SBI at CMU

Larry Wasserman
larry@cmu.edu

Sources

1. Izbicki, Lee, Schafer (2014)
2. Niccolo Dalmasso, Rafael Izbicki, Ann Lee (2020)
3. Masserano, Dorigo, Izbicki, Kuusela, Lee (2023)
4. Yi, Alison, Kuusela (2024)
5. Zhu, Desai, Kuusela, Mikuni, Nachman, Wasserman (2024)
6. Walchessan, Zammit-Mangion, Huser, Kuusela (2024)
7. Walchessan, Lenzi, Kuusela (2024)
8. Stanley, Batlle, Patil, Owhadi, Kuusela (2025)
9. Carzon, Masserano, Ingram, Shen, Ribeiro, Dorigo, Doro, Speagle, Izbicki, Lee (2025)
10. Tomaselli, Ventura, Wasserman (2025)

Outline

Outline

- Brief review of SBI

Outline

- Brief review of SBI
- Summary of CMU work

Outline

- Brief review of SBI
- Summary of CMU work
- Open questions

SBI: Quick Review

SBI: Quick Review

- Two different tasks:

SBI: Quick Review

- Two different tasks:
  - Construct confidence sets (without regularity conditions)

SBI: Quick Review

- Two different tasks:
  - Construct confidence sets (without regularity conditions)
  - Estimate the likelihood function

SBI: Quick Review

- Two different tasks:
  - Construct confidence sets (without regularity conditions)
  - Estimate the likelihood function
- These often get mushed together (for example ABC).

SBI: Quick Review

- Two different tasks:
  - Construct confidence sets (without regularity conditions)
  - Estimate the likelihood function
- These often get mushed together (for example ABC).
- Complex models may fail to satisfy standard regularity conditions which means that the usual (asymptotic) methods can fail. Fortunately, SBI methods don't rely on these regularity conditions.

SBI: Quick Review

- Two different tasks:
  - Construct confidence sets (without regularity conditions)
  - Estimate the likelihood function
- These often get mushed together (for example ABC).
- Complex models may fail to satisfy standard regularity conditions which means that the usual (asymptotic) methods can fail. Fortunately, SBI methods don't rely on these regularity conditions.
- Note: I'll focus on frequentist inference. Not discussing Bayesian inference.

Setting

Setting

- We have data $\mathcal{Y}_{obs}$.

Setting

- We have data $\mathcal{Y}_{obs}$.
- Could be scalars, vectors, functions, images, etc

Setting

- We have data $\mathcal{Y}_{obs}$.
- Could be scalars, vectors, functions, images, etc
- Model:
$$\mathcal{P} = (p_\theta : \ \theta \in \Theta), \quad \Theta \subset \mathbb{R}^d.$$

Setting

- We have data $\mathcal{Y}_{obs}$.
- Could be scalars, vectors, functions, images, etc
- Model:
$$\mathcal{P} = (p_\theta : \ \theta \in \Theta), \quad \Theta \subset \mathbb{R}^d.$$
- Want to infer θ. (Infer $=$ uncertainty quantification)

Setting

- We have data $\mathcal{Y}_{obs}$.
- Could be scalars, vectors, functions, images, etc
- Model:

$$\mathcal{P} = (p_\theta : \ \theta \in \Theta), \quad \Theta \subset \mathbb{R}^d.$$

- Want to infer θ. (Infer = uncertainty quantification)
- In particular, we want a confidence set C such that

$$P_\theta(\theta \in C) \geq 1 - \alpha, \quad \text{for all } \theta.$$

Setting

- We have data $\mathcal{Y}_{obs}$.
- Could be scalars, vectors, functions, images, etc
- Model:

$$\mathcal{P} = (p_\theta : \ \theta \in \Theta), \quad \Theta \subset \mathbb{R}^d.$$

- Want to infer θ. (Infer = uncertainty quantification)
- In particular, we want a confidence set C such that

$$P_\theta(\theta \in C) \geq 1 - \alpha, \quad \text{for all } \theta.$$

- Main assumption: it is easy to simulate from p_θ

Recall: Confidence Sets by Inverting a Test

Recall: Confidence Sets by Inverting a Test

- For each θ: test H_0: $\theta_{true} = \theta$ at level α

Recall: Confidence Sets by Inverting a Test

- For each θ: test H_0: $\theta_{true} = \theta$ at level α
- Test statistic $T = T(\mathcal{Y}, \theta)$.

Recall: Confidence Sets by Inverting a Test

- For each θ: test H_0: $\theta_{true} = \theta$ at level α
- Test statistic $T = T(\mathcal{Y}, \theta)$.
- Reject H_0 if $T > q(\theta)$ where $P_\theta(T > q(\theta)) = \alpha$.

Recall: Confidence Sets by Inverting a Test

- For each θ: test H_0: $\theta_{true} = \theta$ at level α
- Test statistic $T = T(\mathcal{Y}, \theta)$.
- Reject H_0 if $T > q(\theta)$ where $P_\theta(T > q(\theta)) = \alpha$.
- Invert: $C = \{\theta : \textit{not rejected}\}$

Recall: Confidence Sets by Inverting a Test

- For each θ: test H_0: $\theta_{true} = \theta$ at level α
- Test statistic $T = T(\mathcal{Y}, \theta)$.
- Reject H_0 if $T > q(\theta)$ where $P_\theta(T > q(\theta)) = \alpha$.
- Invert: $C = \{\theta : \textit{not rejected}\}$
- then

$$P_\theta(\theta \in C) = 1 - \alpha$$

  for all θ

Inverting a Test

Inverting a Test

- *p*-value version:

$$C = \{\theta : \; p(\theta) \geq \alpha\}$$

where

$$p(\theta) = P_\theta\left(T(\mathcal{Y}(\theta), \theta) \geq T(\mathcal{Y}_{obs}, \theta) \right)$$

and $\mathcal{Y}(\theta) \sim P_\theta$.

Inverting a Test

- *p*-value version:
$$C = \{\theta : \; p(\theta) \geq \alpha\}$$
where
$$p(\theta) = P_\theta\Bigg(T(\mathcal{Y}(\theta), \theta) \geq T(\mathcal{Y}_{obs}, \theta) \Bigg)$$
and $\mathcal{Y}(\theta) \sim P_\theta$.
- Quantile version:
$$C = \{\theta : T(\mathcal{Y}_{obs}, \theta) \leq q(\theta)\}$$
where
$$P_\theta\Bigg(T(\mathcal{Y}(\theta), \theta) \leq q(\theta) \Bigg) = 1 - \alpha.$$

Inverting a Test

- *p*-value version:

$$C = \{\theta : \; p(\theta) \geq \alpha\}$$

where

$$p(\theta) = P_\theta\Bigg(T(\mathcal{Y}(\theta), \theta) \geq T(\mathcal{Y}_{obs}, \theta)\Bigg)$$

and $\mathcal{Y}(\theta) \sim P_\theta$.

- Quantile version:

$$C = \{\theta : T(\mathcal{Y}_{obs}, \theta) \leq q(\theta)\}$$

where

$$P_\theta\Bigg(T(\mathcal{Y}(\theta), \theta) \leq q(\theta)\Bigg) = 1 - \alpha.$$

- Dalmasso et al (2020, 2024) proposed using simulation to do this

SBI Confidence Sets: Version 1: inverting a test

SBI Confidence Sets: Version 1: inverting a test

- Draw $\theta_1, \ldots, \theta_N \sim \pi$. (Not a prior!)

SBI Confidence Sets: Version 1: inverting a test

- Draw $\theta_1, \ldots, \theta_N \sim \pi$. (Not a prior!)
- Draw $\mathcal{Y}(\theta_j) \sim P_{\theta_j}$.

SBI Confidence Sets: Version 1: inverting a test

- Draw $\theta_1, \ldots, \theta_N \sim \pi$. (Not a prior!)
- Draw $\mathcal{Y}(\theta_j) \sim P_{\theta_j}$.
- Let $T_j = T(\mathcal{Y}(\theta_j), \theta_j)$. (Could be one dataset or many.)

SBI Confidence Sets: Version 1: inverting a test

- Draw $\theta_1, \ldots, \theta_N \sim \pi$. (Not a prior!)
- Draw $\mathcal{Y}(\theta_j) \sim P_{\theta_j}$.
- Let $T_j = T(\mathcal{Y}(\theta_j), \theta_j)$. (Could be one dataset or many.)
- Define $Z_j = I\left(T(\mathcal{Y}_{obs}, \theta_j) \geq T(\mathcal{Y}(\theta_j), \theta_j) \right)$.

SBI Confidence Sets: Version 1: inverting a test

- Draw $\theta_1, \ldots, \theta_N \sim \pi$. (Not a prior!)
- Draw $\mathcal{Y}(\theta_j) \sim P_{\theta_j}$.
- Let $T_j = T(\mathcal{Y}(\theta_j), \theta_j)$. (Could be one dataset or many.)
- Define $Z_j = I\Bigg(T(\mathcal{Y}_{obs}, \theta_j) \geq T(\mathcal{Y}(\theta_j), \theta_j) \Bigg)$.
- We have $(\theta_1, Z_1), \ldots, (\theta_N, Z_N)$.

