data in Section 8.

The goodness-of-fit testing for general distributions on Riemannian manifolds is not established, to the best of our knowledge. For tests of uniformity, several methods have been proposed such as the Sobolev test (Chikuse & Jupp, 2004; Giné, 1975; Jupp et al., 2008). However, they are not readily applicable to general disributions. Although there are a few methods applicable to general distributions (Jupp et al., 2005; Jupp & Kume, 2018), they require computation of the normalization constant, which is often computationally intensive. In addition, existing testing procedures cannot be applied to perform interpretable model criticism (Jitkrittum et al., 2016; Kim et al., 2016; Lloyd & Ghahramani, 2015), which would provide an intuitive clarification of the discrepancy between the model and data.

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

Here, we briefly review the goodness-of-fit testing with kernel Stein discrepancy on $\mathbb{R}^d$. 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 $\mathbf{f} = (f_1, \ldots, f_d) : \mathbb{R}^d \to \mathbb{R}^d$, the Stein operator $\mathcal{T}_q$ is defined by

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

From integration by parts on $\mathbb{R}^d$, we obtain the equality, i.e. the Stein's identity $\mathbb{E}_q[\mathcal{T}_q \mathbf{f}] = 0$, under mild regularity conditions. Since Stein 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 operator, kernel Stein discrepancy (KSD) (Gorham & Mackey, 2015; Ley et al., 2017) between two densities p and q is defined as $\mathrm{KSD}(p\|q) = \sup_{\|\mathbf{f}\|_{\mathcal{H}^d} \leq 1} \mathbb{E}_p[\mathcal{T}_q \mathbf{f}]$.

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

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

where $h_q(x, \tilde{x}) = \langle \mathcal{T}_q k(x, \cdot), \mathcal{T}_q k(\tilde{x}, \cdot) \rangle_{\mathcal{H}}$, which does not involve the density p.

Given samples $x_1, \ldots, x_n$ from *unknown* density p on $\mathbb{R}^d$, an empirical estimate of $\mathrm{KSD}^2(p\|q)$ can be obtained by using Eq.(4) in the form of U-statistics, and this estimate is used to test the hypothesis $H_0 : p = q$, where the critical value is determined by bootstrap. In this way, a general method of non-parametric goodness-of-fit test on $\mathbb{R}^d$ is obtained, which does not require computation of the normalization constant.

[1] In this paper, $\mathcal{M}$ may have non-empty boundary ∂M.

3. Stein Operators on $\mathcal{M}$

In this section, we introduce several types of Stein operators for distributions on Riemannian manifolds by using Stokes' theorem. The operators are categorized via the order of differentials of the input functions[2].

3.1. Differential Forms and Stokes' Theorem

To derive Stein operators on Riemannian manifolds, we need to use differential forms and Stokes' theorem. Here, we briefly introduce these concepts. For more detailed and rigorous treatments, see Flanders (1963); Spivak (2018).

Let $\mathcal{M}$ be a smooth d-dimensional Riemannian 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 $\mathcal{M}$. 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 $\mathcal{M}$, 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 volume element is defined as the d-form given by

$$(\det g)^{1/2} \mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^d,$$

where $g = g(x^1, \dots, x^d)$ is the $d \times d$ matrix of the Riemannian metric with respect to $x^1, \dots, x^d$.

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.

Proposition 1 (Stokes' theorem)**.** *Let $\partial\mathcal{M}$ be the boundary of $\mathcal{M}$ and ω be a $(d-1)$-form on $\mathcal{M}$. Then,*

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

Corollary 1. *If $\partial\mathcal{M}$ is empty, then $\int_{\mathcal{M}} \mathrm{d}\omega = 0$ for any $(d-1)$-form ω on $\mathcal{M}$.*

[2]Note that this should be distinguished from the differentials of the (unnormalized) density functions.

Coordinate choice In the following, to facilitate the derivation as well as computation of Stein operators, we assume that there exists a coordinate system $\theta^1, \dots, \theta^d$ on $\mathcal{M}$ that covers $\mathcal{M}$ almost everywhere. For example, spherical coordinates for the hyperspheres and torus, generalized Euler angles (Chikuse, 2012, Section 2.5.1) for the rotation groups, and Givens rotations (Pourzanjani et al., 2017) for the Stiefel manifolds satisfy this assumption.

3.2. First Order Stein Operator

For a smooth probability density q on $\mathcal{M}$ and a smooth function $\mathbf{f} = (f^1, \dots, f^d) : \mathcal{M} \to \mathbb{R}^d$, define a function $\mathcal{A}_q^{(1)}\mathbf{f} : \mathcal{M} \to \mathbb{R}$ by

$$\mathcal{A}_q^{(1)}\mathbf{f} = \sum_{i=1}^{d} \left(\frac{\partial f^i}{\partial \theta^i} + f^i \frac{\partial}{\partial \theta^i} \log(qJ) \right), \tag{5}$$

where $J = (\det g)^{1/2}$ is the volume element. We refer to $\mathcal{A}_q^{(1)}$ as the first order Stein operator. Note that Xu & Matsuda (2020) utilized this operator for goodness-of-fit testing on hyperspheres.

Theorem 1. *If $\partial\mathcal{M}$ is empty or $f^1, \dots, f^d$ vanish on ∂M, then*

$$\mathbb{E}_q[\mathcal{A}_q^{(1)}\mathbf{f}] = 0.$$

If $\mathcal{M}$ is a closed manifold such as torus and rotation group, it does not have boundary by definition and thus the assumption of Theorem 1 holds. If the boundary of $\mathcal{M}$ is non-empty, a discussion relevant to the assumption of Theorem 1 can be found in Liu & Kanamori (2019), which studies density estimation on truncated domains. Note that the assumption of Theorem 1 is similar to Assumption 4 in Barp et al. (2018).

3.3. Second Order Stein Operator

In the context of numerical integration on Riemannian manifolds, Barp et al. (2018) introduced a different type of Stein operator $\mathcal{A}_q^{(2)}$, which we call the second order Stein operator. Specifically, for a smooth probability density q on $\mathcal{M}$ and a smooth function $\tilde{f} : \mathcal{M} \to \mathbb{R}$, define $\mathcal{A}_q^{(2)}\tilde{f} : \mathcal{M} \to \mathbb{R}$ by

$$\mathcal{A}_q^{(2)}\tilde{f} = \sum_{ij} \left(g^{ij} \frac{\partial^2 \tilde{f}}{\partial \theta^i \partial \theta^j} + g^{ij} \frac{\partial \tilde{f}}{\partial \theta^j} \frac{\partial \log qJ}{\partial \theta^i} \right) \tag{6}$$

where we denote the inverse matrix of (g_{ij}) by (g^{ij}) following the convention of Riemmanian geometry.

