Short course on Generalized Fiducial Inference

Parts of this short course are joint work with
T. C.M. Lee (UC Davis), H. Iyer (NIST)
Randy Lai (UC Davis), J. Williams (NCSU), Y. Cui (UPenn)
T. Petty (UNC), G. Li (UNC)

Summer 2020

Jan Hannig[a]

University of North Carolina at Chapel Hill
https://hannig.cloudapps.unc.edu

[a]NSF support acknowledged

Outline

- Introduction
- Definition
- Theoretical Results
- Applications
- Conclusions

Outline

Fiducial?

- Oxford English Dictionary
  - adjective technical (of a point or line) used as a fixed basis of comparison.
  - Origin from Latin fiducia ‘trust, confidence’
- Merriam-Webster dictionary
  1. taken as standard of reference *a fiducial mark*
  2. founded on faith or trust
  3. having the nature of a trust : fiduciary

Aims

- Explain the definition of generalized fiducial distribution

Aims

- Explain the definition of generalized fiducial distribution
- Discuss theoretical results

Aims

- Explain the definition of generalized fiducial distribution
- Discuss theoretical results
- Show successful applications

Aims

- Explain the definition of generalized fiducial distribution
- Discuss theoretical results
- Show successful applications

- My point of view is frequentist
  - Justified using asymptotic theorems and simulations.
  - GFI shows very good repeated sampling performance in applications.

Long, long, long time ago...

- Probabilistic uncertainty via Bayes Theorem:

$$P(\xi|X) = \frac{f(X|\xi)\pi(\xi)}{\int_\Xi f(X|\xi)\pi(\xi)d\xi}.$$

Long, long, long time ago...

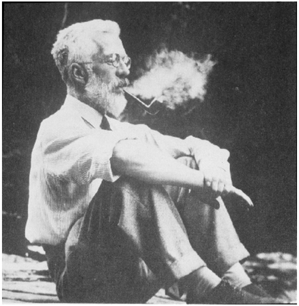
- Probabilistic uncertainty via Bayes Theorem:

$$P(\xi|X) = \frac{f(X|\xi)\pi(\xi)}{\int_{\Xi} f(X|\xi)\pi(\xi)d\xi}.$$

- Bayes-Laplace postulate:

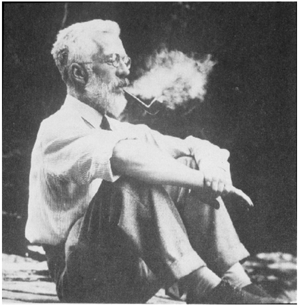
  When nothing is known about the parameter in advance, let the prior be so that all values of the parameter are equally likely.

Long, long, time ago...



“Not knowing the chance of mutually exclusive events and knowing the chance to be equal are two quite different states of knowledge” R. A. Fisher

Long, long, time ago...



"Not knowing the chance of mutually exclusive events and knowing the chance to be equal are two quite different states of knowledge" R. A. Fisher

It was a wild ride after that!

Brief history of fiducial inference

- Fisher (1922, 1930, 1935) no formal definition
- Lindley (1958) fiducial vs Bayes
- Fraser (1966) structural inference
- Dempster (1967) upper and lower probabilities
- Dawid and Stone (1982) theoretical results for simple cases.
- Barnard (1995) pivotal based methods.
- Weerahandi (1989, 1993), Krishnamoorthy generalized inference.

Fiducial Inspired Work in the New Millennium

- Dempster-Shafer calculus; Dempster (2008), Edlefsen, Liu & Dempster (2009)
- Inferential Models; Liu & Martin (2015)
- Confidence Distributions; Xie, Singh & Strawderman (2011), Schweder & Hjort (2016)
- Higher order likelihood, tangent exponential family, $r^\star$, Reid & Fraser (2010)
- Objective Bayesian inference, e.g., reference prior Berger, Bernardo & Sun (2009, 2012).
- Fiducial Inference H, Iyer & Patterson (2006), H (2009, 2013), H & Lee (2009), Taraldsen & Lindqvist (2013), Veronese & Melilli (2015), H, Iyer, Lai & Lee (2016)...

Bird's Eye View of Statistical InferenceI

Bird's Eye View of Statistical InferenceI

- Common: quantify uncertainty using adequate data generating mechanism

Bird's Eye View of Statistical Inferencel

- Common: quantify uncertainty using adequate data generating mechanism
- Difference: math details, interpretation, **replication**

Bird's Eye View of Statistical Inferencel

- Common: quantify uncertainty using adequate data generating mechanism
- Difference: math details, interpretation, **replication**
  - My subjective opinion: If the underlying optimization problem is the same, the methods are the same.

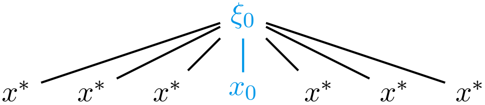
Frequentist

Frequentist

- **Modeling:** collection of distributions $\mathcal{P} = \{P_\xi\}_{\xi \in \Xi}$.

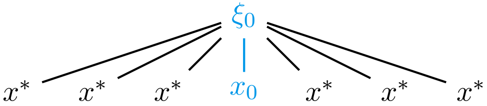
Frequentist

- **Modeling:** collection of distributions $\mathcal{P} = \{P_\xi\}_{\xi \in \Xi}$.
- **Replication:** parameter ξ_0 fixed, data $\boldsymbol{x}$ replicated



Frequentist

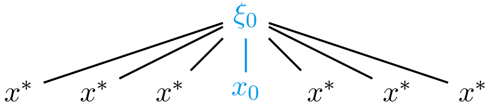
- **Modeling:** collection of distributions $\mathcal{P} = \{P_\xi\}_{\xi \in \Xi}$.
- **Replication:** parameter ξ_0 fixed, data $\boldsymbol{x}$ replicated



- **Issues:**
  - Quality judged by averaging over unobserved data $\boldsymbol{x}^*$ (SLLN + Cournot's principle)

Frequentist

- **Modeling:** collection of distributions $\mathcal{P} = \{P_\xi\}_{\xi \in \Xi}$.
- **Replication:** parameter ξ_0 fixed, data $\boldsymbol{x}$ replicated



- **Issues:**
  - Quality judged by averaging over unobserved data $\boldsymbol{x}^*$ (SLLN + Cournot's principle)
  - Each problem requires its own solution

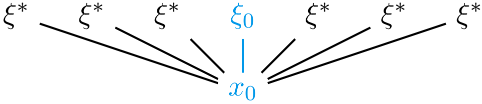
Bayesian

Bayesian

- **Modeling:** One joint distribution $f(\boldsymbol{x}|\xi) \cdot \pi(\xi)$.

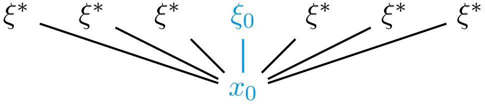
Bayesian

- **Modeling:** One joint distribution $f(\boldsymbol{x}|\xi) \cdot \pi(\xi)$.
- **Replication:** data x_0 fixed, parameter ξ replicated



Bayesian

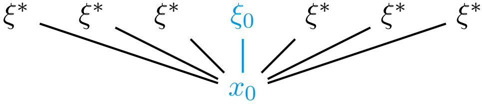
- **Modeling:** One joint distribution $f(\boldsymbol{x}|\xi) \cdot \pi(\xi)$.
- **Replication:** data x_0 fixed, parameter ξ replicated



- **Issues:**
  - Averaging over unused parameters ξ^* needs prior

Bayesian

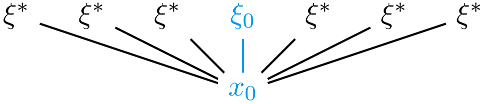
- **Modeling:** One joint distribution $f(\boldsymbol{x}|\xi) \cdot \pi(\xi)$.
- **Replication:** data x_0 fixed, parameter ξ replicated



- **Issues:**
  - Averaging over unused parameters ξ^* needs prior
  - Unique solution using Bayes theorem (conditional probability)

Bayesian

- **Modeling:** One joint distribution $f(\boldsymbol{x}|\xi) \cdot \pi(\xi)$.
- **Replication:** data x_0 fixed, parameter ξ replicated



- **Issues:**
  - Averaging over unused parameters ξ^* needs prior
  - Unique solution using Bayes theorem (conditional probability)
  - Axiomatic system for all of inference, subjective interpretation (de Finetti, Savage).

