# Interpretable Stein Goodness-of-fit Tests on Riemannian Manifolds Supplementary Material

## A. Proofs and Derivations

### Stein's Identity

Proof of Theorem 1

*Proof.* Let $\omega = \sum_{i=1}^{d} f^i \mathrm{d}\theta^{(-i)}$, where $\mathrm{d}\theta^{(-i)} = (-1)^i \mathrm{d}\theta^{i+1} \wedge \cdots \wedge \mathrm{d}\theta^d \wedge \mathrm{d}\theta^1 \cdots \wedge \mathrm{d}\theta^{i-1}$ for $i = 1, \ldots, d$. Then,

$$\mathrm{d}(qJ\omega) = \sum_{i=1}^{d} \left( \frac{\partial f^i}{\partial \theta^i} + f^i \frac{\partial}{\partial \theta^i} \log(qJ) \right) \mathrm{d}\theta^1 \wedge \cdots \wedge \mathrm{d}\theta^d = (qJ\mathcal{A}_q^{(1)}\mathbf{f}) \mathrm{d}\theta^1 \wedge \cdots \wedge \mathrm{d}\theta^d.$$

Therefore, from Theorem 1 and Corollary 1,

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

□

### Quadratic form of mKSD

Proof of Theorem 2

*Proof.* We show that, the mKSD admits the form of taking expectation over $p$ for bivariate functions $h_q^{(c)}$ which is independent of $p$. $h_q^{(c)}$ is also referred as the Stein kernel. The proof utilize the reproducing property of relevant RKHS and the fact that $\mathcal{A}_q^{(c)}$ is a linear functional of relevant test function $f$.

For $c = 1$, the test function is a stack of $d$-dimensional RKHS functions $\mathbf{f} \in \mathcal{H}^d$. $\mathbb{E}_p[\mathcal{A}_q^{(1)}\mathbf{f}]$ is a linear functional of $\mathbf{f} \in \mathcal{H}^d$. Then, from the Riesz representation theorem, there uniquely exists $\mathbf{r} = (r_1, \ldots, r_d) \in \mathcal{H}^d$ such that $\mathbb{E}_p[\mathcal{A}_q^{(1)}\mathbf{f}] = \langle \mathbf{f}, \mathbf{r} \rangle_{\mathcal{H}^d}$. By using the reproducing property of $\mathcal{H}$ associate with kernel $k$, we obtain

$$r_i(x) = \mathbb{E}_{\tilde{x}\sim p} \left[ k(x, \tilde{x}) \frac{\partial}{\partial \tilde{\theta}^i} \log(qJ) + \frac{\partial}{\partial \tilde{\theta}^i} k(x, \tilde{x}) \right], \tag{18}$$

for $i = 1, \ldots, d$. Thus, the maximization in $\mathrm{mKSD}^{(1)}(p\|q)$ is attained by $\mathbf{f} = \mathbf{r}/\|\mathbf{r}\|_{\mathcal{H}^d}$ and $\mathrm{mKSD}^{(1)}(p\|q)^2 = \|\mathbf{r}\|_{\mathcal{H}^d}^2$. Therefore, the quadratic form is obtained after straightforward calculations:

$$\mathrm{mKSD}^{(1)}(p\|q)^2 = \left\langle \mathbb{E}_{x\sim p}[\mathcal{A}_q^{(1)} k(x, \cdot)], \mathbb{E}_{\tilde{x}\sim p}[\mathcal{A}_q^{(1)} k(\tilde{x}, \cdot)] \right\rangle_{\mathcal{H}^d} = \mathbb{E}_{x,\tilde{x}\sim p} \Bigg[ \underbrace{\left\langle \mathcal{A}_q^{(1)} k(x, \cdot), \mathcal{A}_q^{(1)} k(\tilde{x}, \cdot) \right\rangle_{\mathcal{H}^d}}_{h_q^{(1)}(x,\tilde{x})} \Bigg],$$

and the assertion follows.