Proposition 2 (Proposition 1 of Barp et al. (2018))**.** *If $\partial\mathcal{M}$ is empty or $\tilde{f}$ vanishes on ∂M, then*

$$\mathbb{E}_q[\mathcal{A}_q^{(2)}\tilde{f}] = 0.$$

Theorem 2 follows from Theorem 1, because the second order Stein operator in Eq.(6) can be viewed as a special case of the first order Stein operator in Eq.(5) with

$$f^i = \sum_j g^{ij} \frac{\partial \tilde{f}}{\partial \theta^j}. \tag{7}$$

Similar form of the second order Stein operator in Eq.(6) has been studied in Liu & Zhu (2018) for Bayesian inference. On the other hand, Le et al. (2020) arrives at a similar second order Stein operator on Riemannian manifolds via Feller diffusion process in the context of density approximation.

3.4. Zeroth Order Stein Operator

For a smooth probability density q on $\mathcal{M}$ and a function $h : \mathcal{M} \to \mathbb{R}$, define a function $\mathcal{A}_q^{(0)} h : \mathcal{M} \to \mathbb{R}$ by

$$\mathcal{A}_q^{(0)} h = h - \mathbb{E}_q[h], \tag{8}$$

which clearly satisfies $\mathbb{E}_q[\mathcal{A}_q^{(0)} h] = 0$. Since $\mathcal{A}_q^{(0)}$ does not involve any differentials, we call it the zeroth order Stein operator. Compared to the first and second order Stein operators, this operator requires the normalization constant of q, which is often computationally intractable for Riemannian manifolds. We will show later that this operator corresponds to the maximum mean discrepancy (MMD) (Gretton et al., 2007).

4. Goodness-of-fit Tests on $\mathcal{M}$

In this section, we propose goodness-of-fit testing procedures for distributions on Riemannian manifolds based on kernelized discrepancies using the Stein operators in the previous section.

4.1. Manifold Kernel Stein Discrepancies

By using Stein operators introduced in the previous section, we extend kernel Stein discrepancy to distributions on Riemannian manifolds.

Let $\mathcal{H}$ be a RKHS on $\mathcal{M}$ with reproducing kernel k and $\mathcal{H}^d$ be its product. We define the manifold kernel Stein discrepancies (mKSD) of the first, second and zeroth order by

$$\begin{aligned}
\mathrm{mKSD}^{(1)}(p\|q) &= \sup_{\|\mathbf{f}\|_{\mathcal{H}^d} \le 1} \mathbb{E}_p[\mathcal{A}_q^{(1)} \mathbf{f}], \\
\mathrm{mKSD}^{(2)}(p\|q) &= \sup_{\|\tilde{f}\|_{\mathcal{H}} \le 1} \mathbb{E}_p[\mathcal{A}_q^{(2)} \tilde{f}], \\
\mathrm{mKSD}^{(0)}(p\|q) &= \sup_{\|h\|_{\mathcal{H}} \le 1} \mathbb{E}_p[\mathcal{A}_q^{(0)} h],
\end{aligned}$$

respectively. We also define the Stein kernels of first, second and zeroth order by

$$\begin{aligned}
h_q^{(1)}(x, \tilde{x}) &= \left\langle \mathcal{A}_q^{(1)} k(x, \cdot), \mathcal{A}_q^{(1)} k(\tilde{x}, \cdot) \right\rangle_{\mathcal{H}^d}, \\
h_q^{(2)}(x, \tilde{x}) &= \left\langle \mathcal{A}_q^{(2)} k(x, \cdot), \mathcal{A}_q^{(2)} k(\tilde{x}, \cdot) \right\rangle_{\mathcal{H}}, \\
h_q^{(0)}(x, \tilde{x}) &= \left\langle \mathcal{A}_q^{(0)} k(x, \cdot), \mathcal{A}_q^{(0)} k(\tilde{x}, \cdot) \right\rangle_{\mathcal{H}},
\end{aligned}$$

respectively. Then, by algebraic manipulation, we obtain the following.

Theorem 2. *If p and q are smooth densities on $\mathcal{M}$ and the reproducing kernel k of $\mathcal{H}$ is smooth, then*

$$\mathrm{mKSD}^{(c)}(p\|q)^2 = \mathbb{E}_{x,\tilde{x}}[h_q^{(c)}(x, \tilde{x})] \tag{9}$$

for $c = 0, 1, 2$, where $x, \tilde{x} \sim p$ are independent.

From Theorem 2, we can estimate mKSD by using samples from p. This is an important property in goodness-of-fit testing.

The following theorem shows that mKSD is a proper discrepancy measure between distributions on Riemannian manifolds. The proof is given in supplementary material. Let $L(x) = (L_1(x), \ldots, L_d)^\top \in \mathbb{R}^d$ with

$$L_i(x) = \frac{\partial}{\partial \theta^i} \log \frac{q(x)}{p(x)}.$$

Theorem 3. *Let p and q be smooth densities on $\mathcal{M}$. Assume: 1) The kernel k vanishes at $\partial\mathcal{M}$ and is compact universal in the sense of* Carmeli et al. (2010, Definition 2 (ii))*; 2) $\mathbb{E}_{x,\tilde{x}\sim p}[h_q^{(c)}(x, \tilde{x})^2] < \infty$, for $c = 0, 1, 2$; 3) $\mathbb{E}_p \|L(x)\|^2 < \infty$. Then, $\mathrm{mKSD}^{(c)}(p\|q) \ge 0$ and $\mathrm{mKSD}^{(c)}(p\|q) = 0$ if and only if $p = q$.*

Note that different mKSD uses different RKHS as the space of test functions. With the d-dimensional vector valued RKHS $\mathcal{H}^d$, $\mathrm{mKSD}^{(1)}$ takes the supremum over a larger class of functions than $\mathrm{mKSD}^{(2)}$, capturing richer distribution features. Theoretical analysis in testing context will be presented in Section 6.

Equivalence of $\mathrm{mKSD}^{(0)}$ and MMD For a RKHS $\mathcal{H}$, the maximum mean discrepancy (MMD) (Gretton et al., 2007) between p and q is defined by

$$\mathrm{MMD}(p\|q)^2 = \|\mu_p - \mu_q\|_{\mathcal{H}}^2,$$

where μ_p, μ_q are the kernel mean embeddings (Muandet et al., 2017) of p and q, respectively. The following theorem shows that $\mathrm{mKSD}^{(0)}$ is equivalent to MMD.