SBI Confidence Sets: Version 1: inverting a test

- Draw $\theta_1, \ldots, \theta_N \sim \pi$. (Not a prior!)
- Draw $\mathcal{Y}(\theta_j) \sim P_{\theta_j}$.
- Let $T_j = T(\mathcal{Y}(\theta_j), \theta_j)$. (Could be one dataset or many.)
- Define $Z_j = I\Bigg(T(\mathcal{Y}_{obs}, \theta_j) \geq T(\mathcal{Y}(\theta_j), \theta_j) \Bigg)$.
- We have $(\theta_1, Z_1), \ldots, (\theta_N, Z_N)$.
- Regress Z_j on θ_j (nonparametric regression) to get

$$p(\theta_j) = \mathbb{E}[Z_j | \theta_j]$$

which is the p-value for testing

$$H_0 : \theta = \theta_j.$$

SBI Confidence Sets: Version 1: inverting a test

- Draw $\theta_1, \ldots, \theta_N \sim \pi$. (Not a prior!)
- Draw $\mathcal{Y}(\theta_j) \sim P_{\theta_j}$.
- Let $T_j = T(\mathcal{Y}(\theta_j), \theta_j)$. (Could be one dataset or many.)
- Define $Z_j = I\Bigg(T(\mathcal{Y}_{obs}, \theta_j) \geq T(\mathcal{Y}(\theta_j), \theta_j) \Bigg)$.
- We have $(\theta_1, Z_1), \ldots, (\theta_N, Z_N)$.
- Regress Z_j on θ_j (nonparametric regression) to get

  $$p(\theta_j) = \mathbb{E}[Z_j | \theta_j]$$

  which is the p-value for testing

  $$H_0 : \theta = \theta_j.$$

- Invert the test:

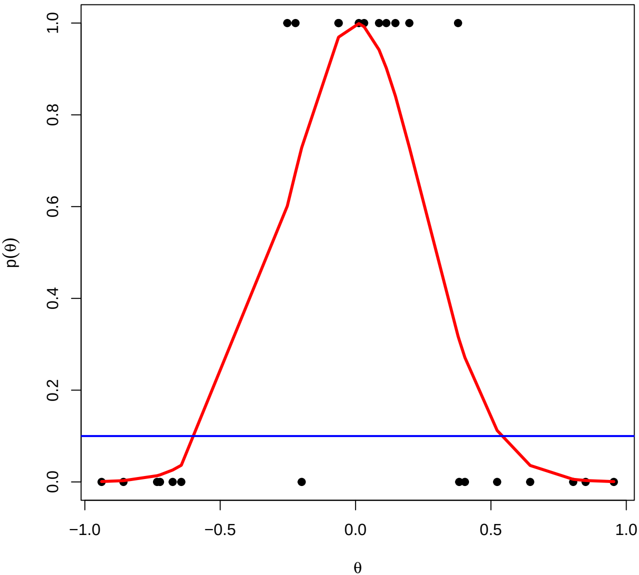
  $$C = \{\theta : \ \widehat{p}(\theta) \geq \alpha\}.$$

SBI Confidence Sets

θ	θ_1	θ_2	$\cdots$	θ_N
$\mathcal{Y}$	$\mathcal{Y}(\theta_1)$	$\mathcal{Y}(\theta_2)$	$\cdots$	$\mathcal{Y}(\theta_N)$
$T(\mathcal{Y}(\theta), \theta)$	T_1	T_2	$\cdots$	T_N
Z	Z_1	Z_2	$\cdots$	Z_N
$\widehat{p}$	$\widehat{p}(\theta_1)$	$\widehat{p}(\theta_2)$	$\cdots$	$\widehat{p}(\theta_N)$

$$C = \left\{\theta : \ \widehat{p}(\mathcal{Y}_{obs}, \theta) \geq 1 - \alpha\right\}.$$

p-value Version



SBI Confidence Sets: version 2: Quantile Regression

SBI Confidence Sets: version 2: Quantile Regression

- Draw $\theta_1, \ldots, \theta_N \sim \pi$.

SBI Confidence Sets: version 2: Quantile Regression

- Draw $\theta_1, \ldots, \theta_N \sim \pi$.
- Draw $\mathcal{Y}(\theta_j) \sim P_{\theta_j}$.

SBI Confidence Sets: version 2: Quantile Regression

- Draw $\theta_1, \ldots, \theta_N \sim \pi$.
- Draw $\mathcal{Y}(\theta_j) \sim P_{\theta_j}$.
- Let $T_j = T(\mathcal{Y}(\theta_j), \theta_j)$.

SBI Confidence Sets: version 2: Quantile Regression

- Draw $\theta_1, \ldots, \theta_N \sim \pi$.
- Draw $\mathcal{Y}(\theta_j) \sim P_{\theta_j}$.
- Let $T_j = T(\mathcal{Y}(\theta_j), \theta_j)$.
- Now we have: $(\theta_1, T_1), \ldots, (\theta_N, T_N)$

SBI Confidence Sets: version 2: Quantile Regression

- Draw $\theta_1, \ldots, \theta_N \sim \pi$.
- Draw $\mathcal{Y}(\theta_j) \sim P_{\theta_j}$.
- Let $T_j = T(\mathcal{Y}(\theta_j), \theta_j)$.
- Now we have: $(\theta_1, T_1), \ldots, (\theta_N, T_N)$
- Perform quantile regression of T_j on θ_j to estimate $q(\theta)$ where

$$P_\theta\big(T(\mathcal{Y}(\theta), \theta) \leq q(\theta)\big) = 1 - \alpha.$$

SBI Confidence Sets: version 2: Quantile Regression

- Draw $\theta_1, \ldots, \theta_N \sim \pi$.
- Draw $\mathcal{Y}(\theta_j) \sim P_{\theta_j}$.
- Let $T_j = T(\mathcal{Y}(\theta_j), \theta_j)$.
- Now we have: $(\theta_1, T_1), \ldots, (\theta_N, T_N)$
- Perform quantile regression of T_j on θ_j to estimate $q(\theta)$ where

$$P_\theta\big(T(\mathcal{Y}(\theta), \theta) \leq q(\theta)\big) = 1 - \alpha.$$

- Return $C = \Big\{\theta : \ T(\mathcal{Y}_{obs}, \theta) \leq \widehat{q}(\theta)\Big\}$.

Quantile Regression

Quantile Regression

- Does not have to be done using neural nets.

Quantile Regression

- Does not have to be done using neural nets.
- Local linear quantile regression: $\widehat{q}(\theta) = \widehat{\beta}_0$ where we minimize

$$\sum_j \rho(T_j - \beta_0 - \beta^T \theta_j) K_h(\theta_j - \theta)$$

where K_h is a kernel with bandwidth h and and ρ is the check loss:

$$\rho(u) = u(1 - \alpha - I(u < 0)).$$

Quantile Regression

- Does not have to be done using neural nets.
- Local linear quantile regression: $\widehat{q}(\theta) = \widehat{\beta}_0$ where we minimize

$$\sum_j \rho(T_j - \beta_0 - \beta^T \theta_j) K_h(\theta_j - \theta)$$

where K_h is a kernel with bandwidth h and and ρ is the check loss:

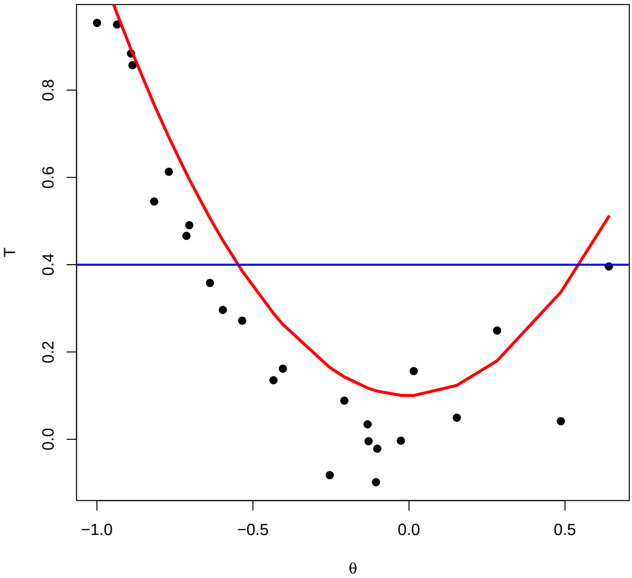
$$\rho(u) = u(1 - \alpha - I(u < 0)).$$

- This is easy and only has one tuning parameter h. And we can get standard errors for $\widehat{q}(\theta)$ easily.

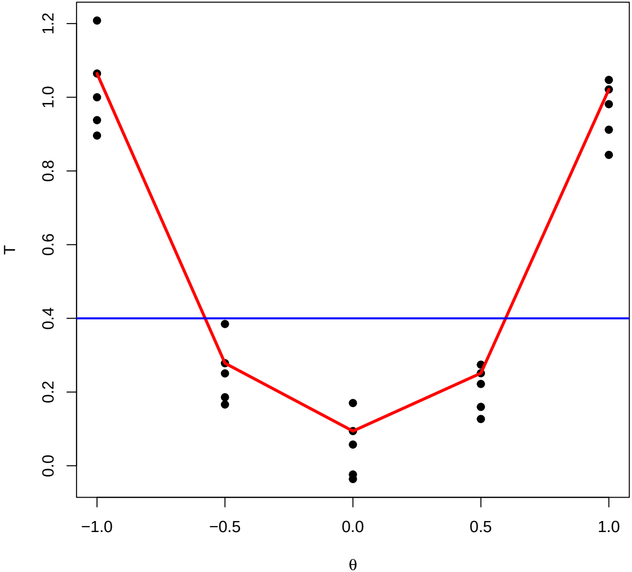
SBI Confidence Sets

θ	θ_1	θ_2	$\cdots$	θ_N
$\mathcal{Y}$	$\mathcal{Y}(\theta_1)$	$\mathcal{Y}(\theta_2)$	$\cdots$	$\mathcal{Y}(\theta_N)$
$T(\mathcal{Y}(\theta), \theta)$	T_1	T_2	$\cdots$	T_N
$\widehat{q}$	$\widehat{q}(\theta_1)$	$\widehat{q}(\theta_2)$	$\cdots$	$\widehat{q}(\theta_N)$

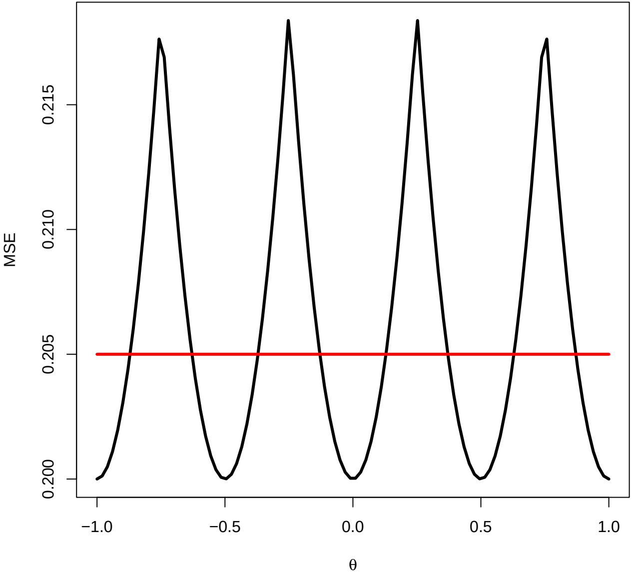
$$C = \left\{ \theta : \ \widehat{T}(\mathcal{Y}_{obs}, \theta) \leq \widehat{q}(\theta) \right\}$$


0.0
0.2
0.4
0.6
0.8
T
−1.0
−0.5
0.0
0.5
θ

With Repetition



MSE with and without repetition



SBI Confidence Sets

SBI Confidence Sets

- No regularity conditions on the model.