For $c = 2$, similar argument applies where the test function is a scalar-valued RKHS $\tilde{f} \in \mathcal{H}$. Instead of Eq.(18), we have $\tilde{r} \in \mathcal{H}$, s.t. $\mathbb{E}_p[\mathcal{A}_q^{(2)}\mathbf{f}] = \langle \tilde{f}, \tilde{r} \rangle_{\mathcal{H}}$ and

$$\tilde{r}(x) = \mathbb{E}_{\tilde{x}\sim p} \left[ \sum_{ij} g^{ij} \left( \frac{\partial}{\partial \tilde{\theta}^j} k(x, \tilde{x}) \frac{\partial}{\partial \tilde{\theta}^i} \log(qJ) + \frac{\partial^2}{\partial \tilde{\theta}^i \partial \tilde{\theta}^j} k(x, \tilde{x}) \right) \right], \tag{19}$$

and the maximization in $\mathrm{mKSD}^{(2)}(p\|q)$ is attained by $\tilde{f} = \tilde{r}/\|\tilde{r}\|_{\mathcal{H}}$; thus $\mathrm{mKSD}^{(2)}(p\|q)^2 = \|\tilde{r}\|_{\mathcal{H}}^2$. The assertion then follows from the similar calculations as above.

For $c = 0$, the quadratic form is readily obtained from derivation of maximum-mean-discrepancy (MMD) (Gretton et al., 2007) form as shown in Theorem 4. Alternatively, for scalar test function $h \in \mathcal{H}$, we can write,

$$\mathrm{mKSD}^{(0)}(p\|q) = \sup_{\|h\|_{\mathcal{H}} \leq 1} \mathbb{E}_p[\mathcal{A}_q^{(0)} h] = \sup_{\|h\|_{\mathcal{H}} \leq 1} \left| \mathbb{E}_p[h] - \mathbb{E}_q[h] \right|,$$

where taking the supreme we get,

$$\begin{aligned}
\mathrm{mKSD}^{(0)}(p\|q)^2 &= \Big\langle \mathbb{E}_p\big[k(x,\cdot) - \mathbb{E}_q[k(x,\cdot)]\big], \mathbb{E}_p\big[k(\tilde{x},\cdot) - \mathbb{E}_q[k(\tilde{x},\cdot)]\big] \Big\rangle_{\mathcal{H}} \\
&= \mathbb{E}_{x,\tilde{x}\sim p} \Big\langle \underbrace{k(x,\cdot) - \mathbb{E}_q[k(x,\cdot)]}_{\mathcal{A}_q^{(0)} k(x,\cdot)}, k(\tilde{x},\cdot) - \mathbb{E}_q[k(\tilde{x},\cdot)] \Big\rangle_{\mathcal{H}}.
\end{aligned}$$

The assertion follows. □

The quadratic form is useful when computing the empirical estimate for the expectation where only samples from unknown distribution $p$ is observed. We also note that $\mathbb{E}_q[k(\tilde{x},\cdot)]$, in general, is not possible to obtain in analytical form, especially when the density $q$ is only given up to normalization. Samples from $q$, if possible to obtain from unnormalized density, can be useful to estimate $\mathcal{A}_q^{(0)} k(x,\cdot)$, where we denote as $\widehat{\mathcal{A}_q^{(0)}} k(x,\cdot)$.

**Characterisation of mKSD**

Proof of Theorem 3

*Proof.* Denote $\mathbf{s}_p^{(c)}(\cdot) = \mathbb{E}_{\tilde{x}\sim p}[\mathcal{A}_q^{(c)} k(\tilde{x},\cdot)] \in \mathcal{F}$ and we can write $\mathrm{mKSD}^{(c)}(p\|q)^2 = \|\mathbf{s}_p(\cdot)\|_{\mathcal{F}}^2 \geq 0$, where $\mathcal{F}$ can be $\mathcal{H}$ for $c = 0, 2$ or $\mathcal{H}^d$ for $c = 1$. If $p = q$, then $\mathrm{mKSD}^{(c)}(p\|q)^2 = 0$ from the Stein identity.