Theorem 4.

$$\mathrm{mKSD}^{(0)}(p\|q) = \mathrm{MMD}(p\|q).$$

Proof. By definition, we have

$$\begin{aligned}\mathrm{mKSD}^{(0)}(p\|q) &= \sup_{\|h\|_{\mathcal{H}}\leq 1} \mathbb{E}_p[\mathcal{A}_q^{(0)}h] \\ &= \sup_{\|h\|_{\mathcal{H}}\leq 1} (\mathbb{E}_p[h] - \mathbb{E}_q[h]).\end{aligned}$$

Hence, taking the supreme in closed form via reproducing property, we obtain

$$\mathrm{mKSD}^{(0)}(p\|q)^2 = \|\mu_p - \mu_q\|_{\mathcal{H}}^2 = \mathrm{MMD}(p\|q)^2.$$

□

An illustrative example To see the differences between the Stein operators, we consider a uniform distribution q on 2-dimensional torus, using spherical coordinate (θ^1, θ^2)[3]. Thus, the first and second order Stein operators can be explicitly written as

$$\mathcal{A}_q^{(1)}\mathbf{f} = \frac{\partial f^1}{\partial \theta^1} + \frac{\partial f^2}{\partial \theta^2}, \quad \mathcal{A}_q^{(2)}\tilde{f} = \frac{\partial^2 \tilde{f}}{\partial \theta^1 \partial \theta^1} + \frac{\partial^2 \tilde{f}}{\partial \theta^2 \partial \theta^2},$$

respectively. This derivation echoes the interpretation of the difference between first order and second order via Eq.(7). To better understand the difference in terms of mKSD, we choose $f^1, f^2, \tilde{f} \in \mathcal{H}_1$ where $\mathcal{H}_1$ denotes the RKHS equipped with product von-Mises kernel of unit bandwidth,

$$k(u,v) = \exp\{u^\top v\} = \exp\{\cos(\theta_u^1 - \theta_v^1) + \cos(\theta_u^2 - \theta_v^2)\}$$

where θ_u, θ_v are θ-parametrisation of u, v respectively. Then, for $\theta_u, \theta_v \sim p$, mKSD^2 has explicit form,

$$\begin{aligned}\mathrm{mKSD}^{(1)}(p\|q)^2 &= \mathbb{E}_{\theta_u,\theta_v}\left[\frac{\partial^2 k(u,v)}{\partial \theta_u^1 \partial \theta_v^1} + \frac{\partial^2 k(u,v)}{\partial \theta_u^2 \partial \theta_v^2}\right] \\ &= \mathbb{E}_{\theta_u,\theta_v}\left[(\xi_1 + \xi_2)k(u,v)\right]\end{aligned}$$

where $\xi_1 = \cos(\theta_u^1 - \theta_v^1) - \sin(\theta_u^1 - \theta_v^1)^2$, and $\xi_2 = \cos(\theta_u^2 - \theta_v^2) - \sin(\theta_u^2 - \theta_v^2)^2$ can be seen as the statistics tracking differences of p, q in each $\mathcal{S}^1$. On the other hand,

$$\begin{aligned}&\mathrm{mKSD}^{(2)}(p\|q)^2 = \\ &\mathbb{E}_{\theta_u,\theta_v}\left[\frac{\partial^4 k(u,v)}{\partial {\theta_u^1}^2 \partial {\theta_v^1}^2} + \frac{\partial^4 k(u,v)}{\partial {\theta_u^2}^2 \partial {\theta_v^2}^2} + \frac{\partial^4 k(u,v)}{\partial {\theta_u^1}^2 \partial {\theta_v^2}^2} + \frac{\partial^4 k(u,v)}{\partial {\theta_u^2}^2 \partial {\theta_v^1}^2}\right] \\ &= \mathbb{E}_{\theta_u,\theta_v}\Big[(\xi_{11} + \xi_{22} + \xi_{12} + \xi_{21})k(u,v)\Big],\end{aligned}$$

where

$$\xi_{ii} = \sin(\theta_u^i - \theta_v^i)\left(\sin(\theta_u^i - \theta_v^i)^2 - 3\cos(\theta_u^i - \theta_v^i) - 1\right),$$

and $\xi_{ij} = \xi_{ii}\xi_{jj}$ for $(i \neq j)$.

[3]In this case, the Jacobian J is a constant as the torus $\mathcal{S}^1 \times \mathcal{S}^1$ is a direct product of two circles. Then, derivative of $\log qJ = 0$ By setting f^1, f^2 and $\tilde{f}$ all belonging to the same RKHS, $\mathcal{H}_1$, we are able to explicitly see an effect of having a larger space of test functions $\mathbf{f} = (f^1, f^2) \in \mathcal{H}_1 \times \mathcal{H}_1$, compared to $\tilde{f} \in \mathcal{H}_1$, through ξ-terms.

The zeroth order operator $\mathcal{A}_q^{(0)}$ does not involve any differential operator while density q is represented via expectation over the test function in Eq.(8). For mean embedding $\mu_q(\cdot) = \mathbb{E}_q[k(x,\cdot)]$, where $\mu_q(x) = \langle k(x,\cdot), \mu_q(\cdot)\rangle_{\mathcal{H}}$ and the constant $c_q = \|\mu_q\|_{\mathcal{H}}^2$,

$$\mathrm{mKSD}^{(0)}(p\|q)^2 = \mathbb{E}_{\theta_u,\theta_v}\left[k(u,v)\right] - 2\mathbb{E}_{\theta_u}\left[\mu_q(u)\right] + c_q.$$

From the derivation[4] for mKSD$^{(0)}$, we see that $k(u,v)$ does not interact with any of the ξ-terms as opposed to differential based Stein discrepancies. Instead, the characterisation is via balancing constant c_q that is representative for density q.

Connections to Euclidean Stein operator One interesting interpretation on the differences between proposed Stein operators $\mathcal{A}_q^{(1)}, \mathcal{A}_q^{(2)}$ and the Euclidean Stein operator in Eq.(3) is that the proposed operators diffuse *along* the shape/surface of the manifold, while the Stein operator in $\mathbb{R}^d$ diffuse to all directions w.r.t. the Cartesian coordinate[5]. For instance, in a 2-dimensional torus that is embedded in $\mathbb{R}^3$, the proposed Stein operator only diffuse on the surface of the torus, not going into or out from the torus; however, an Euclidean Stein operator does so due to diffusion over $\mathbb{R}^3$. This is also crucial for the unnormalised models. The same unnormalised density corresponds to different models when domains are not the same, e.g. unit sphere and sphere with radius 2. Essentially, the *diffusion along manifold* notion makes the Stein's identity holds and gives controlled type-I error for our tests.