Fiducial

Fiducial

- **Modeling:** Data generating algorithm: $\boldsymbol{x} = G(\boldsymbol{u}, \xi)$

Fiducial

- **Modeling:** Data generating algorithm: $\boldsymbol{x} = G(\boldsymbol{u}, \xi)$
- **Replication:** data $\boldsymbol{x}$ & parameter ξ linked through DGA, auxiliary variable $\boldsymbol{u}$ replicated

$$
\begin{array}{cccccccc}
 & \xi^* & \xi^* & \xi^* & \xi_0 & \xi^* & \xi^* & \xi^* \\
u_0 & \vdots & \vdots & \vdots & | & \vdots & \vdots & \vdots \\
 & x^* & x^* & x^* & x_0 & x^* & x^* & x^*
\end{array}
$$

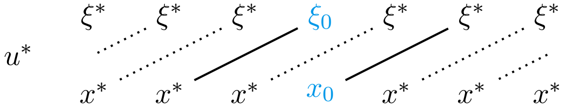
Fiducial

- **Modeling:** Data generating algorithm: $\boldsymbol{x} = G(\boldsymbol{u}, \xi)$
- **Replication:** data x & parameter ξ linked through DGA, auxiliary variable $\boldsymbol{u}$ replicated

$$
\begin{array}{cccccccc}
 & \xi^* & \xi^* & \xi^* & \xi_0 & \xi^* & \xi^* & \xi^* \\
u^* & & & & & & & \\
 & x^* & x^* & x^* & x_0 & x^* & x^* & x^*
\end{array}
$$

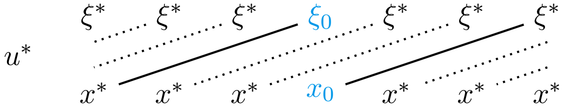
Fiducial

- **Modeling:** Data generating algorithm: $\boldsymbol{x} = G(\boldsymbol{u}, \xi)$
- **Replication:** data $\boldsymbol{x}$ & parameter ξ linked through DGA, auxiliary variable $\boldsymbol{u}$ replicated



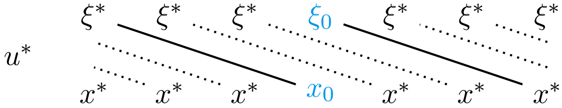
Fiducial

- **Modeling:** Data generating algorithm: $\boldsymbol{x} = G(\boldsymbol{u}, \xi)$
- **Replication:** data $\boldsymbol{x}$ & parameter ξ linked through DGA, auxiliary variable $\boldsymbol{u}$ replicated



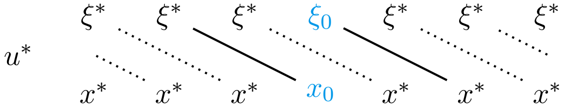
Fiducial

- **Modeling:** Data generating algorithm: $\boldsymbol{x} = G(\boldsymbol{u}, \xi)$
- **Replication:** data x & parameter ξ linked through DGA, auxiliary variable $\boldsymbol{u}$ replicated



Fiducial

- **Modeling:** Data generating algorithm: $\boldsymbol{x} = G(\boldsymbol{u}, \xi)$
- **Replication:** data $\boldsymbol{x}$ & parameter ξ linked through DGA, auxiliary variable $\boldsymbol{u}$ replicated



Fiducial

- **Modeling:** Data generating algorithm: $\boldsymbol{x} = G(\boldsymbol{u}, \xi)$
- **Replication:** data x & parameter ξ linked through DGA, auxiliary variable $\boldsymbol{u}$ replicated

$$\begin{array}{cccccccc} & \xi^* & \xi^* & \xi^* & \xi_0 & \xi^* & \xi^* & \xi^* \\ u^* & & & & & & & \\ & x^* & x^* & x^* & x_0 & x^* & x^* & x^* \end{array}$$

Fiducial

- **Modeling:** Data generating algorithm: $\boldsymbol{x} = G(\boldsymbol{u}, \xi)$
- **Replication:** data x & parameter ξ linked through DGA, auxiliary variable $\boldsymbol{u}$ replicated

$$
\begin{array}{cccccccc}
 & \xi^* & \xi^* & \xi^* & \xi_0 & \xi^* & \xi^* & \xi^* \\
u_0 & \vdots & \vdots & \vdots & | & \vdots & \vdots & \vdots \\
 & x^* & x^* & x^* & x_0 & x^* & x^* & x^*
\end{array}
$$

Fiducial

- **Modeling:** Data generating algorithm: $\boldsymbol{x} = G(\boldsymbol{u}, \xi)$
- **Replication:** data $\boldsymbol{x}$ & parameter ξ linked through DGA, auxiliary variable $\boldsymbol{u}$ replicated

$$\begin{array}{cccccccc} & \xi^* & \xi^* & \xi^* & \xi_0 & \xi^* & \xi^* & \xi^* \\ u_0 & \vdots & \vdots & \vdots & | & \vdots & \vdots & \vdots \\ & x^* & x^* & x^* & x_0 & x^* & x^* & x^* \end{array}$$

- **Issues**
  - Fix either $\boldsymbol{x}_0$ or ξ_0. Under symmetry "fiducial $\longleftrightarrow$ frequentist".

Fiducial

- **Modeling:** Data generating algorithm: $\boldsymbol{x} = G(\boldsymbol{u}, \xi)$
- **Replication:** data x & parameter ξ linked through DGA, auxiliary variable $\boldsymbol{u}$ replicated

$$\begin{array}{cccccccc} & \xi^* & \xi^* & \xi^* & \xi_0 & \xi^* & \xi^* & \xi^* \\ u_0 & \vdots & \vdots & \vdots & | & \vdots & \vdots & \vdots \\ & x^* & x^* & x^* & x_0 & x^* & x^* & x^* \end{array}$$

- **Issues**
  - Fix either $\boldsymbol{x}_0$ or ξ_0. Under symmetry "fiducial $\longleftrightarrow$ frequentist".
  - Break in symmetry: some u^* incompatible with observed $\boldsymbol{x}_0$. Still useful, frequentist properties need to be established.

