

Interpretable Stein Goodness-of-fit Tests on Riemannian Manifolds

Wenkai Xu¹ Takeru Matsuda²

¹Gatsby Computational Neuroscience Unit, London, UK

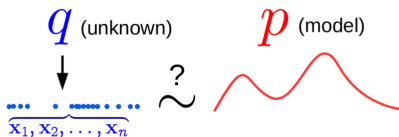
²RIKEN Center for Brain Science, Tokyo, Japan

Key Message from This Talk

Our Tasks for Riemannian manifold \mathcal{M} , we perform,

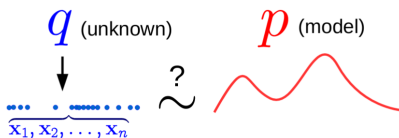
Key Message from This Talk

Our Tasks for Riemannian manifold \mathcal{M} , we perform,
Goodness-of-fit test:



Key Message from This Talk

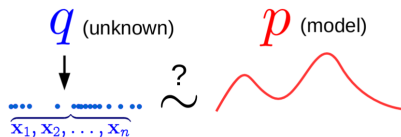
Our Tasks for Riemannian manifold \mathcal{M} , we perform,
Goodness-of-fit test:



Model criticism: if $q \neq p$, how do they differ?

Key Message from This Talk

Our Tasks for Riemannian manifold \mathcal{M} , we perform,
Goodness-of-fit test:



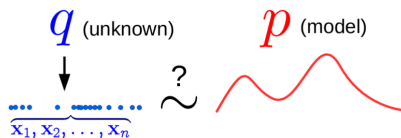
Model criticism: if $q \neq p$, how do they differ?

Our Contributions

- ▶ Develop **kernel Stein** goodness-of-fit testing procedures for **unnormalized densities** on Riemannian manifold

Key Message from This Talk

Our Tasks for Riemannian manifold \mathcal{M} , we perform,
Goodness-of-fit test:



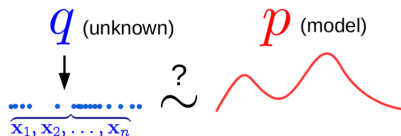
Model criticism: if $q \neq p$, how do they differ?

Our Contributions

- ▶ Develop **kernel Stein** goodness-of-fit testing procedures for **unnormalized densities** on Riemannian manifold
- ▶ Perform corresponding *interpretable* model criticism

Key Message from This Talk

Our Tasks for Riemannian manifold \mathcal{M} , we perform,
Goodness-of-fit test:



Model criticism: if $q \neq p$, how do they differ?

Our Contributions

- ▶ Develop **kernel Stein** goodness-of-fit testing procedures for **unnormalized densities** on Riemannian manifold
- ▶ Perform corresponding *interpretable* model criticism
- ▶ Compare 3 different **kernel Stein** tests with Bahadur efficiency

Unnormalized Densities on Manifold

Density p in (\mathcal{M}, g) has the unnormalized form, e.g.

$$p(X) \propto \text{etr}(\Theta^\top X), \quad X \in \mathcal{M}$$

is called Fisher distribution for rotation group $\mathcal{M} = \text{SO}(d)$ or matrix-Langevin distribution for matrix-valued variables.

Unnormalized Densities on Manifold

Density p in (\mathcal{M}, g) has the unnormalized form, e.g.

$$p(X) \propto \text{etr}(\Theta^\top X), \quad X \in \mathcal{M}$$

is called Fisher distribution for rotation group $\mathcal{M} = \text{SO}(d)$ or matrix-Langevin distribution for matrix-valued variables.

- ▶ Normalization constant $Z = \int_{\mathcal{M}} \text{etr}(\Theta^\top X) dX$ can be hard to compute, especially in high dimensions.

Unnormalized Densities on Manifold

Density p in (\mathcal{M}, g) has the unnormalized form, e.g.

$$p(X) \propto \text{etr}(\Theta^\top X), \quad X \in \mathcal{M}$$

is called Fisher distribution for rotation group $\mathcal{M} = \text{SO}(d)$ or matrix-Langevin distribution for matrix-valued variables.

- ▶ Normalization constant $Z = \int_{\mathcal{M}} \text{etr}(\Theta^\top X) dX$ can be hard to compute, especially in high dimensions.
- ▶ May have non-vanishing boundary $\partial\mathcal{M}$.

Unnormalized Densities on Manifold

Density p in (\mathcal{M}, g) has the unnormalized form, e.g.

$$p(X) \propto \text{etr}(\Theta^\top X), \quad X \in \mathcal{M}$$

is called Fisher distribution for rotation group $\mathcal{M} = \text{SO}(d)$ or matrix-Langevin distribution for matrix-valued variables.