Although $\mathcal{A}_q^{(0)}$ was not discussed explicitly in the Euclidean setting, it remains the same for both Euclidean and Riemannian cases. The manifold shapes and structures are well-taken care of by taking the expectation over relevant distributions in Eq.(8), e.g. in the above example, the expectation is taken over the 2-dimensional torus instead of $\mathbb{R}^3$. As such, the relevant kernel Stein test gives controlled type-I error. However, it is worth to note that the advantage on dealing with unnormalized densities can be violated when computing expectations in closed forms. When samples are available from unnormalized densities, e.g. via MCMC, an estimation for the expectation function can be obtained. By Theorem 4, Stein test for goodness-of-fit degenerates into MMD test for two-sample problems. We further discuss such connections and empirical results in Section 6 and 7.

[4]In this case, the derivation holds for both scalar valued test function $h \in \mathcal{H}$ or $(h_1, h_2) \in \mathcal{H} \times \mathcal{H}$.

[5]The Cartesian coordinate forms the basis directions to guide the diffusion in $\mathbb{R}^d$. Hence, we can see Eq.(3) as a special case of $\mathcal{A}_q^{(1)}$ on Euclidean manifold.

4.2. Goodness-of-fit Testing with mKSDs

Here, we present procedures for testing $H_0 : p = q$ with significance level α based on samples $x_1, \dots, x_n \sim p$.

From Theorem 2, an unbiased estimate of mKSD can be obtained in the form of U-statistics (Lee, 1990):

$$\mathrm{mKSD}_u^{(c)}(p\|q)^2 = \frac{1}{n(n-1)} \sum_{i \neq j} h_q^{(c)}(x_i, x_j). \tag{10}$$

Its asymptotic distribution is obtained via U-statistics theory (Lee, 1990; Van der Vaart, 2000) as follows. We denote the convergence in distribution by $\xrightarrow{d}$.

Theorem 5. *For $c = 0, 1, 2$, the following statements hold.*

1. Under $H_0 : p = q$,

$$n \cdot \mathrm{mKSD}_u^{(c)}(p\|q)^2 \xrightarrow{d} \sum_{j=1}^{\infty} w_j^{(c)}(Z_j^2 - 1), \tag{11}$$

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

$$\int h_q^{(c)}(x, \tilde{x}) \phi_j(\tilde{x}) p(\tilde{x}) \mathrm{d}\tilde{x} = w_j^{(c)} \phi_j(x), \tag{12}$$

where $\phi_j(x) \neq 0$ is the non-trivial eigen-function.

2. Under $H_1 : p \neq q$,

$$\sqrt{n} \cdot \left(\mathrm{mKSD}_u^{(c)}(p\|q)^2 - \mathrm{mKSD}^{(c)}(p\|q)^2 \right) \xrightarrow{d} \mathcal{N}(0, {\sigma_c}^2),$$

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

Based on Theorem 5, we propose two procedures for goodness-of-fit testing.

Wild-bootstrap Test We employ the wild-bootstrap test with the V-statistics (Chwialkowski et al., 2014). The test statisic is given by

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

To approximate its null distribution, we define the wild-bootstrap samples by

$$S_t = \frac{1}{n^2} \sum_{i,j} W_{i,t} W_{j,t} h_q^{(c)}(x_i, x_j), \tag{14}$$

where each $W_{i,t} \in \{-1, 1\}$ is the Rademacher variable of zero mean and unit variance.

The testing procedure is outlined in Algorithm 1. We adopt this algorithm in the following experiments due to its computational efficiency.

Algorithm 1 mKSD test via wild-bootstrap

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 statistic $\mathrm{mKSD}_v{}^{(c)}(p\|q)^2$, Eq.(10).
2: **for** $t = 1 : B$ **do**
3: Sample Rademacher variables $W_{1,t}, \dots, W_{n,t}$.
4: Compute S_t by Eq.(14).
5: **end for**
6: Determine the $(1-\alpha)$-quantile $\gamma_{1-\alpha}$ of $S_1, \dots, S_B$.
Output:
Reject H_0 if $\mathrm{mKSD}_v{}^{(c)}(p\|q)^2 > \gamma_{1-\alpha}$; otherwise do not reject H_0.

Spectrum Test We can also directly approximate the null distribution in Eq.(11) by using the eigenvalues of the Stein kernel matrix (Gretton et al., 2009, Theorem 1). Specifically, let $M^{(c)}$ be the $n \times n$ Stein kernel matrix defined by $(M^{(c)})_{ij} = h_q^{(c)}(x_i, x_j)$ and $\widetilde{w}_1^{(c)}, \dots, \widetilde{w}_n^{(c)}$ be its eigenvalues. Then, we generate the simulated null samples by

$$S_t = \frac{1}{n} \sum_{j=1}^{n} \widetilde{w}_j^{(c)}(Z_{j,t}^2 - 1), \tag{15}$$

where each $Z_{j,t}$ is the standard Gaussian variable. In practice, the spectrum test is more useful when sample size is small where the wild-bootstrap procedure can be less accurate. However, when sample size n is large, computing $\widetilde{w}^{(c)}$ via eigenvalue decomposition requires $O(n^3)$ complexity, which makes the test computationally less efficient.

Kernel choice The performance of kernel-based testing is sensitive to the choice of kernel parameters. We choose the kernel parameters by maximizing an approximation of the test power following Gretton et al. (2012); Jitkrittum et al. (2016); Sutherland et al. (2016). From Theorem 5,

$$D := \sqrt{n} \cdot \frac{\mathrm{mKSD}_u^{(c)}(p\|q)^2 - \mathrm{mKSD}^{(c)}(p\|q)^2}{\sigma_c} \xrightarrow{d} \mathcal{N}(0, 1)$$

under the alternative hypothesis, $H_1 : p \neq q$. Thus, for sufficiently large n, the test power is approximated as

$$\mathbb{P}_{H_1}(n \cdot \mathrm{mKSD}_u^{(c)}(p\|q)^2 > r) \approx \Phi\left(\sqrt{n} \cdot \frac{\mathrm{mKSD}^{(c)}(p\|q)^2}{\sigma_c}\right).$$

where Φ denotes the c.d.f. for the standard Gaussian distribution (Sutherland et al., 2016). Thus, we choose the kernel parameters by maximizing an estimate of $\mathrm{mKSD}^2(p\|q)/\sigma_c$ (Jitkrittum et al., 2017).