Fiducial

- **Modeling:** Data generating algorithm: $\boldsymbol{x} = G(\boldsymbol{u}, \xi)$
- **Replication:** data $\boldsymbol{x}$ & parameter ξ linked through DGA, auxiliary variable $\boldsymbol{u}$ replicated

$$
\begin{array}{cccccccc}
 & \xi^* & \xi^* & \xi^* & \xi_0 & \xi^* & \xi^* & \xi^* \\
u_0 & \vdots & \vdots & \vdots & | & \vdots & \vdots & \vdots \\
 & x^* & x^* & x^* & x_0 & x^* & x^* & x^*
\end{array}
$$

- **Issues**
  - Fix either $\boldsymbol{x}_0$ or ξ_0. Under symmetry "fiducial $\longleftrightarrow$ frequentist".
  - Break in symmetry: some u^* incompatible with observed $\boldsymbol{x}_0$. Still useful, frequentist properties need to be established.
  - Does not satisfy likelihood principle. Philosophical interpretation subject to argument

Outline

Comparison to likelihood

- Density is the function $f(\mathbf{x}, \xi)$, where ξ is fixed and $\mathbf{x}$ is variable.

Comparison to likelihood

- Density is the function $f(\mathbf{x}, \xi)$, where ξ is fixed and $\mathbf{x}$ is variable.
- Likelihood is the function $f(\mathbf{x}, \xi)$, where ξ is variable and $\mathbf{x}$ is fixed.

Comparison to likelihood

- Density is the function $f(\mathbf{x}, \xi)$, where ξ is fixed and $\mathbf{x}$ is variable.
- Likelihood is the function $f(\mathbf{x}, \xi)$, where ξ is variable and $\mathbf{x}$ is fixed.
  - Likelihood as a distribution?

Data generating algorithm

- Data generating algorithm (DGA)

$$\mathbf{X} = \mathbf{G}(\mathbf{U}, \xi),$$

  - U is a random with known distribution (iid $U(0,1)$)
  - Parameter ξ is fixed.
  - Generate $\mathbf{X}$s by generating $\mathbf{U}$s and DGA.
    - This determines sampling distribution

Data generating algorithm

- Data generating equation (DGA)

$$\mathbf{x} = \mathbf{G}(\mathbf{U}^\star, \xi^\star),$$

  - U is a random with known distribution (iid $U(0,1)$)
  - Data $\mathbf{x}$ is fixed
  - Generate ξ^* by generating $\mathbf{U}^\star$s and inverting DGA.
    - This determines fiducial distribution
    - Denote the inverse $Q_{\boldsymbol{x}}(\boldsymbol{U}^*)$.

Example -- Bernoulli trials

- Data generating algorithm

$$X_i = 1\{U_i \leq p\},\ U_i \sim \text{Uniform}(0,1)$$

Generating U_i samples Bernoulli(p).

Example -- Bernoulli trials

- Data generating algorithm

$$X_i = 1\{U_i^\star \leq p^\star\},\ U_i^\star \sim \text{Uniform(0,1)}$$

Estimating U_i by $U_i^\star$ defines fiducial distribution

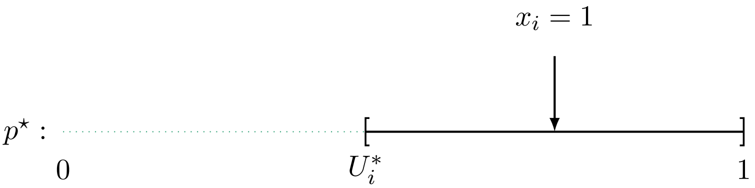
Example -- Bernoulli trials

- Data generating algorithm

$$X_i = 1\{U_i^\star \leq p^\star\},\ U_i^\star \sim \text{Uniform}(0,1)$$

Estimating U_i by $U_i^\star$ defines fiducial distribution

- If $x_i = 1$, then $p^\star \in [U_i^\star, 1]$



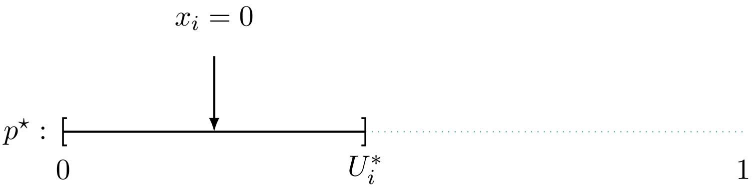
Example -- Bernoulli trials

- Data generating algorithm

$$X_i = 1\{U_i^\star \leq p^\star\},\ U_i^\star \sim \text{Uniform}(0,1)$$

Estimating U_i by $U_i^\star$ defines fiducial distribution

- If $x_i = 0$, then $p^\star \in [0, U_i^\star]$



Example -- Binomial

- Data generating algorithm

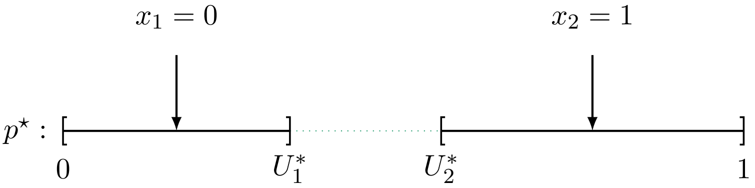
$X_1 = 1\{U_1 \leq p\},\ X_2 = 1\{U_2 \leq p\}$ $\quad$ U_1, U_2 i.i.d. Uniform(0,1)

Example -- Binomial

- Data generating algorithm

  $X_1 = 1\{U_1 \le p\},\ X_2 = 1\{U_2 \le p\}$ $\quad U_1, U_2$ i.i.d. Uniform(0,1)

- If $X_1 = 0, X_2 = 1$ and $U_1^\star < U_2^\star$



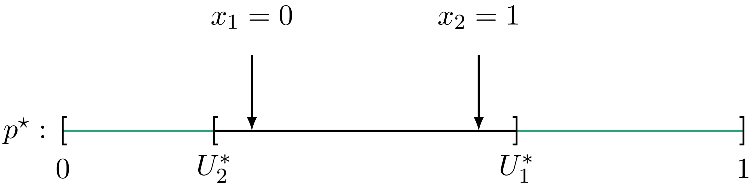
- No solution! Remove $(U_1^\star, U_2^\star)$ inconsistent with data.

Example -- Binomial

- Data generating algorithm

  $X_1 = 1\{U_1 \le p\},\ X_2 = 1\{U_2 \le p\}$ $\qquad$ U_1, U_2 i.i.d. Uniform(0,1)

- If $X_1 = 0, X_2 = 1$ and $U_1^\star > U_2^\star$



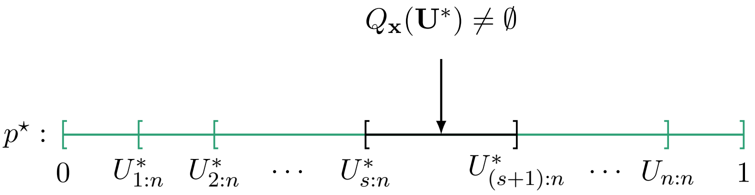
- $(U_1^\star, U_2^\star)$ uniform on $\{U_1^\star > U_2^\star\}$

Example -- Binomial

- $(X_1, \dots X_n) \overset{iid}{\sim}$ Bernoulli(p), $S = \sum_{i=1}^{n} X_i \sim$ Binomial(n, p)

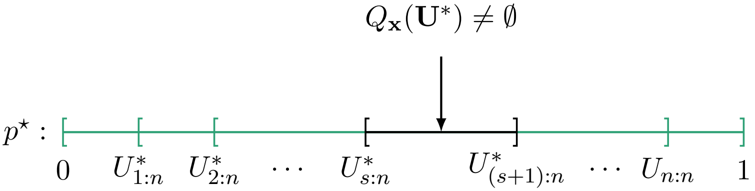
Example -- Binomial

- $(X_1, \ldots X_n) \overset{iid}{\sim}$ Bernoulli(p), $S = \sum_{i=1}^{n} X_i \sim$ Binomial(n, p)
- Condition $\mathbf{U}^*$ on having a solution for p



Example -- Binomial

- $(X_1, \dots X_n) \overset{iid}{\sim}$ Bernoulli(p), $S = \sum_{i=1}^{n} X_i \sim$ Binomial(n, p)
- Condition $\mathbf{U}^*$ on having a solution for p



- Select a point in the interval.
  - A particular choice results in Beta$(s + 1/2, n - s + 1/2)$

Example -- Location Cauchy

- Consider $X_i = \mu + U_i$ where U_i are i.i.d. standard Cauchy.