Conversely, if $\mathrm{mKSD}^{(c)}(p\|q)^2 = 0$, then $\mathbf{s}_p^{(c)}(x) = \mathbf{0}$, a zero vector in $\mathbb{R}^d$ for $c = 1$ and a scalar zero in $\mathbb{R}$ for $c = 0, 2$, $\forall x$, s.t. $p(x) > 0$. Then, from $\log(q/p) = \log(q) - \log(p) = (\log(q) - \log(J)) - (\log(p) - \log(J)) = \log(qJ) - \log(pJ)$, we obtain,

$$\mathbb{E}_{\tilde{x}\sim p}\left[L_i(\tilde{x}) k(\tilde{x}, x)\right] = (\mathbf{s}_p^{(1)})_i(x) - \mathbb{E}_{\tilde{x}\sim p}\left[\mathcal{A}_p^{(1)} k(\tilde{x}, x)\right] = 0,$$

and

$$\mathbb{E}_{\tilde{x}\sim p}\left[L(\tilde{x}) k(\tilde{x}, x)\right] = (\mathbf{s}_p^{(c)})(x) - \mathbb{E}_{\tilde{x}\sim p}\left[\mathcal{A}_p^{(c)} k(\tilde{x}, x)\right] = 0,$$

for $c = 0, 2$, for every $x$ with positive densities. Since $k$ is compact-universal, vanishes at $\partial\mathcal{M}$ and $\mathcal{M}$ is smooth and compact, the injectivity result in Carmeli et al. (2010, Theorem 4(b)) implies that $L_i^{(1)} = 0, \forall i$ (for $c = 1$, $i \in \{1, \ldots, d\}$; for $c = 0, 2$, $i = 1$). Therefore, $\log(q/p)$ is constant on $\mathcal{M}$. Since both $p$ and $q$ are both densities on $\mathcal{M}$ that integrate to one, we conclude $p = q$. □

**Asymptotics of mKSD**

Proof of Theorem 5

*Proof.* To show part 1, it is enough to check the mKSD statistics is degenerate U-statistics under $H_0 : p = q$. By considering test function $f = k(x,\cdot)$ (or its relevant vector-valued form for $c = 1$), Stein identity shows that,

$$\mathbb{E}_{\tilde{x}\sim p}[\mathcal{A}_q^{(c)} k(x, \tilde{x})] = 0, \forall x \in \mathcal{M},$$

so that the variance $\sigma_c^2 = 0$ for $c = 0, 1, 2$. Then the standard results for degenerate U-statistics in (Serfling, 2009, Section 5.5.2) apply and the assertions follow.

In addition, it is interesting to note link the result for $c = 0$ with the asymptotic result in as

$$h_q^{(0)}(x, \tilde{x}) = k(x, \tilde{x}) - \mu_q(x) - \mu_q(\tilde{x}) + c_q,$$

where $c_q = \mathbb{E}_{x,\tilde{x}\sim q} k(x,\tilde{x})$ is a constant, and that $\mu_q(x) = \mathbb{E}_{\tilde{x}\sim q} k(x,\tilde{x})$ being only a function of $x$, $\xi(\tilde{x}) = \mathbb{E}_{x\sim q} k(x,\tilde{x})$ being only a function of $\tilde{x}$, are the mean embedding function for density $q$. The formulation is analogous to the asymptotic results for MMD, as shown in (Gretton et al., 2007, Theorem 8): $h_q^{(0)}(x,\tilde{x})$ is equivalent to the notion of $\tilde{k}(x,\tilde{x})$ in (Gretton et al., 2007).

Part 2 follows as $\sigma_c^2 > 0$ under $H_1 : p \neq q$ by Theorem 3. Apply asymptotic distribution of non-degenerate U-statistics (Serfling, 2009, Section 5.5.1) and the assertions follow. □

**Asymptotics for mFSSD** To compute the empirical version of mFSSD, we consider the empirical version $\mathbf{s}_p(\cdot)$ in Eq.(16) from samples $x_1, \ldots, x_n \sim p$:

$$\widehat{\mathbf{s}}_p(\cdot) = \frac{1}{n}\sum_i [\mathcal{A}_q^{(1)} k(x_i, \cdot)].$$

Then the empirical mFSSD has the form

$$\widehat{\mathrm{mFSSD}}^2 = \frac{1}{dJ}\sum_{i=1}^{d}\sum_{j=1}^{J} (\widehat{\mathbf{s}}_p(v_j))_i^2, \tag{20}$$

for any set of test locations $\{v_j\}_{j=1}^J$.