5. Model Criticism on $\mathcal{M}$

Now, we propose mKSD-based model criticism procedures for distributions on Riemannian manifolds.

When the proposed model does not fit the observed data well, understanding which part of the model misfit the data is of practical interest. The model criticism study can be helpful to better understand the representative prototype (Kim et al., 2016), to criticize prior assumptions in Bayesian settings (Lloyd & Ghahramani, 2015) or to help better training of generative models (Sutherland et al., 2016). With kernel-based non-parametric testing, distributional features can be extracted in the form of *test locations* to represent areas that *"best distinguish"* distributions. The locations where two sample distributions differ the most via MMD are studied in Jitkrittum et al. (2016) and the most "misspecified" locations between samples and models via KSD are studied in Jitkrittum et al. (2017). Recently, Seth et al. (2019) studied the model criticism via latent space, which may intrinsically correspond to Riemannian manifold structures. Such setting can be an interesting application of our development.

Let $\mathbf{s}_p(\cdot) = \mathbb{E}_{\tilde{x}\sim p}[\mathcal{A}_q^{(1)} k(\tilde{x}, \cdot)] \in \mathcal{H}^d$. Adapted from Jitkrittum et al. (2017), we define the manifold Finite Set Stein Discrepancy (mFSSD) as follows. For a small set of J test locations $\{v_1, \dots, v_j\} \in \mathcal{M}$,

$$\mathrm{mFSSD}(p\|q)^2 = \frac{1}{dJ} \sum_{i=1}^{d} \sum_{j=1}^{J} (\mathbf{s}_p(v_j))_i^2, \qquad (16)$$

which can be computed in linear time of sample size n.

Proposition 3 (Theorem 1 Jitkrittum et al. (2017))**.** *Let $V = v_1, \dots, v_J \in \mathcal{M}$ be random vectors drawn i.i.d. from a distribution ν which has a density. Let X be a connected open set in $\mathbb{R}^d$. Assume conditions in Theorem 3 hold. Then, for any $J \geq 1$,ν-almost surely $\mathrm{mFSSD}(p\|q)^2 = 0$ if and only if $p = q$.*

Stein identity of $\mathbf{s}_p(\cdot)$ ensures $\mathrm{mFSSD}^2 = 0$ under H_0 almost surely. To perform model criticism, we extract test locations that give a higher detection rate (i.e., test power) than others. We choose the test locations $V = \{v_j\}_{j=1}^J$ by maximizing the approximate test power:

$$V = \arg\max_{\mathbf{v}} \frac{\mathrm{mFSSD}^2}{\tilde{\sigma}_{H_1}}, \qquad (17)$$

where $\tilde{\sigma}_{H_1}$ is the variance of mFSSD^2 under H_1. More details are shown in Proposition 4 and 5 in the supplementary.

6. Comparison between mKSD Tests

Bahadur efficiency From Theorem 5, mKSD tests are consistent against all alternatives. Thus, to understand which mKSD test is more powerful than others, we investigated their *Bahadur efficiency* (Bahadur et al., 1960), which quantify how fast the p-value goes to zero under alternatives. Here, to focus on the effect of the choice of Stein operator on test performance, we briefly present results for testing of uniformity on the circle $\mathcal{S}^1$ under the von-Mises distribution. See supplementary material for more details. The technique of the proof is adapted from Jitkrittum et al. (2017).

Theorem 6. *(Scaling shift in von-Mises distribution) Let $x \in \mathcal{S}^1$, $q(x) \propto 1$ and $p(x) \propto \exp(\kappa u^\top x)$. Choose the von-Mises kernel of the form $k(x, x') = \exp(x^\top x')$. Denote the approximate Bahadur efficiency between* mKSD *with first and second order Stein operators as*

$$\mathrm{E}_{1,2}(\kappa) := \frac{c^{(\mathrm{mKSD}^{(1)})}(\kappa)}{c^{(\mathrm{mKSD}^{(2)})}(\kappa)},$$

where $\kappa > 0$. Then $\mathrm{E}_{1,2}(\kappa) > 1$.

Adapting from Theorem 5 of Jitkrittum et al. (2017), it suffices to show $\mathrm{mKSD}^{(1)}(p\|q) \geq \mathrm{mKSD}^{(2)}(p\|q)$ and

$$\mathbb{E}_{x,\tilde{x}\sim q}[h_q^{(2)}(x, \tilde{x})^2] > \mathbb{E}_{x,\tilde{x}\sim q}[h_q^{(1)}(x, \tilde{x})^2] > 0.$$

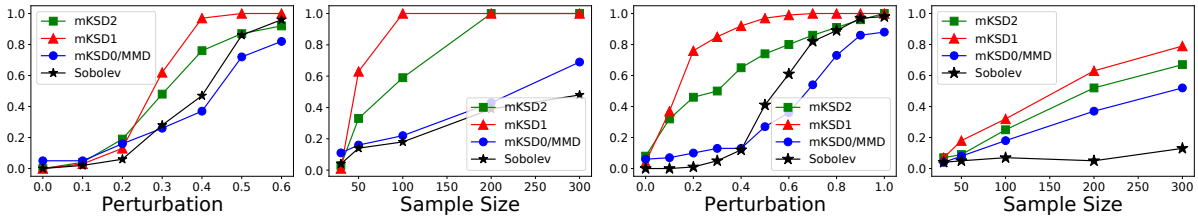
See supplementary material for details.

We provide additional discussion on test efficiencies with $\mathrm{mKSD}^{(0)}$ in the supplementary material. In general, since we cannot compute $\mathbb{E}_p$ in closed form, especially with unnormlized density, we need to perform the test with samples, where sampling error makes the $\mathrm{mKSD}^{(0)}$ test less asymptotically efficient (Jitkrittum et al., 2017; Yang et al., 2019; Xu & Matsuda, 2020).

Computational efficiency Since the Stein kernels $h_q^{(1)}$ and $h_q^{(2)}$ depend on q only through the derivative of $\log q$, mKSD tests with the first and second order Stein operators do not require computation of the normalization constant of q. This is a major computational advantage over existing goodness-of-fit tests on Riemannian manifolds. While the computational cost of $\mathrm{mKSD}_u^{(1)}$ is $O(n^2 d)$, that of $\mathrm{mKSD}^{(2)}$ is $O(n^2 d^3)$ due to the computation of the metric tensor.