Example -- Location Cauchy

- Consider $X_i = \mu + U_i$ where U_i are i.i.d. standard Cauchy.
- Solve:

$$Q_{\boldsymbol{x}}(\boldsymbol{u}) = \begin{cases} x_1 - u_1 & \text{if } x_2 - x_1 = u_2 - u_1^\star, \ldots, x_n - x_1 = u_n - u_1 \\ \emptyset & \text{otherwise} \end{cases}$$

Example -- Location Cauchy

- Consider $X_i = \mu + U_i$ where U_i are i.i.d. standard Cauchy.
- Solve:

$$Q_{\boldsymbol{x}}(\boldsymbol{u}) = \begin{cases} x_1 - u_1 & \text{if } x_2 - x_1 = u_2 - u_1^\star, \ldots, x_n - x_1 = u_n - u_1 \\ \emptyset & \text{otherwise} \end{cases}$$

- Estimate $\boldsymbol{u}$ by

$\mathbf{U}^\star$ truncated to $\{x_2 - x_1 = U_2^\star - U_1^\star, \ldots, x_n - x_1 = U_n^\star - U_1^\star\}$

Example -- Location Cauchy

- Consider $X_i = \mu + U_i$ where U_i are i.i.d. standard Cauchy.
- Solve:

$$Q_{\boldsymbol{x}}(\boldsymbol{u}) = \begin{cases} x_1 - u_1 & \text{if } x_2 - x_1 = u_2 - u_1^\star, \dots, x_n - x_1 = u_n - u_1 \\ \emptyset & \text{otherwise} \end{cases}$$

- Estimate $\boldsymbol{u}$ by

$\mathbf{U}^\star$ truncated to $\{x_2 - x_1 = U_2^\star - U_1^\star, \dots, x_n - x_1 = U_n^\star - U_1^\star\}$

- Fiducial density $r_{\boldsymbol{x}}(\mu) \propto \prod_{i=1}^n (1 + (\mu - x_i)^2)^{-1}$.

Example -- Location Cauchy

- Consider $X_i = \mu + U_i$ where U_i are i.i.d. standard Cauchy.
- Solve:

$$Q_{\boldsymbol{x}}(\boldsymbol{u}) = \begin{cases} x_1 - u_1 & \text{if } x_2 - x_1 = u_2 - u_1^\star, \dots, x_n - x_1 = u_n - u_1 \\ \emptyset & \text{otherwise} \end{cases}$$

- Estimate $\boldsymbol{u}$ by

$\mathbf{U}^\star$ truncated to $\{x_2 - x_1 = U_2^\star - U_1^\star, \dots, x_n - x_1 = U_n^\star - U_1^\star\}$

- Fiducial density $r_{\boldsymbol{x}}(\mu) \propto \prod_{i=1}^n (1 + (\mu - x_i)^2)^{-1}$.
  - Location problem – same as posterior computed using Jeffreys prior

General Definition

- Data generating equation $\boldsymbol{X} = \boldsymbol{G}(\boldsymbol{U}, \xi)$.
  - e.g. $X_i = \mu + \sigma U_i$

General Definition

- Data generating equation $\boldsymbol{X} = \boldsymbol{G}(\boldsymbol{U}, \xi)$.
  - e.g. $X_i = \mu + \sigma U_i$
- **Generalized Fiducial Distribution** defined as distribution of

$$\xi(\mathbf{x}, \mathbf{U}^\star) = \arg\min_{\xi} \|\mathbf{x} - \boldsymbol{G}(\mathbf{U}^\star, \xi)\| \tag{1}$$

where $\boldsymbol{U}^\star$ is truncated to

$$\{\boldsymbol{U}^\star : \ \|\mathbf{x} - \boldsymbol{G}(\mathbf{U}^\star, \xi(\mathbf{x}, \mathbf{U}^\star))\| \leq \varepsilon\}$$

Take a limit as $\varepsilon \downarrow 0$.

General Definition

- Data generating equation $\boldsymbol{X} = \boldsymbol{G}(\boldsymbol{U}, \xi)$.
  - e.g. $X_i = \mu + \sigma U_i$
- **Generalized Fiducial Distribution** defined as distribution of

$$\xi(\mathbf{x}, \mathbf{U}^\star) = \arg\min_{\xi} \|\mathbf{x} - \boldsymbol{G}(\mathbf{U}^\star, \xi)\| \tag{1}$$

where $\boldsymbol{U}^\star$ is truncated to

$$\{\boldsymbol{U}^\star : \|\mathbf{x} - \boldsymbol{G}(\mathbf{U}^\star, \xi(\mathbf{x}, \mathbf{U}^\star))\| \leq \varepsilon\}$$

Take a limit as $\varepsilon \downarrow 0$.

- Similar to ABC; generating from prior replaced by $\min$.

General Definition

- Data generating equation $\boldsymbol{X} = \boldsymbol{G}(\boldsymbol{U}, \xi)$.
  - e.g. $X_i = \mu + \sigma U_i$
- **Generalized Fiducial Distribution** defined as distribution of

$$\xi(\mathbf{x}, \mathbf{U}^\star) = \arg\min_{\xi} \|\mathbf{x} - \boldsymbol{G}(\mathbf{U}^\star, \xi)\| \qquad (1)$$

where $\boldsymbol{U}^\star$ is truncated to

$$\{\boldsymbol{U}^\star : \ \|\mathbf{x} - \boldsymbol{G}(\mathbf{U}^\star, \xi(\mathbf{x}, \mathbf{U}^\star))\| \leq \varepsilon\}$$

Take a limit as $\varepsilon \downarrow 0$.

- Similar to ABC; generating from prior replaced by $\min$.
- Computations?

Explicit limit (1)

- Assume $\mathbf{X} \in \mathbb{R}^n$ is continuous; parameter $\xi \in \mathbb{R}^p$
- The limit in (1) has density (H, Iyer, Lai & Lee, 2016)

$$r_{\mathbf{x}}(\xi) = \frac{f_{\mathbf{X}}(\mathbf{x}|\xi) J(\mathbf{x}, \xi)}{\int_{\Xi} f_{\mathbf{X}}(\mathbf{x}|\xi') J(\mathbf{x}, \xi') \, d\xi'},$$

where $J(\mathbf{x}, \xi) = D\left(\nabla_\xi \mathbf{G}(\mathbf{u}, \xi)|_{\mathbf{u}=\mathbf{G}^{-1}(\mathbf{x},\xi)} \right)$

- $n = p$ gives $D(A) = |\det A|$

Explicit limit (1)

- Assume $\mathbf{X} \in \mathbb{R}^n$ is continuous; parameter $\xi \in \mathbb{R}^p$
- The limit in (1) has density (H, Iyer, Lai & Lee, 2016)

$$r_{\mathbf{x}}(\xi) = \frac{f_{\mathbf{X}}(\mathbf{x}|\xi) J(\mathbf{x}, \xi)}{\int_{\Xi} f_{\mathbf{X}}(\mathbf{x}|\xi') J(\mathbf{x}, \xi')\, d\xi'},$$

where $J(\mathbf{x}, \xi) = D\left(\nabla_\xi \mathbf{G}(\mathbf{u}, \xi)|_{\mathbf{u}=\mathbf{G}^{-1}(\mathbf{x},\xi)} \right)$

