Interpretable Stein Goodness-of-fit Tests on Riemannian Manifolds

Wenkai Xu[1] Takeru Matsuda[2]

[1]Gatsby Computational Neuroscience Unit, London, UK

[2]RIKEN Center for Brain Science, Tokyo, Japan

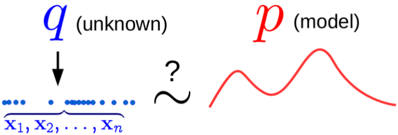
Key Message from This Talk

Our Tasks for Riemannian manifold M, we perform,

Key Message from This Talk

Our Tasks for Riemannian manifold M, we perform,

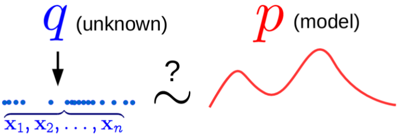
Goodness-of-fit test:



Key Message from This Talk

Our Tasks for Riemannian manifold M, we perform,

Goodness-of-fit test:

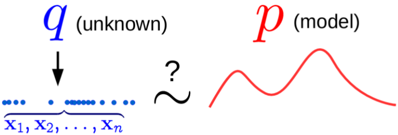


Model criticism: if $q \neq p$, how do they differ?

Key Message from This Talk

Our Tasks for Riemannian manifold M, we perform,

Goodness-of-fit test:



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

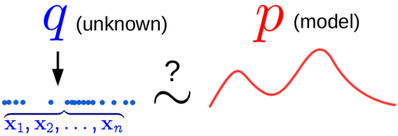
Our Contributions

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

Key Message from This Talk

Our Tasks for Riemannian manifold M, we perform,

Goodness-of-fit test:



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

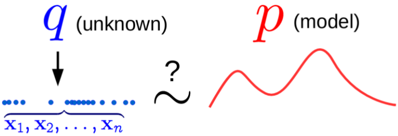
Our Contributions

- 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 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 (M, g) has the unnormalized form, e.g.

$$p(X) \propto \mathrm{etr}(\Theta^\top X), \qquad X \in M$$

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

Unnormalized Densities on Manifold

Density p in (M, g) has the unnormalized form, e.g.

$$p(X) \propto \mathrm{etr}(\Theta^\top X), \qquad X \in M$$

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

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

Unnormalized Densities on Manifold

Density p in (M, g) has the unnormalized form, e.g.

$$p(X) \propto \mathrm{etr}(\Theta^{\top} X), \qquad X \in M$$

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

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

Unnormalized Densities on Manifold

Density p in (M, g) has the unnormalized form, e.g.

$$p(X) \propto \operatorname{etr}(\Theta^\top X), \qquad X \in M$$

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

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

Stein Operators for Manifold

Stein's identity:

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

Stein Operators for Manifold

Stein's identity:

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

For Unnormalized Densities

- First Order:

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

Stein Operators for Manifold

Stein's identity:

$$\mathrm{E}_p[A_p f] = 0$$

For Unnormalized Densities

- First Order:

$$A_p^{(1)}\mathbf{f} = \sum_{i=1}^{d} \left(\frac{\partial f^i}{\partial \;{}^{i}} + f^i \frac{\partial}{\partial \;{}^{i}} \log(pJ) \right), \qquad (1)$$

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

Stein Operators for Manifold

Stein's identity:

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

For Unnormalized Densities

- First Order:

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

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

- Second Order:

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

Stein Operators for Manifold

Stein's identity:

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

For Unnormalized Densities

- First Order:

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

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

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

  $\tilde{f}$ scalar valued test function.

Stein Operators for Manifold

Stein's identity:

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

For Unnormalized Densities

- First Order:

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

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

- Second Order:

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

  $\tilde{f}$ 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:

$$A_{p}^{(0)} h = h \quad \mathrm{E}_{p}[h]; \tag{3}$$

Stein Operators for Manifold

It is also natural to consider

- Zeroth Order:

$$A_p^{(0)} h = h \quad \mathsf{E}_p[h]; \tag{3}$$

  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:

$$A_p^{(0)} h = h \quad \mathsf{E}_p[h]; \tag{3}$$

  as a Stein operator

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

  E_p can't be computed with unnormalized density p.

Stein Operators for Manifold

It is also natural to consider

- Zeroth Order:

$$A_p^{(0)}h = h \quad E_p[h]; \tag{3}$$

  as a Stein operator

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

  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, $H^{(c)}$, as test function class, $c = 0, 1, 2$

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

Manifold Kernel Stein Discrepancy (mKSD)

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

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

Reproducing property gives quadratic form:

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

Manifold Kernel Stein Discrepancy (mKSD)

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

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

Reproducing property gives quadratic form:

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

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, $c = 0, 1, 2$

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

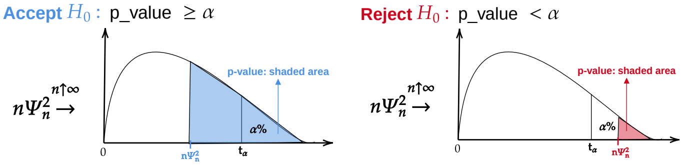
Reproducing property gives quadratic form:

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

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



Interpretable Model Criticism

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

Interpretable Model Criticism

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

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

$$\mathrm{mFSSD}(q\|p; V)^2 = \frac{1}{dJ}\sum_{j=1}^{J}\sum_{i=1}^{d}\left(\mathbb{E}_{\tilde{x}\sim q}[A_p^{(c)} k(\tilde{x}, v_j)]\right)_i^2, \qquad (4)$$

Interpretable Model Criticism

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

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

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

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

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

where $\sigma_{H_1}^2$ denotes variance of mFSSD^2 under H_1.

Interpretable Model Criticism

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

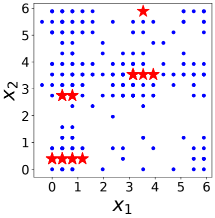
Manifold Finite-Set Stein Discrepancy (mFSSD) is defined by fix J test location $V = \{v_j\}_{j=1}^{J}$

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

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

$$V = \arg\max_{\mathbf{v}} \frac{\mathrm{mFSSD}^2}{\sigma_{H_1}}; \quad (5)$$

where $\sigma_{H_1}^2$ 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 E ciency (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 E ciency (ABE):
the ratio between Bahadur slopes of the tests.

Case study: von-Mises Fisher distribution with scaling difference

$$q(x)\ /\ \exp f\ \ u^{>} x g$$

$$H_0 :\quad = 0; \qquad \text{v.s.} \qquad H_1 :\quad > 0$$

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

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

$$H_0 : \quad = 0; \qquad \text{v.s.} \qquad H_1 : \quad > 0$$

* von-Mises kernel with unit bandwidth:

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

Test Comparisons

Approximate Relative E ciency (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 E ciency (ABE):
the ratio between Bahadur slopes of the tests.
Case study: von-Mises Fisher distribution with scaling difference

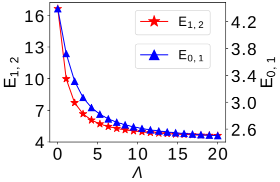
$$q(x) \; / \; \exp f \;\; u^{>} x g$$

$$H_0 : \quad = 0; \qquad \text{v.s.} \qquad H_1 : \quad > 0$$

* von-Mises kernel with unit bandwidth:

$$k(\tilde{x}; x) = \exp f \;\; \tilde{x}^{>} x g$$

* $E_{0;1}$: ABE of mKSD$^{(0)}$ and mKSD$^{(1)}$
* $E_{1;2}$: ABE of mKSD$^{(1)}$ and mKSD$^{(2)}$



Thanks for Your Attention