On the other hand, mKSD test of zeroth order is equivalent to testing whether two sets of samples are from the same distribution by using 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. This procudre requires to sample from the null distribution q on Riemannian manifolds, which is computationally intensive in general. Note that the results in Theorem 5 with $c = 0$ replicate the asymptotic results for MMD (Gretton et al., 2007).

Choosing mKSD tests In overall, testing with $\mathrm{mKSD}^{(1)}$ has its advantage in terms of having a larger space of test



(a) $n = 100$ (b) $\kappa = 0.35$ (c) $n = 100$ (d) $b = 0.20$

Figure 1. Rejection rates at $\alpha = 0.01$: a)-b) for uniform density; c)-d) for Fisher distribution on SO(3)

functions with both asymptotic test efficiency and computational efficiency so that it is recommended to use when available. $\mathrm{mKSD}^{(2)}$ can be slightly easier to compute and parameterize in certain scenarios, although it may sacrifice test power and computational efficiency. $\mathrm{mKSD}^{(0)}$, or namely MMD test from simulated samples, is also applicable when it is possible to sample from the given unnormalized density model on the Riemannian manifolds.

7. Simulation Results

In this section, we show the validity of the proposed mKSD tests by simulation on the rotation group $\mathrm{SO}(3)$. We use the Euler angle (Chikuse, 2012) as the coordinate system. The bootstrap sample size is set to $B = 1000$. The significance level is set to $\alpha = 0.01$. For the $\mathrm{mKSD}^{(0)}$ test (MMD two-sample test), the number of samples from the null is set to be equal to the sample size n. We used the kernel $k(X, Y) = \exp(\gamma \cdot \mathrm{tr}(X^\top Y))$, where the parameter γ was chosen by optimizing the approximate test power. The exponential-trace kernel $k(X, Y) = \exp(\gamma \cdot \mathrm{tr}(X^\top Y))$ for the rotation group is compact universal. To see this, we rewrite the kernel in the form analogous to the Gaussian kernel: $k(X, Y) = \exp(\gamma \cdot \mathrm{tr}(X^\top Y)) = C \cdot \exp(-\frac{1}{2}\gamma \cdot \|X - Y\|_F^2)$, where C is a constant that only depends on d, the dimension of the matrices $X, Y \in \mathrm{SO(d)}$ due to $\mathrm{tr}(X^\top X) = \mathrm{tr}(I_d) = d$ for all $X \in \mathrm{SO(d)}$. Since the Gaussian kernel is universal (Sriperumbudur et al., 2011) and the rotation group $\mathrm{SO(d)}$ is a compact subset of the space of $d \times d$ matrices, the exponential-trace kernel is then also compact-universal from Corollary 3 of Carmeli et al. (2010).

7.1. Uniform distribution

First, we consider testing of uniformity on $\mathrm{SO}(3)$ and compare the performance of the mKSD tests with the Sobolev test (Jupp et al., 2005). We generated samples from the exponential trace distribution $p(X \mid \kappa) \propto \exp(\kappa \cdot \mathrm{tr}(X))$ by the rejection sampling (Hoff, 2009). The uniform distribution corresponds to $\kappa = 0$.

Figure 1 (a) plots the rejection rates with respect to κ for $n = 100$. When $\kappa = 0$, the type-I errors of all tests are well controlled to the significance level $\alpha = 0.01$. The power of all tests increases with increasing κ and converges to one. Figure 1 (b) plots the rejection rates with respect to n for $\kappa = 0.35$. The power of all tests increases with n and converges to one. When the model becomes increasingly different from the null, the mKSD1 is more sensitive to distinguish the difference, with higher power than others.

7.2. Fisher distribution

Next, we consider the Fisher distribution (or matrix-Langevin distribution) $p(X \mid F) \propto \exp(\mathrm{tr}(F^\top X))$ (Chikuse, 2003; Sei et al., 2013). We generated data from $p(X \mid F_0)$ and applied mKSD tests on the null $p(X \mid F_b)$, where $F_b = \begin{pmatrix} 1 & b & 0 \\ b & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$. We compare the mKSD tests with the extended Sobolev test (Jupp et al., 2005), in which we compute the normalization constant by Monte Carlo.

Figure 1 (c) plots the rejection rates with respect to b for $n = 100$. Figure 1 (d) plots the rejection rates with respect to n for $b = 0.2$. From the plot, we see that all tests achieves the correct test level under the null. When the model becomes increasingly different from the null, the mKSD1 is more sensitive to distinguish the difference, with higher power than others. MMD test has lower power than mKSD1 and mKSD2 due to inefficiency from sampling. While the Sobolev test is useful when the null and the alternative are very different, it is not powerful for harder problems where the alternative perturbed very little from the null.

8. Real Data Applications

Finally, we apply the mKSD tests to two real data.

8.1. Vectorcardiogram data

As a real dataset on the rotation group $\mathrm{SO}(3)$, we use the vectorcardiogram data studied by Jupp et al. (2008). The data summarizes vectorcardiogram from normal children where each data point records 3 perpendicular vectors of directions QRS, PRS and T from Frank system for electrical lead placement. Details of this dataset can be found in Downs (1972).

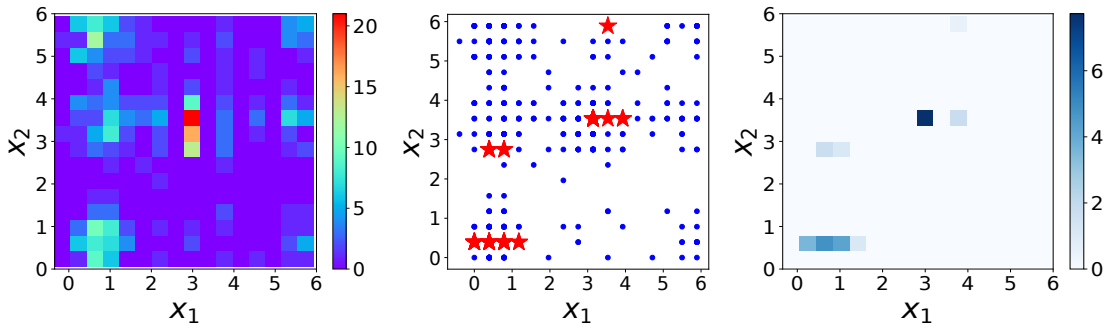


Figure 2. Wind direction data. Left: 2D histogram for wind directions; colorbar shows the counts of data points in each square. Mid: the 10 optimized locations (in red star), without repetition. Right: the objective value in Eq.(17), $\frac{\mathrm{mFSSD}^2}{\tilde{\sigma}_{H_1}}$, on the specified data location of test (i.e. setting J=1); the higher the darker.