- $n = p$ gives $D(A) = |\det A|$
- $\|\cdot\|_2$ gives $D(A) = (\det A^\top A)^{1/2}$
- $\|\cdot\|_\infty$ gives $D(A) = \sum_{\mathbf{i}=(i_1,\dots,i_p)} |\det(A)_{\mathbf{i}}|$
- $\|\cdot\|_1$ gives $D(A) = \sum_{\mathbf{i}=(i_1,\dots,i_p)} w_{\mathbf{i}}\, |\det(A)_{\mathbf{i}}|$

Example -- Uniform(θ, θ^2)

- X_i i.i.d. $U(\theta, \theta^2)$, $\theta > 1$

Example -- Uniform(θ, θ^2)

- X_i i.i.d. $U(\theta, \theta^2)$, $\theta > 1$
  - Data generating algorithm $X_i = \theta + (\theta^2 - \theta)U_i,\ U_i \sim U(0, 1)$.

Example -- Uniform(θ, θ^2)

- X_i i.i.d. $U(\theta, \theta^2)$, $\theta > 1$
  - Data generating algorithm $X_i = \theta + (\theta^2 - \theta)U_i,\ U_i \sim U(0, 1)$.
- $\frac{d}{d\theta}[\theta + (\theta^2 - \theta)U_i] = 1 + (2\theta - 1)U_i$, with $U_i = \frac{X_i - \theta}{\theta^2 - \theta}$.

Example -- Uniform(θ, θ^2)

- X_i i.i.d. $U(\theta, \theta^2)$, $\theta > 1$
  - Data generating algorithm $X_i = \theta + (\theta^2 - \theta)U_i,\ U_i \sim U(0,1)$.
- $\frac{d}{d\theta}[\theta + (\theta^2 - \theta)U_i] = 1 + (2\theta - 1)U_i$, with $U_i = \frac{X_i - \theta}{\theta^2 - \theta}$.
- Jacobian

$$J(\boldsymbol{x}, \theta) = D \begin{pmatrix} 1 + \frac{(2\theta-1)(x_1-\theta)}{\theta^2-\theta} \\ \vdots \\ 1 + \frac{(2\theta-1)(x_n-\theta)}{\theta^2-\theta} \end{pmatrix} = \frac{1}{\theta^2 - \theta} D \begin{pmatrix} x_1(2\theta - 1) - \theta^2 \\ \vdots \\ x_n(2\theta - 1) - \theta^2 \end{pmatrix}$$

Example -- Uniform(θ, θ^2)

- X_i i.i.d. $U(\theta, \theta^2)$, $\theta > 1$
  - Data generating algorithm $X_i = \theta + (\theta^2 - \theta)U_i,\ U_i \sim U(0, 1)$.
- $\frac{d}{d\theta}[\theta + (\theta^2 - \theta)U_i] = 1 + (2\theta - 1)U_i$, with $U_i = \frac{X_i - \theta}{\theta^2 - \theta}$.
- Jacobian

$$J(\boldsymbol{x}, \theta) = D \begin{pmatrix} 1 + \frac{(2\theta-1)(x_1-\theta)}{\theta^2-\theta} \\ \vdots \\ 1 + \frac{(2\theta-1)(x_n-\theta)}{\theta^2-\theta} \end{pmatrix} = \frac{1}{\theta^2 - \theta} D \begin{pmatrix} x_1(2\theta - 1) - \theta^2 \\ \vdots \\ x_n(2\theta - 1) - \theta^2 \end{pmatrix}$$

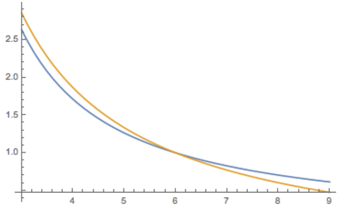
  - $= n\frac{\bar{x}(2\theta-1)-\theta^2}{\theta^2-\theta}$ for L_∞.

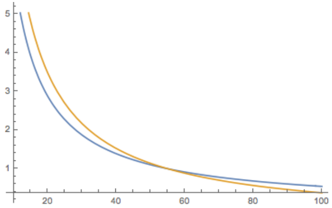
Example -- Uniform(θ, θ^2)

- Reference prior (Berger, Bernardo & Sun, 2009)
$\pi(\theta) = \frac{e^{\psi\left(\frac{2\theta}{2\theta-1}\right)}(2\theta-1)}{\theta^2-\theta}$.

Example -- Uniform(θ, θ^2)

- Reference prior (Berger, Bernardo & Sun, 2009)
  $\pi(\theta) = \frac{e^{\psi\left(\frac{2\theta}{2\theta-1}\right)}(2\theta-1)}{\theta^2-\theta}$.
- reference prior vs fiducial Jacobian



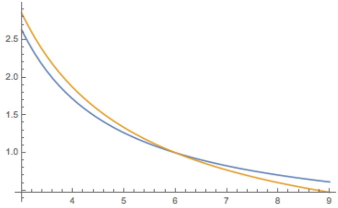


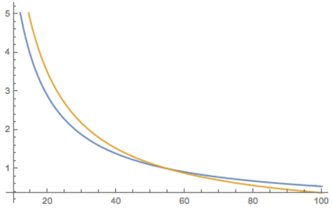
Example -- Uniform(θ, θ^2)

- Reference prior (Berger, Bernardo & Sun, 2009)

  $\pi(\theta) = \frac{e^{\psi\left(\frac{2\theta}{2\theta-1}\right)}(2\theta-1)}{\theta^2-\theta}$.

- reference prior vs fiducial Jacobian





- In simulations fiducial was marginally better than reference prior which was much better than flat prior.

Example -- Linear Regression

- Data generating algorithm $Y = X\beta + \sigma Z$

Example -- Linear Regression

- Data generating algorithm $Y = X\beta + \sigma Z$
- $\frac{d}{d\theta} Y = (X, Z)$ and $Z = (Y - X\beta)/\sigma$.

Example -- Linear Regression

- Data generating algorithm $Y = X\beta + \sigma Z$
- $\frac{d}{d\theta}Y = (X, Z)$ and $Z = (Y - X\beta)/\sigma$.
- Jacobian $J(\boldsymbol{y}, \beta, \sigma) = D\left(\boldsymbol{X}, \frac{\boldsymbol{y}-\boldsymbol{X}\beta}{\sigma}\right) = \sigma^{-1}D(\boldsymbol{X}, \boldsymbol{y})$

Example -- Linear Regression

- Data generating algorithm $Y = X\beta + \sigma Z$
- $\frac{d}{d\theta}Y = (X, Z)$ and $Z = (Y - X\beta)/\sigma$.
- Jacobian $J(\boldsymbol{y}, \beta, \sigma) = D\left(\boldsymbol{X}, \frac{\boldsymbol{y}-\boldsymbol{X}\beta}{\sigma}\right) = \sigma^{-1}D(\boldsymbol{X}, \boldsymbol{y})$
  - $= \sigma^{-1}|\det(X^T X)|^{1/2}(RSS)^{1/2}$ for L_2.