SBI Confidence Sets

- No regularity conditions on the model.
- No need to use asymptotic approximations.

SBI Confidence Sets

- No regularity conditions on the model.
- No need to use asymptotic approximations.
- To clarify: the coverage is

$$P_\theta(\theta \in C) = 1 - \alpha + O_P\left(\frac{1}{N}\right)^{\frac{\gamma}{2\gamma+d}}$$

where γ is the smoothness of $q(\theta)$ and d is the dimension of θ. Note that it is N (number of simulated θ_j's) not n (number of data points).

SBI Confidence Sets

- No regularity conditions on the model.
- No need to use asymptotic approximations.
- To clarify: the coverage is

$$P_\theta(\theta \in C) = 1 - \alpha + O_P\left(\frac{1}{N}\right)^{\frac{\gamma}{2\gamma+d}}$$

  where γ is the smoothness of $q(\theta)$ and d is the dimension of θ. Note that it is N (number of simulated θ_j's) not n (number of data points).
- The statistic $T(\mathcal{Y}, \theta)$ can be anything. This gives us great freedom.

SBI Confidence Sets

- No regularity conditions on the model.
- No need to use asymptotic approximations.
- To clarify: the coverage is

$$P_\theta(\theta \in C) = 1 - \alpha + O_P\left(\frac{1}{N}\right)^{\frac{\gamma}{2\gamma+d}}$$

  where γ is the smoothness of $q(\theta)$ and d is the dimension of θ. Note that it is N (number of simulated θ_j's) not n (number of data points).
- The statistic $T(\mathcal{Y}, \theta)$ can be anything. This gives us great freedom.
- Taking $T(\mathcal{Y}, \theta)$ to be the likelihood $\mathcal{L}(\theta, \mathcal{Y}) = p_\theta(\mathcal{Y})$ is common but not necessary. Not always the best choice.

SBI Confidence Sets

- No regularity conditions on the model.
- No need to use asymptotic approximations.
- To clarify: the coverage is

$$P_\theta(\theta \in C) = 1 - \alpha + O_P\left(\frac{1}{N}\right)^{\frac{\gamma}{2\gamma+d}}$$

  where γ is the smoothness of $q(\theta)$ and d is the dimension of θ. Note that it is N (number of simulated θ_j's) not n (number of data points).
- The statistic $T(\mathcal{Y}, \theta)$ can be anything. This gives us great freedom.
- Taking $T(\mathcal{Y}, \theta)$ to be the likelihood $\mathcal{L}(\theta, \mathcal{Y}) = p_\theta(\mathcal{Y})$ is common but not necessary. Not always the best choice.
- Can include prior information while retaining coverage (later).

Estimating the Likelihood

θ	θ_1	θ_2	$\cdots$	θ_N	θ_{N+1}	θ_{N+2}	$\cdots$	θ_{2N}
$\mathcal{Y}$	$\mathcal{Y}(\theta_1)$	$\mathcal{Y}(\theta_2)$	$\cdots$	$\mathcal{Y}(\theta_N)$	$\mathcal{Y}(\theta_1)$	$\mathcal{Y}(\theta_2)$	$\cdots$	$\mathcal{Y}(\theta_N)$
W	1	1	$\cdots$	1	0	0	$\cdots$	0

Estimating the Likelihood

θ	θ_1	θ_2	$\cdots$	θ_N	θ_{N+1}	θ_{N+2}	$\cdots$	θ_{2N}
$\mathcal{Y}$	$\mathcal{Y}(\theta_1)$	$\mathcal{Y}(\theta_2)$	$\cdots$	$\mathcal{Y}(\theta_N)$	$\mathcal{Y}(\theta_1)$	$\mathcal{Y}(\theta_2)$	$\cdots$	$\mathcal{Y}(\theta_N)$
W	1	1	$\cdots$	1	0	0	$\cdots$	0

- $\theta_{N+1}, \ldots, \theta_{2N}$ are a permuted version of $\theta_1, \ldots, \theta_N$

Estimating the Likelihood

θ	θ_1	θ_2	$\cdots$	θ_N	θ_{N+1}	θ_{N+2}	$\cdots$	θ_{2N}
$\mathcal{Y}$	$\mathcal{Y}(\theta_1)$	$\mathcal{Y}(\theta_2)$	$\cdots$	$\mathcal{Y}(\theta_N)$	$\mathcal{Y}(\theta_1)$	$\mathcal{Y}(\theta_2)$	$\cdots$	$\mathcal{Y}(\theta_N)$
W	1	1	$\cdots$	1	0	0	$\cdots$	0

- $\theta_{N+1}, \ldots, \theta_{2N}$ are a permuted version of $\theta_1, \ldots, \theta_N$
- Now do binary regression:

$$h(\theta, \mathcal{Y}) = P(W = 1 | \theta, \mathcal{Y}).$$

Note: binary regression not classification.

Estimating the Likelihood

θ	θ_1	θ_2	$\cdots$	θ_N	θ_{N+1}	θ_{N+2}	$\cdots$	θ_{2N}
$\mathcal{Y}$	$\mathcal{Y}(\theta_1)$	$\mathcal{Y}(\theta_2)$	$\cdots$	$\mathcal{Y}(\theta_N)$	$\mathcal{Y}(\theta_1)$	$\mathcal{Y}(\theta_2)$	$\cdots$	$\mathcal{Y}(\theta_N)$
W	1	1	$\cdots$	1	0	0	$\cdots$	0

- $\theta_{N+1}, \ldots, \theta_{2N}$ are a permuted version of $\theta_1, \ldots, \theta_N$
- Now do binary regression:

$$h(\theta, \mathcal{Y}) = P(W = 1 | \theta, \mathcal{Y}).$$

  Note: binary regression not classification.
- Then

$$\mathcal{L}(\theta, \mathcal{Y}) \propto \frac{h(\theta, \mathcal{Y})}{1 - h(\theta, \mathcal{Y})}.$$

Choice of Statistic

Choice of Statistic

- A virtue of SBI based confidence sets is that we can use any statistic $T(\mathcal{Y}, \theta)$.

Choice of Statistic

- A virtue of SBI based confidence sets is that we can use any statistic $T(\mathcal{Y}, \theta)$.
- The maximum likelihood estimate is optimal for large sample sizes and under very strict regularity conditions.

Choice of Statistic

- A virtue of SBI based confidence sets is that we can use any statistic $T(\mathcal{Y}, \theta)$.
- The maximum likelihood estimate is optimal for large sample sizes and under very strict regularity conditions.
- But in general, the likelihood is not optimal.

Choice of Statistic

- A virtue of SBI based confidence sets is that we can use any statistic $T(\mathcal{Y}, \theta)$.
- The maximum likelihood estimate is optimal for large sample sizes and under very strict regularity conditions.
- But in general, the likelihood is not optimal.
- Likelihood inference is very sensitive to model misspecification.

Choice of Statistic

- A virtue of SBI based confidence sets is that we can use any statistic $T(\mathcal{Y}, \theta)$.
- The maximum likelihood estimate is optimal for large sample sizes and under very strict regularity conditions.
- But in general, the likelihood is not optimal.
- Likelihood inference is very sensitive to model misspecification.
- More on this later.

Choice of Statistic

Choice of Statistic

- Carzon et al (2025) suggested

$$T(\mathcal{Y}, \theta) = \frac{\int \mathcal{L}(\psi, \mathcal{Y}) f(\psi) d\psi}{\mathcal{L}(\mathcal{Y}, \theta)}$$

Choice of Statistic

- Carzon et al (2025) suggested

$$T(\mathcal{Y}, \theta) = \frac{\int \mathcal{L}(\psi, \mathcal{Y}) f(\psi) d\psi}{\mathcal{L}(\mathcal{Y}, \theta)}$$

- Here, the prior f allows us to focus on part of the parameter space.

Choice of Statistic

- Carzon et al (2025) suggested

$$T(\mathcal{Y}, \theta) = \frac{\int \mathcal{L}(\psi, \mathcal{Y}) f(\psi) d\psi}{\mathcal{L}(\mathcal{Y}, \theta)}$$

- Here, the prior f allows us to focus on part of the parameter space.
- Bayesian flavor, but still has frequentist coverage:

$$\inf_{\theta} P_{\theta}(\theta \in C) = 1 - \alpha$$

Choice of Statistic

- Carzon et al (2025) suggested

$$T(\mathcal{Y}, \theta) = \frac{\int \mathcal{L}(\psi, \mathcal{Y}) f(\psi) d\psi}{\mathcal{L}(\mathcal{Y}, \theta)}$$

- Here, the prior f allows us to focus on part of the parameter space.
- Bayesian flavor, but still has frequentist coverage:

$$\inf_{\theta} P_{\theta}(\theta \in C) = 1 - \alpha$$

- Smaller confidence intervals if f is focused near the true value.