- ▶ Normalization constant $Z = \int_{\mathcal{M}} \text{etr}(\Theta^\top X) dX$ can be hard to compute, especially in high dimensions.
- ▶ May have non-vanishing boundary $\partial\mathcal{M}$.
- ▶ Multi-variate statistical procedures for Euclidean manifold does not apply.

Stein Operators for Manifold

Stein's identity:

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

Stein Operators for Manifold

Stein's identity:

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

For Unnormalized Densities

► First Order:

$$\mathcal{A}_p^{(1)} \mathbf{f} = \sum_{i=1}^d \left(\frac{\partial f^i}{\partial \theta^i} + f^i \frac{\partial}{\partial \theta^i} \log(pJ) \right), \quad (1)$$

Stein Operators for Manifold

Stein's identity:

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

For Unnormalized Densities

► First Order:

$$\mathcal{A}_p^{(1)} \mathbf{f} = \sum_{i=1}^d \left(\frac{\partial f^i}{\partial \theta^i} + f^i \frac{\partial}{\partial \theta^i} \log(pJ) \right), \quad (1)$$

$\mathbf{f} = (f^1, \dots, f^d)$: vector-valued test function;

Stein Operators for Manifold

Stein's identity:

$$\mathbb{E}_{\mathbf{p}}[\mathcal{A}_{\mathbf{p}}f] = 0$$

For Unnormalized Densities

► First Order:

$$\mathcal{A}_{\mathbf{p}}^{(1)}\mathbf{f} = \sum_{i=1}^d \left(\frac{\partial f^i}{\partial \theta^i} + f^i \frac{\partial}{\partial \theta^i} \log(\mathbf{p}J) \right), \quad (1)$$

$\mathbf{f} = (f^1, \dots, f^d)$: vector-valued test function;

► Second Order:

$$\mathcal{A}_{\mathbf{p}}^{(2)}\tilde{f} = \sum_{ij} \left(g^{ij} \frac{\partial^2 \tilde{f}}{\partial \theta^i \partial \theta^j} + g^{ij} \frac{\partial \tilde{f}}{\partial \theta^j} \frac{\partial \log \mathbf{p}J}{\partial \theta^i} \right) \quad (2)$$

Stein Operators for Manifold

Stein's identity:

$$\mathbb{E}_{\mathbf{p}}[\mathcal{A}_{\mathbf{p}}f] = 0$$

For Unnormalized Densities

► First Order:

$$\mathcal{A}_{\mathbf{p}}^{(1)}\mathbf{f} = \sum_{i=1}^d \left(\frac{\partial f^i}{\partial \theta^i} + f^i \frac{\partial}{\partial \theta^i} \log(\mathbf{p}J) \right), \quad (1)$$

$\mathbf{f} = (f^1, \dots, f^d)$: vector-valued test function;

► Second Order:

$$\mathcal{A}_{\mathbf{p}}^{(2)}\tilde{f} = \sum_{ij} \left(g^{ij} \frac{\partial^2 \tilde{f}}{\partial \theta^i \partial \theta^j} + g^{ij} \frac{\partial \tilde{f}}{\partial \theta^j} \frac{\partial \log \mathbf{p}J}{\partial \theta^i} \right) \quad (2)$$

\tilde{f} scalar valued test function.

Stein Operators for Manifold

Stein's identity:

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

For Unnormalized Densities

► First Order:

$$\mathcal{A}_p^{(1)} \mathbf{f} = \sum_{i=1}^d \left(\frac{\partial f^i}{\partial \theta^i} + f^i \frac{\partial}{\partial \theta^i} \log(pJ) \right), \quad (1)$$

$\mathbf{f} = (f^1, \dots, f^d)$: vector-valued test function;

► Second Order:

$$\mathcal{A}_p^{(2)} \tilde{f} = \sum_{ij} \left(g^{ij} \frac{\partial^2 \tilde{f}}{\partial \theta^i \partial \theta^j} + g^{ij} \frac{\partial \tilde{f}}{\partial \theta^j} \frac{\partial \log pJ}{\partial \theta^i} \right) \quad (2)$$

\tilde{f} scalar valued test function.

The Connection:

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

Stein Operators for Manifold

It is also natural to consider

► Zeroth Order:

$$\mathcal{A}_p^{(0)} h = h - \mathbb{E}_p[h], \quad (3)$$

Stein Operators for Manifold

It is also natural to consider

- ▶ Zeroth Order:

$$\mathcal{A}_p^{(0)} h = h - \mathbb{E}_p[h], \quad (3)$$

as a Stein operator

h can be both scalar-valued or vector-valued;

Stein Operators for Manifold

It is also natural to consider

- ▶ Zeroth Order:

$$\mathcal{A}_p^{(0)} h = h - \mathbb{E}_p[h], \quad (3)$$

as a Stein operator

h can be both scalar-valued or vector-valued;

\mathbb{E}_p can't be computed with **unnormalized density** p .