**Proposition 4.** *Assume the conditions in Theorem 3 hold, and $\mathbb{E}_{x\sim p}[\|\mathbf{s}_p(x)\|^2] < \infty$. Under $H_1 : p \neq q$,*

$$\sqrt{n}\cdot\left(\widehat{\mathrm{mFSSD}}^2 - \mathrm{mFSSD}^2\right) \xrightarrow{d} \mathcal{N}(0, \tilde{\sigma}_{H_1}^2),$$

*where $\tilde{\sigma}_{H_1}^2$ denotes the variance for $\widehat{\mathrm{mFSSD}}^2$.*

*Proof.* With the assumed regularity conditions, Eq.(20) is in the form of the non-degenerate U-statistics with $\tilde{\sigma}_{H_1}^2 > 0$. The asymptotic normality follows from (Serfling, 2009, Section 5.5.1), similarly described in (Jitkrittum et al., 2017, Proposition 2). □

The asymptotic normality for $\widehat{\mathrm{mFSSD}}^2$ in Proposition 4 enables derivation of the approximate test power, similarly as described in Section 4.2 for kernel choice.

**Proposition 5.** [Approximate test power of $n\cdot\widehat{\mathrm{mFSSD}}^2$] *Under $H_1$, for large $n$ and fixed $r$, the test power is*

$$\mathbb{P}_{H_1}(n\cdot\widehat{\mathrm{mFSSD}}^2 > r) \approx 1 - \Phi\left(\frac{r}{\sqrt{n}\tilde{\sigma}_{H_1}^2} - \sqrt{n}\frac{\mathrm{mFSSD}^2}{\tilde{\sigma}_{H_1}^2}\right),$$

*where $\Phi$ denotes the cumulative distribution function of the standard normal distribution, and $\tilde{\sigma}_{H_1}^2$ is defined in Proposition 4.*

Due to $\sqrt{n}$ scaling in Proposition 4, maximising the approximate test power for $n\cdot\widehat{\mathrm{mFSSD}}^2$ can be approximated by maximizing $\frac{\mathrm{mFSSD}^2}{\tilde{\sigma}_{H_1}^2}$ to obtain optimal test locations under the alternative $H_1 : p \neq q$, which is described in Section 5.

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

for $V = \{v_j\}_{j=1}^J$ as the set of test locations to be optimised.

## B. More on Bahadur Efficiency

In this section, we introduce the relevant concepts to study Approximate Relative Efficiency (ARE) between two tests, characterised by *Bahadur slope* (Bahadur et al., 1960) and corresponding *Bahadur efficiency*.

### B.1. Approximate Bahadur Slope

We first define Bahadur slope for general tests (Gleser, 1966) and its applications in kernel-based tests (Jitkrittum et al., 2017; Garreau et al., 2017). Consider the test procedure with null hypothesis $H_0 : \omega \in \Omega_0$ and the alternative $H_1 : \omega \in \Omega \backslash \Omega_0$, where $\Omega$ and $\Omega_0$ are arbitrary sets. Denote $T_n$ as the test statistic computed from a sample of size $n$.

**Definition 1.** *For $\omega_0 \in \Omega_0$, let F be the asymptotic null distribution*

$$F(t) = \lim_{n\to\infty} \mathbb{P}_{\omega_0}(T_n < t)$$

*which is assumed to be continuous and common $\forall \omega_0 \in \Omega_0$. Assume that there exists a continuous strictly increasing function $\rho : (0,\infty) \to (0,\infty)$ s.t $\lim_{n\to\infty} \rho(n) = \infty$. Denote*

$$c(\omega) = -2\,\mathrm{plim}_{n\to\infty} \frac{\log(1 - F(T_n))}{\rho(n)}, \tag{21}$$

*for some bounded non-negative function c such that $c(\omega_0) = 0$ when $\omega_0 \in \Omega_0$. The function $c(\omega)$ is known as* approximate Bahadur slope.