Choice of Statistic

- Carzon et al (2025) suggested

$$T(\mathcal{Y}, \theta) = \frac{\int \mathcal{L}(\psi, \mathcal{Y}) f(\psi) d\psi}{\mathcal{L}(\mathcal{Y}, \theta)}$$

- Here, the prior f allows us to focus on part of the parameter space.
- Bayesian flavor, but still has frequentist coverage:

$$\inf_{\theta} P_\theta(\theta \in C) = 1 - \alpha$$

- Smaller confidence intervals if f is focused near the true value.
- These can be seen as a SBI version of FAB (Frequentist Assisted Bayes); see also Hoff (2020, 2023).

Choice of Statistic

Choice of Statistic

- Masserano et al (2023) introduced WALDO

Choice of Statistic

- Masserano et al (2023) introduced WALDO
- Use posterior (based on a prior)

$$T = (\widehat{\theta} - \theta)^T V^{-1} (\widehat{\theta} - \theta)$$

Choice of Statistic

- Masserano et al (2023) introduced WALDO
- Use posterior (based on a prior)

$$T = (\widehat{\theta} - \theta)^T V^{-1} (\widehat{\theta} - \theta)$$

- Here:

$$\widehat{\theta} = \mathbb{E}[\theta | data]$$
$$V = \mathrm{Var}[\theta | data]$$

Choice of Statistic

- Masserano et al (2023) introduced WALDO
- Use posterior (based on a prior)

$$T = (\widehat{\theta} - \theta)^T V^{-1} (\widehat{\theta} - \theta)$$

- Here:

$$\widehat{\theta} = \mathbb{E}[\theta | data]$$
$$V = \mathrm{Var}[\theta | data]$$

- Now use the SBI algorithm with this statistic

Choice of Statistic

- Masserano et al (2023) introduced WALDO
- Use posterior (based on a prior)

$$T = (\widehat{\theta} - \theta)^T V^{-1} (\widehat{\theta} - \theta)$$

- Here:

$$\widehat{\theta} = \mathbb{E}[\theta | data]$$
$$V = \mathrm{Var}[\theta | data]$$

- Now use the SBI algorithm with this statistic
- Allows prior information but preserves frequentist coverage

Nuisance Parameters

Nuisance Parameters

- Often $\theta = (\psi, \gamma)$ where ψ is the parameter of interest and γ is a nuisance parameters.

Nuisance Parameters

- Often $\theta = (\psi, \gamma)$ where ψ is the parameter of interest and γ is a nuisance parameters.
- Can use profile likelihood $\sup_\gamma \mathcal{L}(\psi, \gamma)$.

Nuisance Parameters

- Often $\theta = (\psi, \gamma)$ where ψ is the parameter of interest and γ is a nuisance parameters.
- Can use profile likelihood $\sup_\gamma \mathcal{L}(\psi, \gamma)$.
- Can use integrated (focused) likelihood $\int \mathcal{L}(\psi, \gamma) f(\gamma) d\gamma$.

Nuisance Parameters

- Often $\theta = (\psi, \gamma)$ where ψ is the parameter of interest and γ is a nuisance parameters.
- Can use profile likelihood $\sup_\gamma \mathcal{L}(\psi, \gamma)$.
- Can use integrated (focused) likelihood $\int \mathcal{L}(\psi, \gamma) f(\gamma) d\gamma$.
- Projection: find confidence set B for (ψ, γ) and take $C = \{\psi : (\psi, \gamma) \in B \text{ for some } \gamma\}$.

Nuisance Parameters

- Often $\theta = (\psi, \gamma)$ where ψ is the parameter of interest and γ is a nuisance parameters.
- Can use profile likelihood $\sup_\gamma \mathcal{L}(\psi, \gamma)$.
- Can use integrated (focused) likelihood $\int \mathcal{L}(\psi, \gamma) f(\gamma) d\gamma$.
- Projection: find confidence set B for (ψ, γ) and take $C = \{\psi : \ (\psi, \gamma) \in B \text{ for some } \gamma\}$.
- Berger-Boos (1994): first infer nuisance parameter and use restricted projection. See Stanley et al (2025).

Diagnostic

Diagnostic

- Draw new samples $(\theta'_1, \mathcal{Y}(\theta'_1)), \ldots, (\theta'_B, \mathcal{Y}(\theta'_B))$.

Diagnostic

- Draw new samples $(\theta'_1, \mathcal{Y}(\theta'_1)), \ldots, (\theta'_B, \mathcal{Y}(\theta'_B))$.
- Use $W_j = I(\theta'_j \in C'_j)$ to estimate the coverage $P_\theta(\theta \in C)$.

Diagnostic

- Draw new samples $(\theta'_1, \mathcal{Y}(\theta'_1)), \ldots, (\theta'_B, \mathcal{Y}(\theta'_B))$.
- Use $W_j = I(\theta'_j \in C'_j)$ to estimate the coverage $P_\theta(\theta \in C)$.
- This can also be used to assess other features of the method such as size of the confidence sets.

Diagnostic

- Draw new samples $(\theta'_1, \mathcal{Y}(\theta'_1)), \ldots, (\theta'_B, \mathcal{Y}(\theta'_B))$.
- Use $W_j = I(\theta'_j \in C'_j)$ to estimate the coverage $P_\theta(\theta \in C)$.
- This can also be used to assess other features of the method such as size of the confidence sets.
- We could also use this to help choose between different test statistics.

Spatial Statistics I

Walchessen, Lenzi, Kuusela 2024

Spatial Statistics I

Walchessen, Lenzi, Kuusela 2024

- Spatial models can have intractable likelihood functions.

Spatial Statistics I

Walchessen, Lenzi, Kuusela 2024

- Spatial models can have intractable likelihood functions.
- Set of spatial locations $\mathcal{S}$.

Spatial Statistics I

Walchessen, Lenzi, Kuusela 2024

- Spatial models can have intractable likelihood functions.
- Set of spatial locations $\mathcal{S}$.
- Process $\{Y(s): \ s \in \mathcal{S}\}$.

Spatial Statistics I

Walchessen, Lenzi, Kuusela 2024

- Spatial models can have intractable likelihood functions.
- Set of spatial locations $\mathcal{S}$.
- Process $\{Y(s) : \ s \in \mathcal{S}\}$.
- Two examples: Gaussian process and the Brown-Resnick process.

Spatial Statistics I

Walchessen, Lenzi, Kuusela 2024

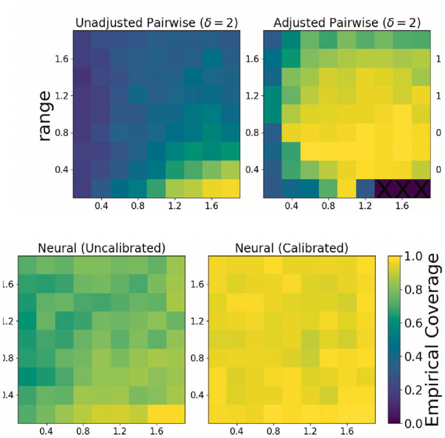
- Spatial models can have intractable likelihood functions.
- Set of spatial locations $\mathcal{S}$.
- Process $\{Y(s) : \ s \in \mathcal{S}\}$.
- Two examples: Gaussian process and the Brown-Resnick process.
- Even the Gaussian case can be hard since exact likelihood calculations involve inverting a large matrix

Spatial Statistics I

Walchessen, Lenzi, Kuusela 2024

- Spatial models can have intractable likelihood functions.
- Set of spatial locations $\mathcal{S}$.
- Process $\{Y(s) : \; s \in \mathcal{S}\}$.
- Two examples: Gaussian process and the Brown-Resnick process.
- Even the Gaussian case can be hard since exact likelihood calculations involve inverting a large matrix
- The Brown-Resnick process is intractable

Coverage Results



Spatial Statistics II

Walchessen, Zammit-Mangion, Huser, Kuusela (2025)

Spatial Statistics II

Walchessen, Zammit-Mangion, Huser, Kuusela (2025)

- Now suppose we want to simulate new observations from the estimated process.

Spatial Statistics II

Walchessen, Zammit-Mangion, Huser, Kuusela (2025)

- Now suppose we want to simulate new observations from the estimated process.
- Want to draw

$$Y \sim p(y|Y; \widehat{\theta})$$

Spatial Statistics II

Walchessen, Zammit-Mangion, Huser, Kuusela (2025)

- Now suppose we want to simulate new observations from the estimated process.
- Want to draw
$$Y \sim p(y|Y; \widehat{\theta})$$
- Method: diffusion model

Spatial Statistics II

Walchessen, Zammit-Mangion, Huser, Kuusela (2025)

- Now suppose we want to simulate new observations from the estimated process.
- Want to draw

$$Y \sim p(y|Y;\widehat{\theta})$$

- Method: diffusion model
- Have data $Y_1, \ldots, Y_n$

Spatial Statistics II

Walchessen, Zammit-Mangion, Huser, Kuusela (2025)

- Now suppose we want to simulate new observations from the estimated process.
- Want to draw

$$Y \sim p(y|Y;\widehat{\theta})$$

- Method: diffusion model
- Have data $Y_1, \ldots, Y_n$
- Evolve the data to noise $Z_1, \ldots, Z_n$

Spatial Statistics II

Walchessen, Zammit-Mangion, Huser, Kuusela (2025)

- Now suppose we want to simulate new observations from the estimated process.
- Want to draw
$$Y \sim p(y|Y; \widehat{\theta})$$
- Method: diffusion model
- Have data $Y_1, \ldots, Y_n$
- Evolve the data to noise $Z_1, \ldots, Z_n$
- Estimate the reverse process

Spatial Statistics II

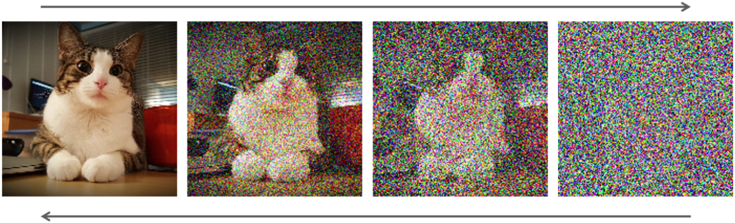
Walchessen, Zammit-Mangion, Huser, Kuusela (2025)