Stein Operators for Manifold

It is also natural to consider

- ▶ Zeroth Order:

$$\mathcal{A}_p^{(0)} h = h - \mathbb{E}_p[h], \quad (3)$$

as a Stein operator

h can be both scalar-valued or vector-valued;

\mathbb{E}_p can't be computed with **unnormalized density** p .

Samples from **unnormalized density**;

The goodness-of-fit problem turns into two-sample problem:
compare samples from unknown data q with generated samples
from model p .

Manifold Kernel Stein Discrepancy (mKSD)

Consider appropriate RKHS, $\mathcal{H}^{(c)}$, as test function class, $c = 0, 1, 2$

$$\text{mKSD}^{(c)}(q||p) = \sup_{\|f\|_{\mathcal{H}^{(c)}} \leq 1} \mathbb{E}_q[\mathcal{A}_p^{(c)} f]$$

Manifold Kernel Stein Discrepancy (mKSD)

Consider appropriate RKHS, $\mathcal{H}^{(c)}$, as test function class, $c = 0, 1, 2$

$$\text{mKSD}^{(c)}(q||p) = \sup_{\|f\|_{\mathcal{H}^{(c)}} \leq 1} \mathbb{E}_q[\mathcal{A}_p^{(c)} f]$$

Reproducing property gives quadratic form:

$$\text{mKSD}^{(c)}(q||p)^2 = \mathbb{E}_{x, \tilde{x} \sim q}[h_p^{(c)}(x, \tilde{x})]$$

Manifold Kernel Stein Discrepancy (mKSD)

Consider appropriate RKHS, $\mathcal{H}^{(c)}$, as test function class, $c = 0, 1, 2$

$$\text{mKSD}^{(c)}(q\|p) = \sup_{\|f\|_{\mathcal{H}^{(c)}} \leq 1} \mathbb{E}_q[\mathcal{A}_p^{(c)} f]$$

Reproducing property gives quadratic form:

$$\text{mKSD}^{(c)}(q\|p)^2 = \mathbb{E}_{x, \tilde{x} \sim q} [h_p^{(c)}(x, \tilde{x})]$$

Empirical estimate

$$\Psi_n^2 = \frac{1}{n^2} \sum_{i,j} [h_p^{(c)}(x_i, \tilde{x}_j)]$$

Manifold Kernel Stein Discrepancy (mKSD)

Consider appropriate RKHS, $\mathcal{H}^{(c)}$, as test function class, $c = 0, 1, 2$

$$\text{mKSD}^{(c)}(q||p) = \sup_{\|f\|_{\mathcal{H}^{(c)}} \leq 1} \mathbb{E}_q[\mathcal{A}_p^{(c)} f]$$

Reproducing property gives quadratic form:

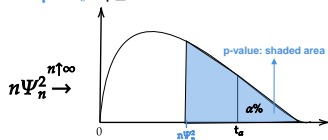
$$\text{mKSD}^{(c)}(q||p)^2 = \mathbb{E}_{x, \tilde{x} \sim q} [h_p^{(c)}(x, \tilde{x})]$$

Empirical estimate

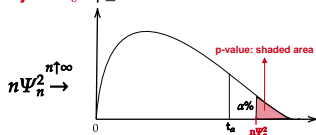
$$\Psi_n^2 = \frac{1}{n^2} \sum_{i,j} [h_p^{(c)}(x_i, \tilde{x}_j)]$$

Goodness-of-fit Testing on \mathcal{M} with mKSD

Accept H_0 : p_value $\geq \alpha$



Reject H_0 : p_value $< \alpha$



Interpretable Model Criticism

Key idea: interpret the distribution difference via test locations on \mathcal{M} where p and q differ the most.

Interpretable Model Criticism

Key idea: interpret the distribution difference via test locations on \mathcal{M} where p and q differ the most.

Manifold Finite-Set Stein Discrepancy (mFSSD) is defined by fix J test location $V = \{v_j\}_{j=1}^J$

$$\text{mFSSD}(q\|p; V)^2 = \frac{1}{dJ} \sum_{j=1}^J \sum_{i=1}^d (\mathbb{E}_{\tilde{x} \sim q}[\mathcal{A}_p^{(c)} k(\tilde{x}, v_j)])_i^2, \quad (4)$$

Interpretable Model Criticism

Key idea: interpret the distribution difference via test locations on \mathcal{M} where p and q differ the most.