**Definition 2.** *Let $\mathcal{D}(a,t)$ be a class of all continuous cumulative distribution functions (CDF) F such that $-2\log(1 - F(x)) = ax^t(1 + o(1))$, as $x \to \infty$ for $a > 0$ and $t > 0$.*

**Proposition 6.** *The approximate Bahadur slope (ABS) for the tests with $\mathrm{mKSD}^{(c)}$, $c = 0, 1, 2$ is*

$$c^{(\mathrm{mKSD}^{(c)})} := \frac{\mathbb{E}_p[h_q^{(c)}(x,\tilde{x})]}{\mathbb{E}_q[h_q^{(c)}(x,\tilde{x})^2]^{\frac{1}{2}}},$$

*where $h_q^{(c)}(x,\tilde{x})$ is the Stein kernel for $\mathrm{mKSD}^{(c)}$, and $\rho(n) = n$.*

*Proof.* Using Theorem 9 and Theorem 11 in (Jitkrittum et al., 2017), we know that $n \cdot \mathrm{mKSD}_u^{(c)}(p\|q)^2$ in Eq.(10) is in the class of $\mathcal{D}(a = 1/\omega_c, t = 1)$ for $\omega_c^2$ is the variance of the statistic. By Stein identity, $\mathbb{E}_{x\sim q}\mathbb{E}_{\tilde{x}\sim q}\left[h_q^{(c)}(x,\tilde{x})\right]^2 = 0$. Hence, using second point in Theorem 9 (Jitkrittum et al., 2017) and choosing $\rho = n$, we know that $n \cdot \mathrm{mKSD}_u^{(c)}(p\|q)^2 \backslash \rho(n) \to \mathrm{mKSD}^{(c)}(p\|q)^2$ by weak law of large numbers. □

### B.2. Asymptotic Relative Efficiencies Between Tests with Different $\mathcal{A}_q$s

Asymptotic Relative Efficiency (ARE) between two statistical testing procedures measures how fast the p-values of one test shrinks to 0, relatively to the other's. If it is faster, for given problem under the alternative, it is more sensitive to pick up the alternative, where we call the test more "statistically efficient". With ABS, we are ready to define approximate Bahadur efficiency.

**Definition 3.** *Given two sequences of test statistics, $T_n^{(1)}$ and $T_n^{(2)}$ and their ABS $c^{(1)}$ and $c^{(2)}$, the approximate Bahadur efficiency of $T_n^{(1)}$ relative to $T_n^{(2)}$ is*

$$\mathrm{E}(\omega_A) := \frac{c^{(1)}(\omega_A)}{c^{(2)}(\omega_A)} \tag{22}$$

*for $\omega_A \in \Omega \backslash \Omega_0$, in the space of alternative models.*

If $\mathrm{E}(\omega_\mathrm{A}) > 1$, then $T_n^{(1)}$ is asymptotically more efficient than $T_n^{(2)}$ in the sense of Bahadur, for the particular problem specified by $\omega_A \in \Omega \backslash \Omega_0$.

### B.3. The Case Study on Circular distribution $\mathcal{S}^1$

Proof of Theorem 6

*Proof.* To compute $\mathrm{E}_{1,2}(\kappa)$, we can rewrite the following:

$$\mathrm{E}_{1,2}(\kappa) = \frac{\mathbb{E}_p[h_q^{(1)}(x,\tilde{x})]}{\mathbb{E}_p[h_q^{(2)}(x,\tilde{x})]} \cdot \frac{\mathbb{E}_q[h_q^{(2)}(x,\tilde{x})^2]^{\frac{1}{2}}}{\mathbb{E}_q[h_q^{(1)}(x,\tilde{x})^2]^{\frac{1}{2}}}$$