Example -- Linear Regression

- Data generating algorithm $Y = X\beta + \sigma Z$
- $\frac{d}{d\theta}Y = (X, Z)$ and $Z = (Y - X\beta)/\sigma$.
- Jacobian $J(\boldsymbol{y}, \beta, \sigma) = D\left(\boldsymbol{X}, \frac{\boldsymbol{y}-\boldsymbol{X}\beta}{\sigma}\right) = \sigma^{-1}D(\boldsymbol{X}, \boldsymbol{y})$
    - $= \sigma^{-1}|\det(X^TX)|^{1/2}(RSS)^{1/2}$ for L_2.
- Same as independence Jeffreys, *explicit* normalizing constant

Example -- Generalized Pareto

- $X_i = G(U_i, \gamma, \sigma) = \sigma \frac{U_i^{-\gamma} - 1}{\gamma}$
  - Models excedances over a large threshold.

Example -- Generalized Pareto

- $X_i = G(U_i, \gamma, \sigma) = \sigma \frac{U_i^{-\gamma} - 1}{\gamma}$
  - Models excedances over a large threshold.
- Likelihood $f(\mathbf{x}, \gamma, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\left(1 + \frac{\gamma x_i}{\sigma}\right)^{1+1/\gamma}}$.

Example -- Generalized Pareto

- $X_i = G(U_i, \gamma, \sigma) = \sigma \frac{U_i^{-\gamma} - 1}{\gamma}$
  - Models excedances over a large threshold.
- Likelihood $f(\mathbf{x}, \gamma, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\left(1+\frac{\gamma x_i}{\sigma}\right)^{1+1/\gamma}}$.
- Jacobian evaluated at $u_i = \left(1 + \frac{\gamma x_i}{\sigma}\right)^{-1/\gamma}$
  - $\frac{d}{d\sigma} G(u_i, \gamma, \sigma) = \frac{x_i}{\sigma}$.

Example -- Generalized Pareto

- $X_i = G(U_i, \gamma, \sigma) = \sigma \frac{U_i^{-\gamma} - 1}{\gamma}$
  - Models excedances over a large threshold.
- Likelihood $f(\mathbf{x}, \gamma, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\left(1 + \frac{\gamma x_i}{\sigma}\right)^{1+1/\gamma}}$.
- Jacobian evaluated at $u_i = \left(1 + \frac{\gamma x_i}{\sigma}\right)^{-1/\gamma}$
  - $\frac{d}{d\sigma} G(u_i, \gamma, \sigma) = \frac{x_i}{\sigma}$.
  - $\frac{d}{d\gamma} G(u_i, \gamma, \sigma) = -\frac{x_i}{\gamma} + \frac{\sigma\left(1 + \frac{\gamma x_i}{\sigma}\right) \log\left(1 + \frac{\gamma x_i}{\sigma}\right)}{\gamma^2}$.

Example -- Generalized Pareto

- $X_i = G(U_i, \gamma, \sigma) = \sigma \frac{U_i^{-\gamma} - 1}{\gamma}$
  - Models excedances over a large threshold.
- Likelihood $f(\mathbf{x}, \gamma, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\left(1+\frac{\gamma x_i}{\sigma}\right)^{1+1/\gamma}}$.
- Jacobian evaluated at $u_i = \left(1 + \frac{\gamma x_i}{\sigma}\right)^{-1/\gamma}$
  - $\frac{d}{d\sigma} G(u_i, \gamma, \sigma) = \frac{x_i}{\sigma}$.
  - $\frac{d}{d\gamma} G(u_i, \gamma, \sigma) = -\frac{x_i}{\gamma} + \frac{\sigma\left(1+\frac{\gamma x_i}{\sigma}\right)\log\left(1+\frac{\gamma x_i}{\sigma}\right)}{\gamma^2}$.
  - $J(\mathbf{x}, \gamma, \sigma) = \gamma^{-2} D \begin{pmatrix} x_1 & \left(1 + \frac{\gamma x_1}{\sigma}\right) \log\left(1 + \frac{\gamma x_1}{\sigma}\right) \\ \vdots & \vdots \\ x_n & \left(1 + \frac{\gamma x_n}{\sigma}\right) \log\left(1 + \frac{\gamma x_n}{\sigma}\right) \end{pmatrix}$

Example -- Generalized Pareto

- $X_i = G(U_i, \gamma, \sigma) = \sigma \frac{U_i^{-\gamma} - 1}{\gamma}$
  - Models excedances over a large threshold.
- Likelihood $f(\mathbf{x}, \gamma, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\left(1 + \frac{\gamma x_i}{\sigma}\right)^{1+1/\gamma}}$.
- Jacobian evaluated at $u_i = \left(1 + \frac{\gamma x_i}{\sigma}\right)^{-1/\gamma}$
  - $\frac{d}{d\sigma} G(u_i, \gamma, \sigma) = \frac{x_i}{\sigma}$.
  - $\frac{d}{d\gamma} G(u_i, \gamma, \sigma) = -\frac{x_i}{\gamma} + \frac{\sigma\left(1+\frac{\gamma x_i}{\sigma}\right)\log\left(1+\frac{\gamma x_i}{\sigma}\right)}{\gamma^2}$.
  - $J(\mathbf{x}, \gamma, \sigma) = \gamma^{-2} D \begin{pmatrix} x_1 & \left(1 + \frac{\gamma x_1}{\sigma}\right) \log\left(1 + \frac{\gamma x_1}{\sigma}\right) \\ \vdots & \vdots \\ x_n & \left(1 + \frac{\gamma x_n}{\sigma}\right) \log\left(1 + \frac{\gamma x_n}{\sigma}\right) \end{pmatrix}$
    - $= \sum_{i<j} \left| \frac{x_j\left(1+\frac{\gamma x_i}{\sigma}\right)\log\left(1+\frac{\gamma x_i}{\sigma}\right) - x_i\left(1+\frac{\gamma x_j}{\sigma}\right)\log\left(1+\frac{\gamma x_j}{\sigma}\right)}{\gamma^2} \right|$ for L_∞.

Exercise

Derive a GFD for:

- Weibull distribution

Exercise

Derive a GFD for:

- Weibull distribution
- Negative Binomial Distribution (compare to Binomial)

Exercise

Derive a GFD for:

- Weibull distribution
- Negative Binomial Distribution (compare to Binomial)
- T distribution (might not have a pretty form)

Exercise

Derive a GFD for:

- Weibull distribution
- Negative Binomial Distribution (compare to Binomial)
- T distribution (might not have a pretty form)
- Your favorite model

Exercise

Derive a GFD for:

- Weibull distribution
- Negative Binomial Distribution (compare to Binomial)
- T distribution (might not have a pretty form)
- Your favorite model

Open problem:

- Derive Jacobian formula on manifolds

Outline

Important Observations (Bayesian)

- GFD is always proper

Important Observations (Bayesian)

- GFD is always proper
- GFD is invariant to re-parametrizations (same as Jeffreys)

Important Observations (Bayesian)

- GFD is always proper
- GFD is invariant to re-parametrizations (same as Jeffreys)
- GFD is *not* invariant to smooth transformation of the data if $n > p$

Important Observations (Bayesian)

- GFD is always proper
- GFD is invariant to re-parametrizations (same as Jeffreys)
- GFD is *not* invariant to smooth transformation of the data if $n > p$
- Consequently:
  - GFD does not satisfy likelihood principle.

Important Observations (Bayesian)

- GFD is always proper
- GFD is invariant to re-parametrizations (same as Jeffreys)
- GFD is *not* invariant to smooth transformation of the data if $n > p$
- Consequently:
  - GFD does not satisfy likelihood principle.
  - Adding a multiple of a column to another column does not alter $D(A)$. Row operations not allowed!