Manifold Finite-Set Stein Discrepancy (mFSSD) is defined by fix J test location $V = \{v_j\}_{j=1}^J$

$$\text{mFSSD}(q\|p; V)^2 = \frac{1}{dJ} \sum_{j=1}^J \sum_{i=1}^d (\mathbb{E}_{\tilde{x} \sim q} [\mathcal{A}_p^{(c)} k(\tilde{x}, v_j)])_i^2, \quad (4)$$

Optimize the test locations w.r.t.

approximate test power:

$$V = \arg \max_{\mathbf{v}} \frac{\text{mFSSD}^2}{\tilde{\sigma}_{H_1}}, \quad (5)$$

where $\tilde{\sigma}_{H_1}^2$ denotes variance of mFSSD^2 under H_1 .

Interpretable Model Criticism

Key idea: interpret the distribution difference via test locations on \mathcal{M} where p and q differ the most.

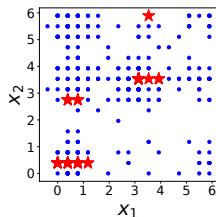
Manifold Finite-Set Stein Discrepancy (mFSSD) is defined by fix J test location $V = \{v_j\}_{j=1}^J$

$$\text{mFSSD}(q\|p; V)^2 = \frac{1}{dJ} \sum_{j=1}^J \sum_{i=1}^d (\mathbb{E}_{\tilde{x} \sim q} [\mathcal{A}_p^{(c)} k(\tilde{x}, v_j)])_i^2, \quad (4)$$

Optimize the test locations w.r.t.
approximate test power:

$$V = \arg \max_{\mathbf{v}} \frac{\text{mFSSD}^2}{\tilde{\sigma}_{H_1}^2}, \quad (5)$$

where $\tilde{\sigma}_{H_1}^2$ denotes variance of mFSSD^2 under H_1 .



Best 10 test locations for
wind direction data

Test Comparisons

Approximate Relative Efficiency (ARE) between two tests:
how fast the p-values of one test shrinks to 0, relatively to the other's (the faster the more sensitive to pick up the alternative)

Test Comparisons

Approximate Relative Efficiency (ARE) between two tests:
how fast the p-values of one test shrinks to 0, relatively to the other's (the faster the more sensitive to pick up the alternative)

Approximate Bahadur Efficiency (ABE):
the ratio between Bahadur slopes of the tests.

Test Comparisons

Approximate Relative Efficiency (ARE) between two tests:
how fast the p-values of one test shrinks to 0, relatively to the other's (the faster the more sensitive to pick up the alternative)

Approximate Bahadur Efficiency (ABE):
the ratio between Bahadur slopes of the tests.

Case study: von-Mises Fisher distribution with scaling difference

$$q(x) \propto \exp\{\kappa u^\top x\}$$

$$H_0 : \kappa = 0, \quad \text{v.s.} \quad H_1 : \kappa > 0$$

Test Comparisons

Approximate Relative Efficiency (ARE) between two tests:
how fast the p-values of one test shrinks to 0, relatively to the other's (the faster the more sensitive to pick up the alternative)

Approximate Bahadur Efficiency (ABE):

the ratio between Bahadur slopes of the tests.

Case study: von-Mises Fisher distribution with scaling difference

$$q(x) \propto \exp\{\kappa u^\top x\}$$

$$H_0 : \kappa = 0, \quad \text{v.s.} \quad H_1 : \kappa > 0$$

* von-Mises kernel with unit bandwidth:

$$k(\tilde{x}, x) = \exp\{\kappa \tilde{x}^\top x\}$$

Test Comparisons

Approximate Relative Efficiency (ARE) between two tests:
how fast the p-values of one test shrinks to 0, relatively to the other's (the faster the more sensitive to pick up the alternative)

Approximate Bahadur Efficiency (ABE):

the ratio between Bahadur slopes of the tests.

Case study: von-Mises Fisher distribution with scaling difference

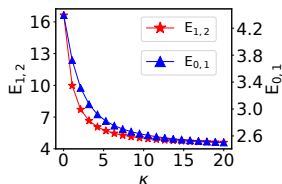
$$q(x) \propto \exp\{\kappa u^\top x\}$$

$$H_0 : \kappa = 0, \quad \text{v.s.} \quad H_1 : \kappa > 0$$

- * von-Mises kernel with unit bandwidth:

$$k(\tilde{x}, x) = \exp\{\kappa \tilde{x}^\top x\}$$

- * $E_{0,1}$: ABE of $\text{mKSD}^{(0)}$ and $\text{mKSD}^{(1)}$
- * $E_{1,2}$: ABE of $\text{mKSD}^{(1)}$ and $\text{mKSD}^{(2)}$



Thanks for Your Attention