We fitted a Fisher distribution $p(X \mid F) \propto \exp(\mathrm{tr}(F^\top X))$ to 28 data points of children aged between 2 to 10 (Jupp et al., 2005) and obtained the estimate

$$\hat{F} = 5.63 \times \begin{pmatrix} 0.583 & 0.629 & 0.514 \\ 0.660 & -0.736 & 0.151 \\ 0.473 & 0.252 & -0.844 \end{pmatrix}.$$

We use this value as the null model to be tested. Table 1 presents the p-values of each test. We apply kernel of the form $k(X, Y) = \exp(\gamma \cdot \mathrm{tr}(X^\top Y))$ as used in Section 7. All mKSD tests show strong evidence to reject the fitted model at $\alpha = 0.05$. However, the Sobolev test, with p-value being 0.126, is not powerful enough to reject the null at the same test level.

Table 1. p-values for vectorcardiogram data.

mKSD1	mKSD2	mKSD0/MMD	Sobolev
0.004	0.000	0.010	0.126

8.2. Wind direction data

As a real data on torus, we consider wind direction in Tokyo on 00:00 (x_1) and 12:00 (x_2) for each day in 2018[6]. Thus, the sample size is $n = 365$. The data were discretized into 16 directions, such as north-northeast. Figure 2 presents a 16×16 histogram of raw data.

We consider the goodness-of-fit testing procedures for the bivariate von-Mises distribution in Eq.(2) using mKSD. We apply the product von-Mises kernel as described in the illustrative example for torus in Section 4, $k(u, v) = \exp\{\gamma_1 \cos(\theta_u^1 - \theta_v^1) + \gamma_2 \cos(\theta_u^2 - \theta_v^2)\}$, where the bandwidth parameters γ_1 and γ_2 were chosen by optimizing the approximate test power[7]. In addition, we pay particular attention on positive definiteness when choosing our kernel. As Feragen et al. (2015) pointed out, not all geodesic distance, $d(u, v)$, induces a positive-definite kernel $k(u, v) = \exp(-\gamma \cdot d(u, v))$ on manifold. Adapting results on strictly positive functions in hyperspheres (Gneiting et al., 2013), the above chosen kernel is positive definite for torus as the product spherical coordinate on $\mathcal{S}^1$. By using noise contrastive estimation (Gutmann & Hyvärinen, 2012), Uehara et al. (2020) fitted the bivariate von-Mises distribution to the wind direction data and obtained the estimate for parameter set $\eta = (\kappa_1, \kappa_2, \mu_1, \mu_2, \lambda_{12})$,

$$\widehat{\eta} = (0.7170, 0.3954, 1.1499, 1.1499, -1.1274).$$

By setting this fitted model to the null model, the p-value obtained using mKSD1 is 0.434, which indicates that the model is a good fit for the data.

In addition, we fitted a simpler model with no interactions between x_1 and x_2, i.e. λ_{12} is set to zero in Eq.(2) so that the model reduces to the product of two von-Mises distribution on each direction. The p-value by mKSD1 is 0.002, which is a strong evidence to reject the null model. In other words, there is a significant interaction between wind direction on 00:00 and 12:00. We then carried out model criticism by mFFSD statistic in Eq.(16) with optimized test location via maximizing approximate test power. Choosing the number of test locations $J = 10$, we plot the optimized locations in Figure 2. It provides information about dependence between wind direction at midnight and noon.

Concluding Remark In this study, we develop goodness-of-fit procedures and model criticism methods for general distributions on Riemannian manifolds. As mKSDs are proper discrepancy measures under mild assumptions, the connections and comparisons of topologies induced from different mKSDs are interesting future direction.

Acknowledgement

T.M. was supported by JSPS KAKENHI Grant Number 19K20220 and JST Moonshot Grant Number JPMJMS2024. W.X. was supported by Gatsby Charitable Foundation.

[6]Data available on Japan Meteorological Agency website https://www.data.jma.go.jp/obd/stats/etrn/.

[7]The optimization objective is, $\frac{\mathrm{mKSD}^2}{\mathrm{Var}(\mathrm{mKSD}^2)}$, which is similar to Eq.(17) where the test statistics is mKSD^2 instead of mFSSD^2.

References

Bahadur, R. R. et al. Stochastic comparison of tests. *Annals of Mathematical Statistics*, 31(2):276–295, 1960.

Barbour, A. D. and Chen, L. H. Y. *An introduction to Stein's method*, volume 4. World Scientific, 2005.

Barp, A., Oates, C., Porcu, E., and Girolami, M. A riemannian-stein kernel method. *arXiv preprint arXiv:1810.04946*, 2018.

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

Carmeli, C., De Vito, E., Toigo, A., and Umanitá, V. Vector valued reproducing kernel hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61, 2010.

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

Chikuse, Y. Concentrated matrix langevin distributions. *Journal of Multivariate Analysis*, 2(85):375–394, 2003.

Chikuse, Y. *Statistics on special manifolds*, volume 174. Springer Science & Business Media, 2012.

Chikuse, Y. and Jupp, P. E. A test of uniformity on shape spaces. *Journal of multivariate analysis*, 88(1):163–176, 2004.

Chwialkowski, K., Strathmann, H., and Gretton, A. A kernel test of goodness of fit. In *International Conference on Machine Learning*, pp. 2606–2615, 2016.

Chwialkowski, K. P., Sejdinovic, D., and Gretton, A. A wild bootstrap for degenerate kernel tests. In *Advances in neural information processing systems*, pp. 3608–3616, 2014.

Downs, T. D. Orientation statistics. *Biometrika*, 59(3): 665–676, 1972.

Dryden, I. L. and Mardia, K. V. *Statistical shape analysis: with applications in R*, volume 995. John Wiley & Sons, 2016.

Feragen, A., Lauze, F., and Hauberg, S. Geodesic exponential kernels: When curvature and linearity conflict. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3032–3042, 2015.

Fernandez, T., Rivera, N., Xu, W., and Gretton, A. Kernelized stein discrepancy tests of goodness-of-fit for time-to-event data. In *International Conference on Machine Learning*, pp. 3112–3122. PMLR, 2020.