- Now suppose we want to simulate new observations from the estimated process.
- Want to draw

$$Y \sim p(y|Y;\widehat{\theta})$$

- Method: diffusion model
- Have data $Y_1, \ldots, Y_n$
- Evolve the data to noise $Z_1, \ldots, Z_n$
- Estimate the reverse process
- sample from noise and evolve backwards

Diffusion



Inverse Problems

Batlle et al 2024, Stanley et al 2025

Inverse Problems

Batlle et al 2024, Stanley et al 2025

- $Y = f(x) + \epsilon$, $\epsilon \sim N(0, \Sigma)$

Inverse Problems

Batlle et al 2024, Stanley et al 2025

- $Y = f(x) + \epsilon$, $\epsilon \sim N(0, \Sigma)$
- x is high dimensional

Inverse Problems

Batlle et al 2024, Stanley et al 2025

- $Y = f(x) + \epsilon$, $\epsilon \sim N(0, \Sigma)$
- x is high dimensional
- Constraints: $x \in \mathcal{X}$

Inverse Problems

Batlle et al 2024, Stanley et al 2025

- $Y = f(x) + \epsilon$, $\epsilon \sim N(0, \Sigma)$
- x is high dimensional
- Constraints: $x \in \mathcal{X}$
- Infer $\varphi(x)$

Inverse Problems

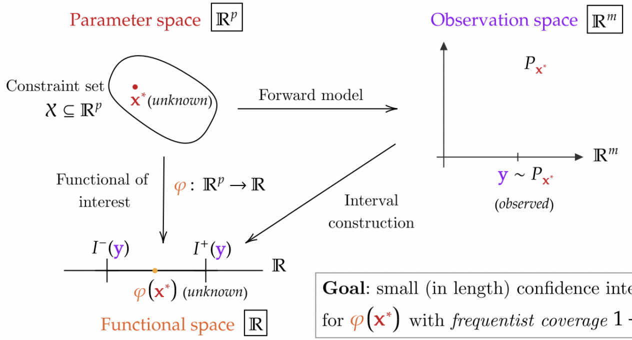
Batlle et al 2024, Stanley et al 2025

- $Y = f(x) + \epsilon$, $\epsilon \sim N(0, \Sigma)$
- x is high dimensional
- Constraints: $x \in \mathcal{X}$
- Infer $\varphi(x)$
- $\inf_{x \in \mathcal{X}} P(\varphi(x) \in C) \geq 1 - \alpha$

Inverse Problems

Batlle et al 2024, Stanley et al 2025

- $Y = f(x) + \epsilon$, $\epsilon \sim N(0, \Sigma)$
- x is high dimensional
- Constraints: $x \in \mathcal{X}$
- Infer $\varphi(x)$
- $\inf_{x \in \mathcal{X}} P(\varphi(x) \in C) \geq 1 - \alpha$
- simulation used to estimate certain quantiles



Goal: small (in length) confidence interval for $\varphi(\mathbf{x}^*)$ with *frequentist coverage* $1-\alpha$

Omnifold

Andreassen et al (2020) and Zhu et al (2024)

Omnifold

Andreassen et al (2020) and Zhu et al (2024)

- Observe

$$Y_1, \ldots, Y_n \sim p(y) = \int k(y|x) f(x) dx$$

Omnifold

Andreassen et al (2020) and Zhu et al (2024)

- Observe

$$Y_1, \ldots, Y_n \sim p(y) = \int k(y|x) f(x) dx$$

- $k(y|x)$ is unknown.

Omnifold

Andreassen et al (2020) and Zhu et al (2024)

- Observe
$$Y_1, \ldots, Y_n \sim p(y) = \int k(y|x) f(x) dx$$
- $k(y|x)$ is unknown.
- Also observe $(X_1^*, Y_1^*), \ldots, (X_N^*, Y_N^*) \sim k(y|x) r(x)$

Omnifold

Andreassen et al (2020) and Zhu et al (2024)

- Observe
$$Y_1, \ldots, Y_n \sim p(y) = \int k(y|x) f(x) dx$$
- $k(y|x)$ is unknown.
- Also observe $(X_1^*, Y_1^*), \ldots, (X_N^*, Y_N^*) \sim k(y|x) r(x)$
- Want to infer $f(x)$ or $\nu(x) = f(x)/r(x)$.

Omnifold

Andreassen et al (2020) and Zhu et al (2024)

- Observe
$$Y_1, \ldots, Y_n \sim p(y) = \int k(y|x) f(x) dx$$
- $k(y|x)$ is unknown.
- Also observe $(X_1^*, Y_1^*), \ldots, (X_N^*, Y_N^*) \sim k(y|x) r(x)$
- Want to infer $f(x)$ or $\nu(x) = f(x)/r(x)$.
- Iterative solution (Multhei, Mainz, Schorr 1987, Kondor 1983, Shepp and Vardi 1982)

Omnifold

Andreassen et al (2020) and Zhu et al (2024)

- Observe
$$Y_1, \ldots, Y_n \sim p(y) = \int k(y|x) f(x) dx$$
- $k(y|x)$ is unknown.
- Also observe $(X_1^*, Y_1^*), \ldots, (X_N^*, Y_N^*) \sim k(y|x) r(x)$
- Want to infer $f(x)$ or $\nu(x) = f(x)/r(x)$.
- Iterative solution (Multhei, Mainz, Schorr 1987, Kondor 1983, Shepp and Vardi 1982)
- $$f^{(k+1)}(x) = f^{(k)}(x) \int \frac{p(y)}{\int k(y|x') f^{(k)}(x') dx'} k(y|x) dy$$

Omnifold

Omnifold

- Andreassen et al (2020) invented a simulation-based version

Omnifold

- Andreassen et al (2020) invented a simulation-based version
- $r^{(k)}(y) = \frac{p(y)}{q^{(k)}(y)} \quad q^{(k)}(y) = \int \nu^{(k)}(x')k(y|x')dx'$

Omnifold

- Andreassen et al (2020) invented a simulation-based version
- $r^{(k)}(y) = \frac{p(y)}{q^{(k)}(y)} \quad q^{(k)}(y) = \int \nu^{(k)}(x')k(y|x')dx'$
- $\nu^{(k+1)}(x) = \nu^{(k)}(x)\frac{q^{(k)}(x)}{q(x)} \quad q^{(k)}(x) = \int r^{(k)}(y)k(y|x)dx$

Omnifold

- Andreassen et al (2020) invented a simulation-based version
- $r^{(k)}(y) = \frac{p(y)}{q^{(k)}(y)} \quad q^{(k)}(y) = \int \nu^{(k)}(x')k(y|x')dx'$
- $\nu^{(k+1)}(x) = \nu^{(k)}(x)\frac{q^{(k)}(x)}{q(x)} \quad q^{(k)}(x) = \int r^{(k)}(y)k(y|x)dx$
- These density ratios are estimated using classifiers

Omnifold

- Andreassen et al (2020) invented a simulation-based version
- $r^{(k)}(y) = \frac{p(y)}{q^{(k)}(y)} \quad q^{(k)}(y) = \int \nu^{(k)}(x')k(y|x')dx'$
- $\nu^{(k+1)}(x) = \nu^{(k)}(x)\frac{q^{(k)}(x)}{q(x)} \quad q^{(k)}(x) = \int r^{(k)}(y)k(y|x)dx$
- These density ratios are estimated using classifiers
- Zhu et al (2024) includes nuisance parameters

Model Mispecification

Tomasselli , Ventura, Wasserman (2025)

Model Mispecification

Tomasselli , Ventura, Wasserman (2025)

- Model $\mathcal{P} = \{p_\theta : \ \theta \in \Theta\}$.

Model Mispecification

Tomasselli , Ventura, Wasserman (2025)

- Model $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$.
- Do not assume that $P \in \mathcal{P}$.

Model Mispecification

Tomasselli , Ventura, Wasserman (2025)

- Model $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$.
- Do not assume that $P \in \mathcal{P}$.
- Choose a discrepancy $d(p, q)$.

Model Mispecification

Tomasselli , Ventura, Wasserman (2025)

- Model $\mathcal{P} = \{p_\theta : \ \theta \in \Theta\}$.
- Do not assume that $P \in \mathcal{P}$.
- Choose a discrepancy $d(p, q)$.
- Projection parameter: θ_* which minimizes $d(p, p_\theta)$ i.e.

$$\theta_* = \mathrm{argmin}_\theta d(p, p_\theta)$$

Model Mispecification

Tomasselli , Ventura, Wasserman (2025)

- Model $\mathcal{P} = \{p_\theta : \ \theta \in \Theta\}$.
- Do not assume that $P \in \mathcal{P}$.
- Choose a discrepancy $d(p, q)$.
- Projection parameter: θ_* which minimizes $d(p, p_\theta)$ i.e.

$$\theta_* = \mathrm{argmin}_\theta d(p, p_\theta)$$

- The model is an approximation

Discrepancies

Discrepancies

- Kullback-Leibler: $d(p, q) = \int p \log(p/q)$

Discrepancies

- Kullback-Leibler: $d(p, q) = \int p \log(p/q)$
- When the model is misspecified, the maximum likelihood (and Bayes) estimate converges to the Kullback-Leibler projection

Discrepancies

- Kullback-Leibler: $d(p, q) = \int p \log(p/q)$
- When the model is misspecified, the maximum likelihood (and Bayes) estimate converges to the Kullback-Leibler projection
- But this is very non-robust

Discrepancies

- Kullback-Leibler: $d(p, q) = \int p \log(p/q)$
- When the model is misspecified, the maximum likelihood (and Bayes) estimate converges to the Kullback-Leibler projection
- But this is very non-robust
- Suppose

$$p = (1 - \epsilon)N(0, 1) + \epsilon Q_a$$

where Q_a is centered at a and ϵ is tiny.