The second term only involves integrals over $q(x) \propto 1$, which is independent of $\kappa$ and we can solve it as $\frac{\mathbb{E}_q[h_q^{(2)}(x,\tilde{x})^2]^{\frac{1}{2}}}{\mathbb{E}_q[h_q^{(1)}(x,\tilde{x})^2]^{\frac{1}{2}}} = 1.692 > 1$. For the first term, the ratio is monotonic decreasing w.r.t. $\kappa > 0$ and $\frac{\mathbb{E}_p[h_q^{(1)}(x,\tilde{x})]}{\mathbb{E}_p[h_q^{(2)}(x,\tilde{x})]}$ is lower bounded by 2 due to exponential-trace kernel and $\mathcal{S}^1$ embedded in $\mathbb{R}^2$. Hence, for $\kappa > 0$, $\mathrm{E}_{1,2} > 1$. □

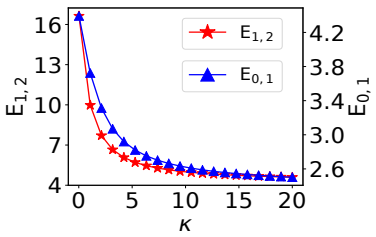



*Figure 3.* Relative Test Efficiency

We can apply similar approach to compare the relative test efficiency $\mathrm{E}_{0,1}(\kappa)$ between $\mathrm{mKSD}^{(0)}$ and $\mathrm{mKSD}^{(1)}$. We plot numerical solutions in Figure 3. From Figure 3, we see that $\mathbb{E}_{1,2}$ and $\mathbb{E}_{0,1}$ both greater than 1 for $\kappa \in (0, 10)$. For further increase of $\kappa$, there is a trend for both relative efficiencies stabilising at some value greater than 1. Theoretical analysis for such limiting behaviour is of an interesting future topic. Although Figure 3 shows that $\mathbb{E}_{0,1}(\kappa) > 1$ for small perturbation from the null, i.e. $\kappa \in (0, 10)$ which suggest the relative efficiency of $\mathrm{mKSD}^{(0)}$ is higher than the first order test $\mathrm{mKSD}^{(1)}$, it is usually not possible to compute MMD analytically and the normalized density is required.

Intuitively, with sampling error of order $\sqrt{n}$ and $\rho(n) = n$ is chosen to compute Bahadur slope, the MMD computed from samples are less efficient to perform goodness-of-fit test compared to mKSD tests that directly access the unnormalized density, as shown in Figure 1. Similar findings are also observed in other settings where MMD is considered to perform goodness-of-fit tests (Liu et al., 2016; Jitkrittum et al., 2017; Yang et al., 2018; 2019; Xu & Matsuda, 2020). In addition, correctly sampling from Riemannian manifold is non-trivial and can be time-consuming for sample-based tests.