Classical Result (n=1, p=1; Frequentist)

Data generating algorithm $S = G_S(\mathbf{U}, \xi)$ (1-dimensional statistic)

Classical Result (n=1, p=1; Frequentist)

Data generating algorithm $S = G_S(\mathbf{U}, \xi)$ (1-dimensional statistic)

1. $G_S(\mathbf{u}, \xi)$ is non-decreasing in ξ for all $\mathbf{u}$
2. For all $\mathbf{u}$ and s the inverse $Q_s(\mathbf{u}) = \{\xi : s = G_S(\mathbf{u}, \xi)\} \neq \emptyset$.
3. For all ξ the cdf $F_S(s, \xi)$ is continuous.

Classical Result (n=1, p=1; Frequentist)

Data generating algorithm $S = G_S(\mathbf{U}, \xi)$ (1-dimensional statistic)

1. $G_S(\mathbf{u}, \xi)$ is non-decreasing in ξ for all $\mathbf{u}$
2. For all $\mathbf{u}$ and s the inverse $Q_s(\mathbf{u}) = \{\xi : s = G_S(\mathbf{u}, \xi)\} \neq \emptyset$.
3. For all ξ the cdf $F_S(s, \xi)$ is continuous.

Then one has "unique" fiducial distribution and exact coverage of one-sided confidence intervals, i.e.,

$$P(Q_s(\mathbf{U}^\star) \leq \xi) = 1 - F_S(s, \xi).$$

Classical Result (n=1, p=1; Frequentist)

Data generating algorithm $S = G_S(\mathbf{U}, \xi)$ (1-dimensional statistic)

1. $G_S(\mathbf{u}, \xi)$ is non-decreasing in ξ for all $\mathbf{u}$
2. For all $\mathbf{u}$ and s the inverse $Q_s(\mathbf{u}) = \{\xi : s = G_S(\mathbf{u}, \xi)\} \neq \emptyset$.
3. For all ξ the cdf $F_S(s, \xi)$ is continuous.

Then one has "unique" fiducial distribution and exact coverage of one-sided confidence intervals, i.e.,

$$P(Q_s(\mathbf{U}^\star) \leq \xi) = 1 - F_S(s, \xi).$$

- If $S \sim \xi_0$ then $1 - F_S(S, \xi_0) \sim U(0, 1)$ – fiducial p-value.

Exact frequentist coverage

- Set $P(Q_s(\mathbf{U}^\star) \leq C_\alpha(s)) = 1 - \alpha$.

Exact frequentist coverage

- Set $P(Q_s(\mathbf{U}^\star) \leq C_\alpha(s)) = 1 - \alpha$.
- Coverage of upper confidence limit:

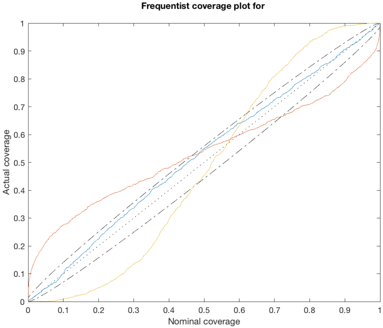
$$
\begin{aligned}
P_\xi(\xi \leq C_\alpha(S)) &= P_\xi(P(Q_S(\mathbf{U}^\star) \leq \xi | S) \leq 1 - \alpha) \\
&= P(U(0,1) \leq 1 - \alpha) = 1 - \alpha
\end{aligned}
$$

Exact frequentist coverage

- Set $P(Q_s(\mathbf{U}^\star) \leq C_\alpha(s)) = 1 - \alpha$.
- Coverage of upper confidence limit:

$$P_\xi(\xi \leq C_\alpha(S)) = P_\xi(P(Q_S(\mathbf{U}^\star) \leq \xi | S) \leq 1 - \alpha)$$
$$= P(U(0,1) \leq 1 - \alpha) = 1 - \alpha$$

- This is general: simulate m fiducial p-values



exact
conservative
liberal

Exact frequentist coverage $(p > 1)$

- Which set of fiducial probability $1 - \alpha$ will be confidence set?

Exact frequentist coverage $(p > 1)$

- Which set of fiducial probability $1 - \alpha$ will be confidence set?
- Follow pivots and Inferential Models (Liu & Martin, 2015)

Exact frequentist coverage $(p > 1)$

- Which set of fiducial probability $1 - \alpha$ will be confidence set?
- Follow pivots and Inferential Models (Liu & Martin, 2015)
  - Select a $P(\boldsymbol{U}^* \in \mathcal{U}) = \alpha$
  - Set $Q_S(\mathcal{U})$ has both fiducial probability and confidence $1 - \alpha$

Exact frequentist coverage $(p > 1)$

- Which set of fiducial probability $1-\alpha$ will be confidence set?
- Follow pivots and Inferential Models (Liu & Martin, 2015)
  - Select a $P(\boldsymbol{U}^* \in \mathcal{U}) = \alpha$
  - Set $Q_S(\mathcal{U})$ has both fiducial probability and confidence $1-\alpha$
  - $p = 1$ above: $\xi \in (-\infty, C(s)) \longleftrightarrow u \in (\alpha, 1)$

Exact frequentist coverage $(p > 1)$

- Which set of fiducial probability $1-\alpha$ will be confidence set?
- Follow pivots and Inferential Models (Liu & Martin, 2015)
  - Select a $P(\boldsymbol{U}^* \in \mathcal{U}) = \alpha$
  - Set $Q_S(\mathcal{U})$ has both fiducial probability and confidence $1-\alpha$
  - $p = 1$ above: $\xi \in (-\infty, C(s)) \longleftrightarrow u \in (\alpha, 1)$
- Comments:
  - Links sets of $1-\alpha$ fiducial probability for different $\boldsymbol{X}$.

Exact frequentist coverage ($p > 1$)

- Which set of fiducial probability $1 - \alpha$ will be confidence set?
- Follow pivots and Inferential Models (Liu & Martin, 2015)
  - Select a $P(\boldsymbol{U}^* \in \mathcal{U}) = \alpha$
  - Set $Q_S(\mathcal{U})$ has both fiducial probability and confidence $1 - \alpha$
  - $p = 1$ above: $\xi \in (-\infty, C(s)) \longleftrightarrow u \in (\alpha, 1)$
- Comments:
  - Links sets of $1 - \alpha$ fiducial probability for different $\boldsymbol{X}$.
  - Reverse: Map $C(S)$ of fiducial probability $1 - \alpha$ to $\mathcal{U}$.
    If invariant in $\boldsymbol{X}$ then exact coverage.

Example -- Bliss-Fieller-Creasy's Problem

- $X \sim N(\mu_x, 1),\ Y \sim N(\mu_y, 1)$, parameter of interest $\eta = \frac{\mu_x}{\mu_y}$

Example -- Bliss-Fieller-Creasy's Problem

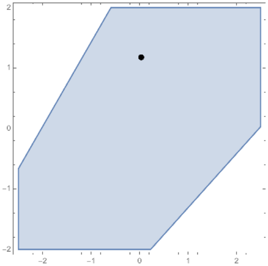
- $X \sim N(\mu_x, 1),\ Y \sim N(\mu_y, 1)$, parameter of interest $\eta = \frac{\mu_x}{\mu_y}$
- DGA 1: $X = \mu_x + U_x,\ Y = \mu_y + U_y$
- Marginal Fiducial RV $Q_{x,y}(U_x^*, U_y^*) = \frac{x - U_x^*}{y - U_y^*}$

Example -- Bliss-Fieller-Creasy's Problem

- $X \sim N(\mu_x, 1),\ Y \sim N(\mu_y, 1)$, parameter of interest $\eta = \frac{\mu_x}{\mu_y}$
- DGA 1: $X = \mu_x + U_x,\ Y = \mu_y + U_y$
- Marginal Fiducial RV $Q_{x,y}(U_x^*, U_y^*) = \frac{x-U_x^*}{y-U_y^*}$
- Consider equal tailed regions of fiducial probability 90%:
  - When $|\mu_y| >> 0$ good frequentist performance, when $\mu_y \approx 0$ poor performance.