Discrepancies

- Kullback-Leibler: $d(p, q) = \int p \log(p/q)$
- When the model is misspecified, the maximum likelihood (and Bayes) estimate converges to the Kullback-Leibler projection
- But this is very non-robust
- Suppose

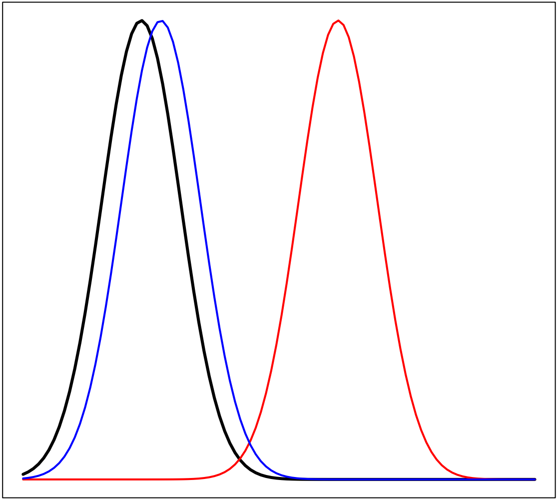
$$p = (1 - \epsilon)N(0, 1) + \epsilon Q_a$$

  where Q_a is centered at a and ϵ is tiny.
- Projection is $N(\mu(a), 1)$.

Discrepancies

- Kullback-Leibler: $d(p, q) = \int p \log(p/q)$
- When the model is misspecified, the maximum likelihood (and Bayes) estimate converges to the Kullback-Leibler projection
- But this is very non-robust
- Suppose
$$p = (1 - \epsilon)N(0, 1) + \epsilon Q_a$$
where Q_a is centered at a and ϵ is tiny.
- Projection is $N(\mu(a), 1)$.
- As $a \to \infty$, $\mu(a) \to \infty$.

Kullback-Leibler Projection



Better Discrepancies

Better Discrepancies

- Hellinger: $d(p, q) = \int (\sqrt{p} - \sqrt{q})^2$

Better Discrepancies

- Hellinger: $d(p,q) = \int (\sqrt{p} - \sqrt{q})^2$
- Density Power Divergence (DPD):

$$d(p,q) = \int \left[q^{1+\gamma}(x) - \left(1 + \frac{1}{\gamma}\right) p(x) q^{\gamma}(x) + \frac{1}{\gamma} \int p^{1+\gamma}(x) \right]$$

where $0 < \gamma \leq 1$.

Better Discrepancies

- Hellinger: $d(p,q) = \int (\sqrt{p} - \sqrt{q})^2$
- Density Power Divergence (DPD):

$$d(p,q) = \int \left[q^{1+\gamma}(x) - \left(1 + \frac{1}{\gamma}\right) p(x) q^{\gamma}(x) + \frac{1}{\gamma} \int p^{1+\gamma}(x) \right]$$

  where $0 < \gamma \leq 1$.
- Gives KL as $\gamma \to 0$.

Better Discrepancies

- Hellinger: $d(p,q) = \int(\sqrt{p} - \sqrt{q})^2$
- Density Power Divergence (DPD):

$$d(p,q) = \int \left[q^{1+\gamma}(x) - \left(1 + \frac{1}{\gamma}\right) p(x) q^{\gamma}(x) + \frac{1}{\gamma} \int p^{1+\gamma}(x) \right]$$

  where $0 < \gamma \leq 1$.
- Gives KL as $\gamma \to 0$.
- Becomes $\int(p-q)^2$ when $\gamma = 1$

Better Discrepancies

- Hellinger: $d(p,q) = \int(\sqrt{p} - \sqrt{q})^2$
- Density Power Divergence (DPD):

$$d(p,q) = \int \left[q^{1+\gamma}(x) - \left(1 + \frac{1}{\gamma}\right) p(x) q^{\gamma}(x) + \frac{1}{\gamma} \int p^{1+\gamma}(x) \right]$$

  where $0 < \gamma \leq 1$.
- Gives KL as $\gamma \to 0$.
- Becomes $\int(p-q)^2$ when $\gamma = 1$
- γ trades off efficiency vs robustness

Better Discrepancies

Better Discrepancies

- Kernel distance (MMD):

$$d^2(p, q) = \mathbb{E}[K(X, X')] - 2\mathbb{E}[K(X, Y)] + \mathbb{E}[K(Y, Y')]$$

where $K(x, y)$ is a symmetric Kernel and $X, X' \sim p$ and $Y, Y' \sim q$.

Better Discrepancies

- Kernel distance (MMD):

$$d^2(p, q) = \mathbb{E}[K(X, X')] - 2\mathbb{E}[K(X, Y)] + \mathbb{E}[K(Y, Y')]$$

  where $K(x, y)$ is a symmetric Kernel and $X, X' \sim p$ and $Y, Y' \sim q$.
- This is equivalent to

$$d^2(p, q) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}_P(f(X)) - \mathbb{E}_Q(f(Y)) \right|$$

  where $\mathcal{F}$ is a reproducing kernel Hilbert space

Better Discrepancies

- Kernel distance (MMD):

$$d^2(p,q) = \mathbb{E}[K(X,X')] - 2\mathbb{E}[K(X,Y)] + \mathbb{E}[K(Y,Y')]$$

  where $K(x,y)$ is a symmetric Kernel and $X, X' \sim p$ and $Y, Y' \sim q$.
- This is equivalent to

$$d^2(p,q) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}_P(f(X)) - \mathbb{E}_Q(f(Y)) \right|$$

  where $\mathcal{F}$ is a reproducing kernel Hilbert space

Which Discrepancy?

	robust	efficient	avoids density estimation	no tuning parameter
KL	$\times$	$\surd$	$\times$	$\surd$
Hellinger	$\surd$	$\surd$	$\times$	$\surd$
DPD	$\surd$	$\times$	$\times$	$\approx$
Kernel	$\surd$	$\times$	$\surd$	$\times$

Model Misspecification: Inference

Model Misspecification: Inference

- Recall that the projection parameter as:

$$\theta_* = \text{argmin}_\theta d(p, p_\theta).$$

Model Misspecification: Inference

- Recall that the projection parameter as:

$$\theta_* = \mathrm{argmin}_\theta d(p, p_\theta).$$

- Goal: get a confidence set for θ_*

Model Misspecification: Inference

- Recall that the projection parameter as:

$$\theta_* = \mathrm{argmin}_\theta d(p, p_\theta).$$

- Goal: get a confidence set for θ_*
- We cannot get a confidence set by inverting a test. The hypothesis $H_0 : \theta = \theta_{true}$ is always false.

Model Misspecification: Inference

- Recall that the projection parameter as:

$$\theta_* = \mathrm{argmin}_\theta d(p, p_\theta).$$

- Goal: get a confidence set for θ_*
- We cannot get a confidence set by inverting a test. The hypothesis $H_0 : \theta = \theta_{true}$ is always false.
- Inverted sets can get smaller and smaller as sample size increases. Due to rejecting all θ eventually. False impression of accuracy.

Estimation

Estimation

- To be concrete, let's use DPD with $\gamma = 1$ which is

$$d^2(p, p_\theta) = \int (p_\theta(x) - p(x))^2$$
$$= \int p_\theta^2(x)dx - 2\int p_\theta(x)p(x) + \text{constant} = \psi(\theta) + \text{constant}$$

Estimation

- To be concrete, let's use DPD with $\gamma = 1$ which is

$$\begin{aligned} d^2(p, p_\theta) &= \int (p_\theta(x) - p(x))^2 \\ &= \int p_\theta^2(x) dx - 2 \int p_\theta(x) p(x) + \text{constant} = \psi(\theta) + \text{constant} \end{aligned}$$

- Draw $W_1, \ldots, W_\ell \sim g$, and use a classifier to estimate $r_\theta(y) = p_\theta(y)/g(y)$ for a reference density g.

Estimation

- To be concrete, let's use DPD with $\gamma = 1$ which is

$$d^2(p, p_\theta) = \int (p_\theta(x) - p(x))^2$$
$$= \int p_\theta^2(x)dx - 2\int p_\theta(x)p(x) + \text{constant} = \psi(\theta) + \text{constant}$$

- Draw $W_1, \ldots, W_\ell \sim g$, and use a classifier to estimate $r_\theta(y) = p_\theta(y)/g(y)$ for a reference density g.
- Estimate:

$$\widehat{\psi}(\theta) = \frac{1}{m}\sum_i \widehat{r}_\theta(Y_i(\theta))g(Y_i(\theta)) - \frac{2}{n}\sum_i \widehat{r}_\theta(Y_i)g(Y_i).$$

Estimation

- To be concrete, let's use DPD with $\gamma = 1$ which is

$$d^2(p, p_\theta) = \int (p_\theta(x) - p(x))^2$$
$$= \int p_\theta^2(x)dx - 2\int p_\theta(x)p(x) + \text{constant} = \psi(\theta) + \text{constant}$$

- Draw $W_1, \ldots, W_\ell \sim g$, and use a classifier to estimate $r_\theta(y) = p_\theta(y)/g(y)$ for a reference density g.
- Estimate:

$$\widehat{\psi}(\theta) = \frac{1}{m}\sum_i \widehat{r}_\theta(Y_i(\theta))g(Y_i(\theta)) - \frac{2}{n}\sum_i \widehat{r}_\theta(Y_i)g(Y_i).$$

- $\widehat{\theta}_*$ minimizes $\widehat{\psi}(\theta_j)$.

Confidence Set

Confidence Set

- The usual confidence set is

$$\left\{\theta : (\widehat{\theta} - \theta)^T V^{-1} (\widehat{\theta} - \theta) \leq \chi^2_{d,\alpha}\right\}$$

where $V = n^{-1} = BMB^T$, $B = n^{-1} = \sum_i \psi_i \psi_i^T$, $A = -n^{-1} \sum_i \psi_i$ and ψ_i is gradient of discrepancy estimate.