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

Garreau, D., Jitkrittum, W., and Kanagawa, M. Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*, 2017.

Giné, E. Invariant tests for uniformity on compact riemannian manifolds based on sobolev norms. *The Annals of statistics*, pp. 1243–1266, 1975.

Girolami, M., Calderhead, B., and Chin, S. A. Riemannian manifold hamiltonian monte carlo. *arXiv preprint arXiv:0907.1100*, 2009.

Gleser, L. J. The comparison of multivariate tests of hypothesis by means of bahadur efficiency. *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 157–174, 1966.

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

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

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. J. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pp. 513–520, 2007.

Gretton, A., Fukumizu, K., Harchaoui, Z., and Sriperumbudur, B. K. A fast, consistent kernel two-sample test. In *Advances in neural information processing systems*, pp. 673–681, 2009.

Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., and Sriperumbudur, B. K. Optimal kernel choice for large-scale two-sample tests. In *Advances in neural information processing systems*, pp. 1205–1213, 2012.

Gutmann, M. U. and Hyvärinen, A. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13:307–361, 2012.

Hoff, P. D. Simulation of the matrix bingham–von mises–fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, 18(2):438–456, 2009.

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

Jitkrittum, W., Szabó, Z., Chwialkowski, K. P., and Gretton, A. Interpretable distribution features with maximum testing power. In *Advances in Neural Information Processing Systems*, pp. 181–189, 2016.

Jitkrittum, W., Xu, W., Szabó, Z., Fukumizu, K., and Gretton, A. A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems*, pp. 262–271, 2017.

Jitkrittum, W., Kanagawa, H., Sangkloy, P., Hays, J., Schölkopf, B., and Gretton, A. Informative features for model comparison. In *Advances in Neural Information Processing Systems*, pp. 808–819, 2018.

Jitkrittum, W., Kanagawa, H., and Schölkopf, B. Testing goodness of fit of conditional density models with kernels. *arXiv preprint arXiv:2002.10271*, 2020.

Jupp, P. and Kume, A. Measures of goodness of fit obtained by canonical transformations on riemannian manifolds. *arXiv preprint arXiv:1811.04866*, 2018.

Jupp, P. et al. Sobolev tests of goodness of fit of distributions on compact riemannian manifolds. *The Annals of Statistics*, 33(6):2957–2966, 2005.

Jupp, P. et al. Data-driven sobolev tests of uniformity on compact riemannian manifolds. *The Annals of Statistics*, 36(3):1246–1260, 2008.

Jupp, P. E., Mardia, K. V., et al. Maximum likelihood estimators for the matrix von mises-fisher and bingham distributions. *The Annals of Statistics*, 7(3):599–606, 1979.

Kanagawa, H., Jitkrittum, W., Mackey, L., Fukumizu, K., and Gretton, A. A kernel stein test for comparing latent variable models. *arXiv preprint arXiv:1907.00586*, 2019.

Kim, B., Khanna, R., and Koyejo, O. Examples are not enough, learn to criticize! criticism for interpretability. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 2288–2296, 2016.

Klein, N., Orellana, J., Brincat, S. L., Miller, E. K., Kass, R. E., et al. Torus graphs for multivariate phase coupling analysis. *Annals of Applied Statistics*, 14(2):635–660, 2020.

Kobayashi, S. and Nomizu, K. *Foundations of differential geometry*, volume 1. New York, London, 1963.

Le, H., Lewis, A., Bharath, K., and Fallaize, C. A diffusion approach to stein's method on riemannian manifolds. *arXiv preprint arXiv:2003.11497*, 2020.

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

Ley, C. and Verdebout, T. *Modern directional statistics*. Chapman and Hall/CRC, 2017.

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

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

Liu, Q., Lee, J., and Jordan, M. A kernelized stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning*, pp. 276–284, 2016.

Liu, S. and Kanamori, T. Estimating density models with complex truncation boundaries. *arXiv preprint arXiv:1910.03834*, 2019.

Lloyd, J. R. and Ghahramani, Z. Statistical model criticism using kernel two sample tests. *Advances in Neural Information Processing Systems*, 28:829–837, 2015.

Ma, Y.-A., Chen, T., and Fox, E. A complete recipe for stochastic gradient mcmc. In *Advances in Neural Information Processing Systems*, pp. 2917–2925, 2015.

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

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

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

Pawlowsky-Glahn, V. and Buccianti, A. *Compositional data analysis: Theory and applications*. John Wiley & Sons, 2011.

Pourzanjani, A. A., Jiang, R. M., Mitchell, B., Atzberger, P. J., and Petzold, L. R. General bayesian inference over the stiefel manifold via the givens representation. *arXiv preprint arXiv:1710.09443*, 2017.

Sei, T., Shibata, H., Takemura, A., Ohara, K., and Takayama, N. Properties and applications of fisher distribution on the rotation group. *Journal of Multivariate Analysis*, 116: 440–455, 2013.

Serfling, R. J. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009.

Seth, S., Murray, I., Williams, C. K., et al. Model criticism in latent space. *Bayesian Analysis*, 14(3):703–725, 2019.

Singh, H., Hnizdo, V., and Demchuk, E. Probabilistic model for two dependent circular variables. *Biometrika*, 89:719–723, 2002.

Song, L., Huang, J., Smola, A., and Fukumizu, K. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 961–968, 2009.

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

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12 (Jul):2389–2410, 2011.

Sutherland, D. J., Tung, H.-Y., Strathmann, H., De, S., Ramdas, A., Smola, A., and Gretton, A. Generative models and model criticism via optimized maximum mean discrepancy. *arXiv preprint arXiv:1611.04488*, 2016.

Uehara, M., Matsuda, T., and Kim, J. K. Imputation estimators for unnormalized models with missing data. In *International Conference on Artificial Intelligence and Statistics*, pp. 831–841. PMLR, 2020.

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Xu, W. and Matsuda, T. A stein goodness-of-fit test for directional distributions. In *International Conference on Artificial Intelligence and Statistics*, pp. 831–841. PMLR, 2020.

Xu, W. and Reinert, G. A stein goodness-of-test for exponential random graph models. In *International Conference on Artificial Intelligence and Statistics*, pp. 415–423. PMLR, 2021.

Yang, J., Liu, Q., Rao, V., and Neville, J. Goodness-of-fit testing for discrete distributions via stein discrepancy. In *International Conference on Machine Learning*, pp. 5557–5566, 2018.

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