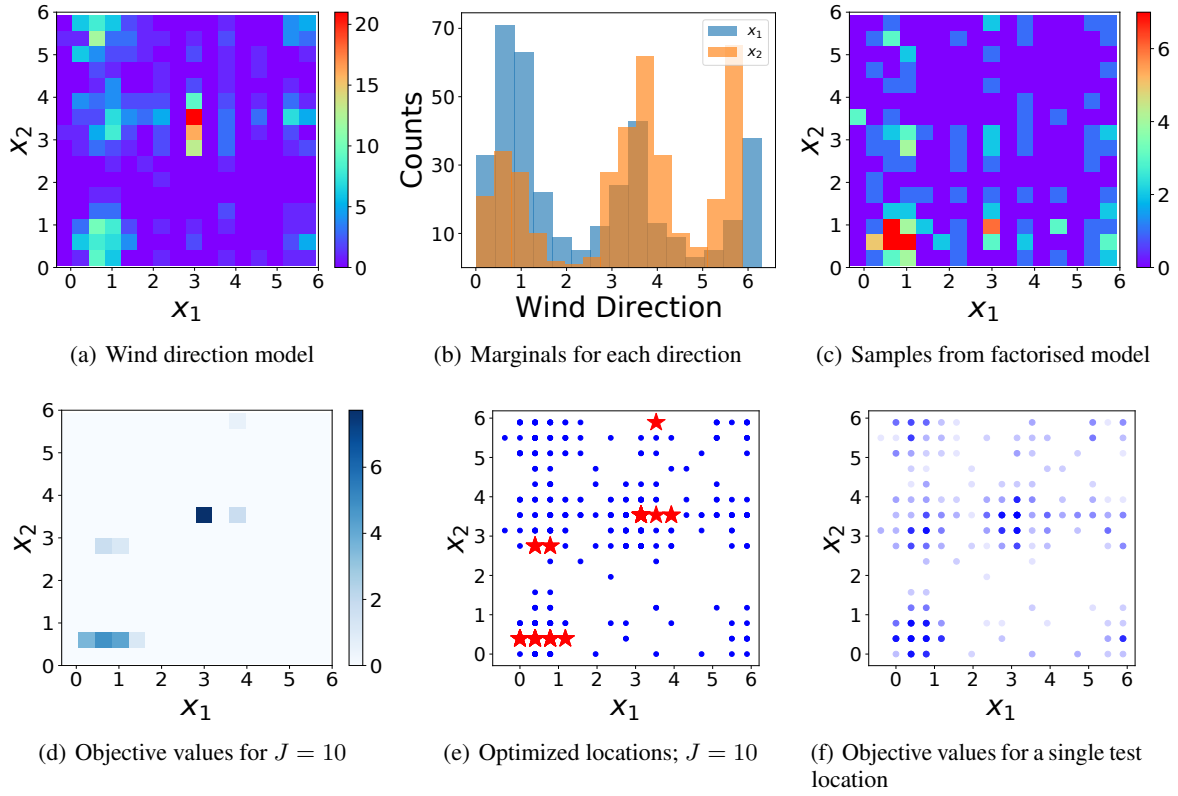
## C. More on Model Criticism

In this section, we provide additional details on model criticism for wind data present in Section 8.2. We fitted the model in Eq.(2) by using noise contrastive estimation (Uehara et al., 2020) and our test does not find evidence to reject the fitted model, suggesting a good fit for the wind direction data. In addition, we consider the model without interaction term between two direcitons:

$$\widetilde{q}(x_1, x_2 \mid \widetilde{\eta}) \propto \exp\{\kappa_1 \cos(x_1 - \mu_1) + \kappa_2 \cos(x_2 - \mu_2)\}, \tag{23}$$

which is equivalent to model in Eq.(2) by imposing $\lambda_{12} = 0$. This model can be viewed as product of marginal distributions of $x_1$ and $x_2$ and we refer as factorised model. Our test reject the null at test level $\alpha = 0.05$ suggesting a poor fit of the factorised model.

To further visualize the difference between models in Eq.(2) and Eq.(23), we plot histogram of each wind direction in Figure 4(b) and samples from the factorised model $\widetilde{q}$ in Figure 4(c) where no interactions are present between $x_1$ and $x_2$. Compare with the wind direction data, shown again in Figure 4(a), we can see that Figure 4(c) differs the most at the regions of $\tilde{x} = (x_1, x_2) = (2.8, \pi)$ (data model denser) and $\tilde{x}' = (x_1, x_2) = (1, 1)$ ($\tilde{q}$ model denser). Such difference is captured by our optimized test locations from mFSSD in Figure 4(e), where $\tilde{x}$ is at the region with 3 stars in a row and $\tilde{x}'$ is around the region with 4-stars in a row. It shows the effectiveness of mFSSD in distinguishing the differences between distributions. As $\tilde{q}$ is referred as imposing data model in Eq.(2) to be 0, a negative $\lambda_{12} = -1.1274 < 0$ in the data model implies that positive $\sin(x_1 - \mu_1)\sin(x_2 - \mu_2)$ is less dense. With $\mu_1 = 1.1499 = \mu_2$, $\sin(x_1 - \mu_1)\sin(x_2 - \mu 2)$ is positive around the region the $\tilde{x}'$ making the data model less dense, as shown in Figure 4(a) and 4(c).




(a) Wind direction model (b) Marginals for each direction (c) Samples from factorised model

(d) Objective values for $J = 10$ (e) Optimized locations; $J = 10$ (f) Objective values for a single test location

*Figure 4.* Visualizing the fitted model and rejected model for wind direction data.