Confidence Set

- The usual confidence set is

$$\left\{\theta : (\widehat{\theta} - \theta)^T V^{-1} (\widehat{\theta} - \theta) \leq \chi^2_{d,\alpha}\right\}$$

  where $V = n^{-1} = BMB^T$, $B = n^{-1} = \sum_i \psi_i \psi_i^T$, $A = -n^{-1} \sum_i \psi_i$ and ψ_i is gradient of discrepancy estimate.
- But this depends on regularity conditions and the derivatives might be intractable.

Relative Test Confidence Set

Park, Balakrishnan, Wasserman (2023), Takatsu and Kuchibhotla (2025) and Chang and Kuchibhotla (2024)

Relative Test Confidence Set

Park, Balakrishnan, Wasserman (2023), Takatsu and Kuchibhotla (2025) and Chang and Kuchibhotla (2024)

- Split data into $\mathcal{D}_1$ and $\mathcal{D}_2$

Relative Test Confidence Set

Park, Balakrishnan, Wasserman (2023), Takatsu and Kuchibhotla (2025) and Chang and Kuchibhotla (2024)

- Split data into $\mathcal{D}_1$ and $\mathcal{D}_2$
- $\mathcal{D}_1 \to \widehat{\theta}$.

Relative Test Confidence Set

Park, Balakrishnan, Wasserman (2023), Takatsu and Kuchibhotla (2025) and Chang and Kuchibhotla (2024)

- Split data into $\mathcal{D}_1$ and $\mathcal{D}_2$
- $\mathcal{D}_1 \to \widehat{\theta}$.
- Use $\mathcal{D}_2$ to test:
  For every θ test: $H_0 : d(p, p_\theta) \leq d(p, p_{\widehat{\theta}})$.

Relative Test Confidence Set

Park, Balakrishnan, Wasserman (2023), Takatsu and Kuchibhotla (2025) and Chang and Kuchibhotla (2024)

- Split data into $\mathcal{D}_1$ and $\mathcal{D}_2$
- $\mathcal{D}_1 \to \widehat{\theta}$.
- Use $\mathcal{D}_2$ to test:
  For every θ test: $H_0 : d(p, p_\theta) \leq d(p, p_{\widehat{\theta}})$.
- Now

$$T = \widehat{d}(p, p_\theta) - \widehat{d}(p, p_{\widehat{\theta}}) = \frac{1}{n}\sum_i W_i - \frac{1}{m}\sum_i V_i \approx N(\mu, \sigma^2_\theta)$$

Relative Test Confidence Set

Park, Balakrishnan, Wasserman (2023), Takatsu and Kuchibhotla (2025) and Chang and Kuchibhotla (2024)

- Split data into $\mathcal{D}_1$ and $\mathcal{D}_2$
- $\mathcal{D}_1 \to \widehat{\theta}$.
- Use $\mathcal{D}_2$ to test:
  For every θ test: $H_0 : d(p, p_\theta) \leq d(p, p_{\widehat{\theta}})$.
- Now

$$T = \widehat{d}(p, p_\theta) - \widehat{d}(p, p_{\widehat{\theta}}) = \frac{1}{n}\sum_i W_i - \frac{1}{m}\sum_i V_i \approx N(\mu, \sigma_\theta^2)$$

- This is $\approx$ Normal without regularity conditions on the model.

Relative Test Confidence Set

Park, Balakrishnan, Wasserman (2023), Takatsu and Kuchibhotla (2025) and Chang and Kuchibhotla (2024)

- Split data into $\mathcal{D}_1$ and $\mathcal{D}_2$
- $\mathcal{D}_1 \to \widehat{\theta}$.
- Use $\mathcal{D}_2$ to test:
  For every θ test: $H_0 : d(p, p_\theta) \leq d(p, p_{\widehat{\theta}})$.
- Now

$$T = \widehat{d}(p, p_\theta) - \widehat{d}(p, p_{\widehat{\theta}}) = \frac{1}{n}\sum_i W_i - \frac{1}{m}\sum_i V_i \approx N(\mu, \sigma^2_\theta)$$

- This is $\approx$ Normal without regularity conditions on the model.
- Reject if $T > z_\alpha \widehat{\sigma}_\theta$

Relative Test Confidence Set

Park, Balakrishnan, Wasserman (2023), Takatsu and Kuchibhotla (2025) and Chang and Kuchibhotla (2024)

- Split data into $\mathcal{D}_1$ and $\mathcal{D}_2$
- $\mathcal{D}_1 \to \widehat{\theta}$.
- Use $\mathcal{D}_2$ to test:
  For every θ test: $H_0 : d(p, p_\theta) \leq d(p, p_{\widehat{\theta}})$.
- Now

$$T = \widehat{d}(p, p_\theta) - \widehat{d}(p, p_{\widehat{\theta}}) = \frac{1}{n} \sum_i W_i - \frac{1}{m} \sum_i V_i \approx N(\mu, \sigma^2_\theta)$$

- This is $\approx$ Normal without regularity conditions on the model.
- Reject if $T > z_\alpha \widehat{\sigma}_\theta$
- $C = \{\theta : T_\theta < z_\alpha \widehat{\sigma}_\theta\}$.

Relative Test Confidence Set

Park, Balakrishnan, Wasserman (2023), Takatsu and Kuchibhotla (2025) and Chang and Kuchibhotla (2024)

- Split data into $\mathcal{D}_1$ and $\mathcal{D}_2$
- $\mathcal{D}_1 \to \widehat{\theta}$.
- Use $\mathcal{D}_2$ to test:
  For every θ test: $H_0 : d(p, p_\theta) \leq d(p, p_{\widehat{\theta}})$.
- Now

$$T = \widehat{d}(p, p_\theta) - \widehat{d}(p, p_{\widehat{\theta}}) = \frac{1}{n}\sum_i W_i - \frac{1}{m}\sum_i V_i \approx N(\mu, \sigma_\theta^2)$$

- This is $\approx$ Normal without regularity conditions on the model.
- Reject if $T > z_\alpha \widehat{\sigma}_\theta$
- $C = \{\theta : T_\theta < z_\alpha \widehat{\sigma}_\theta\}$.
- $P(\theta_* \in C) \approx 1 - \alpha$.

Example: Mixture Model

Example: Mixture Model

- $p(y) = \lambda N(\mu_1, \sigma) + (1 - \lambda) N(\mu_2, \sigma)$

Example: Mixture Model

- $p(y) = \lambda N(\mu_1, \sigma) + (1 - \lambda) N(\mu_2, \sigma)$
- This model is not identified. Regularity conditions fail.

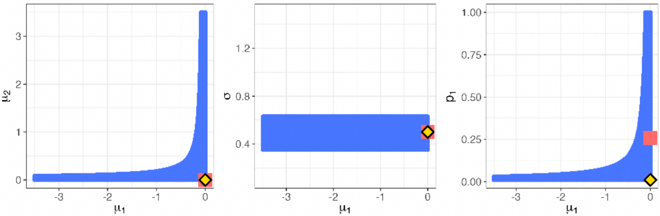
Example: Mixture Model

- $p(y) = \lambda N(\mu_1, \sigma) + (1 - \lambda) N(\mu_2, \sigma)$
- This model is not identified. Regularity conditions fail.
- Standard methods for confidence sets don't work.

Example: Mixture Model

- $p(y) = \lambda N(\mu_1, \sigma) + (1 - \lambda) N(\mu_2, \sigma)$
- This model is not identified. Regularity conditions fail.
- Standard methods for confidence sets don't work.
- Also, we want to allow of misspecification

Mixture Model: Using Discrepancy



Robustness by Tilting

- Protect from model misspecification by expanding the model.

Robustness by Tilting

- Protect from model misspecification by expanding the model.
- Choose basis functions $b_1, \ldots, b_k$.

Robustness by Tilting

- Protect from model misspecification by expanding the model.
- Choose basis functions $b_1, \ldots, b_k$.
- Expand the model p_θ to

$$p_{\theta,\beta}(y) = \frac{p_\theta(y) e^{\sum_j \beta_j b_j(y)}}{\int p_\theta(u) e^{\sum_j \beta_j b_j(u)} du}.$$

Robustness by Tilting

- Protect from model misspecification by expanding the model.
- Choose basis functions $b_1, \ldots, b_k$.
- Expand the model p_θ to

$$p_{\theta,\beta}(y) = \frac{p_\theta(y) e^{\sum_j \beta_j b_j(y)}}{\int p_\theta(u) e^{\sum_j \beta_j b_j(u)} du}.$$

- SBI can be used to get the profile likelihood

$$\mathcal{L}(\theta) = \sup_\beta \mathcal{L}(\theta, \beta)$$

Robustness by Tilting

- Protect from model misspecification by expanding the model.
- Choose basis functions $b_1, \ldots, b_k$.
- Expand the model p_θ to

$$p_{\theta,\beta}(y) = \frac{p_\theta(y) e^{\sum_j \beta_j b_j(y)}}{\int p_\theta(u) e^{\sum_j \beta_j b_j(u)} du}.$$

- SBI can be used to get the profile likelihood

$$\mathcal{L}(\theta) = \sup_\beta \mathcal{L}(\theta, \beta)$$

- Use $\mathcal{L}(\theta)$ to get confidence set for θ.

Robustness by Tilting

- Protect from model misspecification by expanding the model.
- Choose basis functions $b_1, \ldots, b_k$.
- Expand the model p_θ to

$$p_{\theta,\beta}(y) = \frac{p_\theta(y) e^{\sum_j \beta_j b_j(y)}}{\int p_\theta(u) e^{\sum_j \beta_j b_j(u)} du}.$$

- SBI can be used to get the profile likelihood

$$\mathcal{L}(\theta) = \sup_\beta \mathcal{L}(\theta, \beta)$$

- Use $\mathcal{L}(\theta)$ to get confidence set for θ.
- Requires Newton-Raphson to get $\widehat{\beta}(\theta)$ which maximizes $\sup_\beta \mathcal{L}(\theta, \beta)$ for each θ.

Model Approximation Using a Varying Coefficient Model

Model Approximation Using a Varying Coefficient Model

- When $p_\theta(y)$ is intractable, it may be useful, for interpretability, to have an approximate, closed form expression for p_θ.

Model Approximation Using a Varying Coefficient Model

- When $p_\theta(y)$ is intractable, it may be useful, for interpretability, to have an approximate, closed form expression for p_θ.
- Let $b_1, \ldots, b_k$ be basis functions.

Model Approximation Using a Varying Coefficient Model

- When $p_\theta(y)$ is intractable, it may be useful, for interpretability, to have an approximate, closed form expression for p_θ.
- Let $b_1, \ldots, b_k$ be basis functions.
- Let $f(\theta) = (f_1(\theta), \ldots, f_k(\theta))$.

Model Approximation Using a Varying Coefficient Model

- When $p_\theta(y)$ is intractable, it may be useful, for interpretability, to have an approximate, closed form expression for p_θ.
- Let $b_1, \ldots, b_k$ be basis functions.
- Let $f(\theta) = (f_1(\theta), \ldots, f_k(\theta))$.
- Define

$$p(y;\theta, f) = \sum_r f_r(\theta) b_r(y)$$

Model Approximation Using a Varying Coefficient Model

- When $p_\theta(y)$ is intractable, it may be useful, for interpretability, to have an approximate, closed form expression for p_θ.
- Let $b_1, \ldots, b_k$ be basis functions.
- Let $f(\theta) = (f_1(\theta), \ldots, f_k(\theta))$.
- Define

$$p(y; \theta, f) = \sum_r f_r(\theta) b_r(y)$$

- Find f to minimize

$$\int (p_\theta(y) - p(y; \theta, f))^2 dy.$$

Model Approximation Using a Varying Coefficient Model

- When $p_\theta(y)$ is intractable, it may be useful, for interpretability, to have an approximate, closed form expression for p_θ.
- Let $b_1, \ldots, b_k$ be basis functions.
- Let $f(\theta) = (f_1(\theta), \ldots, f_k(\theta))$.
- Define

$$p(y; \theta, f) = \sum_r f_r(\theta) b_r(y)$$

- Find f to minimize

$$\int (p_\theta(y) - p(y; \theta, f))^2 dy.$$

- Then

$$\widehat{f}(\theta_j) = B^{-1} \overline{b}_{\theta_j}$$

where

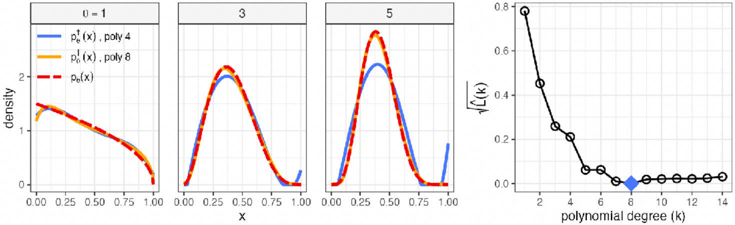
$$\overline{b}_{\theta_j} = \frac{1}{m} \sum_i b_{\theta_j}(Y_i(\theta_j)).$$

Model Approximation Using a Varying Coefficient Model

- When $p_\theta(y)$ is intractable, it may be useful, for interpretability, to have an approximate, closed form expression for p_θ.
- Let $b_1, \ldots, b_k$ be basis functions.
- Let $f(\theta) = (f_1(\theta), \ldots, f_k(\theta))$.
- Define
$$p(y; \theta, f) = \sum_r f_r(\theta) b_r(y)$$
- Find f to minimize
$$\int (p_\theta(y) - p(y; \theta, f))^2 dy.$$
- Then
$$\widehat{f}(\theta_j) = B^{-1} \overline{b}_{\theta_j}$$
where
$$\overline{b}_{\theta_j} = \frac{1}{m} \sum_i b_{\theta_j}(Y_i(\theta_j)).$$
- Then $\widehat{f}(\theta)$ is obtained from $\widehat{f}(\theta_1), \ldots, \widehat{f}(\theta_N)$ by smoothing.

Model Approximation

Red = true. Blue = approx



Active learning

Active learning

- We want to draw $\theta_1, \theta_2, \ldots,$ sequentially and zoom in on the confidence set C.

Active learning

- We want to draw $\theta_1, \theta_2, \ldots,$ sequentially and zoom in on the confidence set C.
- This is critical when θ is high dimensional.

Active learning

- We want to draw $\theta_1, \theta_2, \ldots,$ sequentially and zoom in on the confidence set C.
- This is critical when θ is high dimensional.
- Let $C = \{\theta : \ pv(\theta) \geq \alpha\}$ and $\widehat{C} = \{\theta : \ \widehat{pv}(\theta) \geq \alpha\}$.

Active learning

- We want to draw $\theta_1, \theta_2, \ldots,$ sequentially and zoom in on the confidence set C.
- This is critical when θ is high dimensional.
- Let $C = \{\theta : \; pv(\theta) \geq \alpha\}$ and $\widehat{C} = \{\theta : \; \widehat{pv}(\theta) \geq \alpha\}$.
- Let

$$P\left(I(\theta \in \widehat{C}) \neq I(\theta \in C) \right) \approx \Phi\left(-\frac{|\alpha - pv(\theta)|}{s(\theta)} \right) \equiv e(\theta).$$

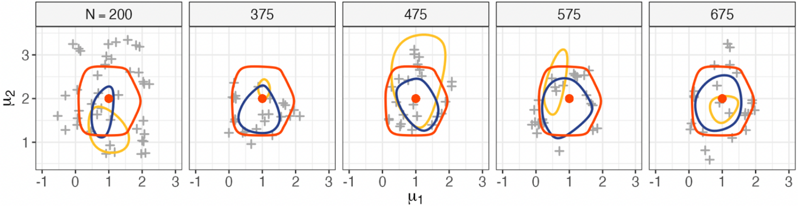
Active learning

- We want to draw $\theta_1, \theta_2, \ldots,$ sequentially and zoom in on the confidence set C.
- This is critical when θ is high dimensional.
- Let $C = \{\theta : \; pv(\theta) \geq \alpha\}$ and $\widehat{C} = \{\theta : \; \widehat{pv}(\theta) \geq \alpha\}$.
- Let

$$P\left(I(\theta \in \widehat{C}) \neq I(\theta \in C) \right) \approx \Phi\left(-\frac{|\alpha - pv(\theta)|}{s(\theta)} \right) \equiv e(\theta).$$

- Minimize R by choosing θ_{j+1} where $e(\theta)$ is large.

Example



Goodness of Fit

Goodness of Fit

- Test

$$H_0 : \inf_{\theta} d(p, p_{\theta}) = 0$$

Goodness of Fit

- Test
$$H_0 : \inf_\theta d(p, p_\theta) = 0$$
- The p-value is
$$p = \sup_\theta p(\theta)$$
where
$$p(\theta) = P_\theta(T_n(\theta) \geq T_n)$$
$$T_n(\theta) = \inf_\psi d(P_\psi, P_n(\theta)), \quad T_n = \inf_\psi d(P_\psi, P_n)$$

Goodness of Fit

- Test
$$H_0 : \inf_\theta d(p, p_\theta) = 0$$
- The p-value is
$$p = \sup_\theta p(\theta)$$
where
$$p(\theta) = P_\theta(T_n(\theta) \geq T_n)$$
$$T_n(\theta) = \inf_\psi d(P_\psi, P_n(\theta)), \quad T_n = \inf_\psi d(P_\psi, P_n)$$
- Getting the critical value (while allowing for non-regularity) is difficult in general.

Goodness of Fit

- Test
$$H_0 : \inf_{\theta} d(p, p_{\theta}) = 0$$
- The p-value is
$$p = \sup_{\theta} p(\theta)$$
where
$$p(\theta) = P_{\theta}(T_n(\theta) \geq T_n)$$
$$T_n(\theta) = \inf_{\psi} d(P_{\psi}, P_n(\theta)), \quad T_n = \inf_{\psi} d(P_{\psi}, P_n)$$
- Getting the critical value (while allowing for non-regularity) is difficult in general.
- This can be SBI-ified.

Open Questions and Challenges

Open Questions and Challenges

- High-dimensional parameter space Θ. We can do high dimensional quantile regression but we need to know where to look.

Open Questions and Challenges

- High-dimensional parameter space Θ. We can do high dimensional quantile regression but we need to know where to look.
- Active learning. This could be the cure for high dimensional problems.

Open Questions and Challenges

- High-dimensional parameter space Θ. We can do high dimensional quantile regression but we need to know where to look.
- Active learning. This could be the cure for high dimensional problems.
- Reducing reducing sensitivity to nuisance parameters?
  Traditionally: compute the score statistic for ψ and subtract its projection onto the score for the nuisance parameter.
  SBI?

Open Questions and Challenges

- High-dimensional parameter space Θ. We can do high dimensional quantile regression but we need to know where to look.
- Active learning. This could be the cure for high dimensional problems.
- Reducing reducing sensitivity to nuisance parameters?
  Traditionally: compute the score statistic for ψ and subtract its projection onto the score for the nuisance parameter.
  SBI?
- Choosing statistic T?

Open Questions and Challenges

- High-dimensional parameter space Θ. We can do high dimensional quantile regression but we need to know where to look.
- Active learning. This could be the cure for high dimensional problems.
- Reducing reducing sensitivity to nuisance parameters?
  Traditionally: compute the score statistic for ψ and subtract its projection onto the score for the nuisance parameter.
  SBI?
- Choosing statistic T?
- Semiparametric and nonparametric inference.

Open Questions and Challenges

- High-dimensional parameter space Θ. We can do high dimensional quantile regression but we need to know where to look.
- Active learning. This could be the cure for high dimensional problems.
- Reducing reducing sensitivity to nuisance parameters?
  Traditionally: compute the score statistic for ψ and subtract its projection onto the score for the nuisance parameter.
  SBI?
- Choosing statistic T?
- Semiparametric and nonparametric inference.

THE END