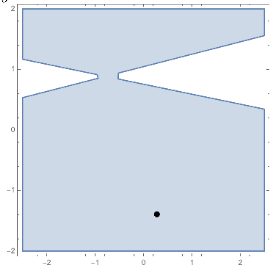
Example -- Bliss-Fieller-Creasy's Problem

- $X \sim N(\mu_x, 1)$, $Y \sim N(\mu_y, 1)$, parameter of interest $\eta = \frac{\mu_x}{\mu_y}$
- DGA 1: $X = \mu_x + U_x$, $Y = \mu_y + U_y$
- Marginal Fiducial RV $Q_{x,y}(U_x^*, U_y^*) = \frac{x - U_x^*}{y - U_y^*}$
- Consider equal tailed regions of fiducial probability 90%:
  - When $|\mu_y| >> 0$ good frequentist performance, when $\mu_y \approx 0$ poor performance.
  - Visualize $\mathcal{U} = \{(u_x, u_y) : \ c_1 \leq \frac{x - u_x^*}{y - u_y^*} \leq c_2\}$



$\mu_{x0} = 10, \mu_{y0} = 10$

coverage≈90%

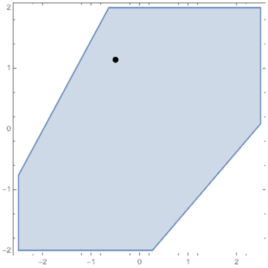


$\mu_{x0} = 0.5, \mu_{y0} = 0.5$

coverage≈96%

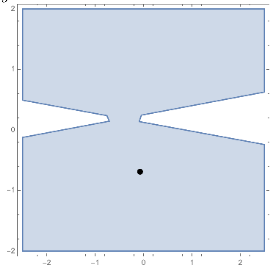
Example -- Bliss-Fieller-Creasy's Problem

- $X \sim N(\mu_x, 1)$, $Y \sim N(\mu_y, 1)$, parameter of interest $\eta = \frac{\mu_x}{\mu_y}$
- DGA 1: $X = \mu_x + U_x$, $Y = \mu_y + U_y$
- Marginal Fiducial RV $Q_{x,y}(U_x^*, U_y^*) = \frac{x - U_x^*}{y - U_y^*}$
- Consider equal tailed regions of fiducial probability 90%:
  - When $|\mu_y| >> 0$ good frequentist performance, when $\mu_y \approx 0$ poor performance.
  - Visualize $\mathcal{U} = \{(u_x, u_y) : \ c_1 \leq \frac{x - u_x^*}{y - u_y^*} \leq c_2\}$



$\mu_{x0} = 10, \mu_{y0} = 10$

coverage≈90%

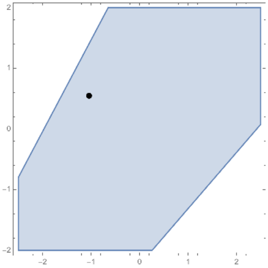


$\mu_{x0} = 0.5, \mu_{y0} = 0.5$

coverage≈96%

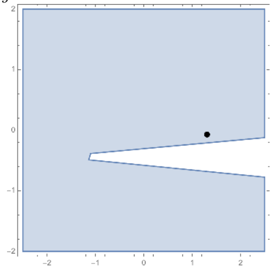
Example -- Bliss-Fieller-Creasy's Problem

- $X \sim N(\mu_x, 1)$, $Y \sim N(\mu_y, 1)$, parameter of interest $\eta = \frac{\mu_x}{\mu_y}$
- DGA 1: $X = \mu_x + U_x,\ Y = \mu_y + U_y$
- Marginal Fiducial RV $Q_{x,y}(U_x^*, U_y^*) = \frac{x - U_x^*}{y - U_y^*}$
- Consider equal tailed regions of fiducial probability 90%:
  - When $|\mu_y| >> 0$ good frequentist performance, when $\mu_y \approx 0$ poor performance.
  - Visualize $\mathcal{U} = \{(u_x, u_y) :\ c_1 \leq \frac{x - u_x^*}{y - u_y^*} \leq c_2\}$



$\mu_{x0} = 10, \mu_{y0} = 10$

coverage≈90%



$\mu_{x0} = 0.5, \mu_{y0} = 0.5$

coverage≈96%

Example -- Better Solution

- DGA 2: $X = \eta\mu_y + \frac{-\eta U_1 + U_2}{\sqrt{1+\eta^2}}$, $Y = \mu_y + \frac{U_1 + \eta U_2}{\sqrt{1+\eta^2}}$

Example -- Better Solution

- DGA 2: $X = \eta\mu_y + \frac{-\eta U_1 + U_2}{\sqrt{1+\eta^2}}$, $Y = \mu_y + \frac{U_1 + \eta U_2}{\sqrt{1+\eta^2}}$
- Fiducial density $r_{x,y}(\eta) = \frac{e^{-\frac{(x-\eta y)^2}{2(\eta^2+1)}} |y + x\eta|}{\sqrt{2\pi}(\eta^2+1)^{3/2}}$
  - Jacobian $J(x, y, \eta, \mu_y) = \frac{|y+x\eta|}{1+\eta^2}$

Example -- Better Solution

- DGA 2: $X = \eta\mu_y + \frac{-\eta U_1 + U_2}{\sqrt{1+\eta^2}}$, $Y = \mu_y + \frac{U_1 + \eta U_2}{\sqrt{1+\eta^2}}$
- Fiducial density $r_{x,y}(\eta) = \frac{e^{-\frac{(x-\eta y)^2}{2(\eta^2+1)}}|y+x\eta|}{\sqrt{2\pi}(\eta^2+1)^{3/2}}$
  - Jacobian $J(x,y,\eta,\mu_y) = \frac{|y+x\eta|}{1+\eta^2}$
- Fieller picked the set of fiducial probability 90% corresponding to $\mathcal{U} = \{|u_1| < 1.96\}$

Example -- Better Solution

- DGA 2: $X = \eta\mu_y + \frac{-\eta U_1 + U_2}{\sqrt{1+\eta^2}}$, $Y = \mu_y + \frac{U_1 + \eta U_2}{\sqrt{1+\eta^2}}$
- Fiducial density $r_{x,y}(\eta) = \frac{e^{-\frac{(x-\eta y)^2}{2(\eta^2+1)}}|y+x\eta|}{\sqrt{2\pi}(\eta^2+1)^{3/2}}$
  - Jacobian $J(x, y, \eta, \mu_y) = \frac{|y+x\eta|}{1+\eta^2}$
- Fieller picked the set of fiducial probability 90% corresponding to $\mathcal{U} = \{|u_1| < 1.96\}$
  - Pros: Fiducial sets are linked, exact coverage is guaranteed
  - Cons: The shape of the sets is strange (interval, complement of interval, whole real line)

Example -- Better Solution

- DGA 2: $X = \eta\mu_y + \frac{-\eta U_1 + U_2}{\sqrt{1+\eta^2}}$, $Y = \mu_y + \frac{U_1 + \eta U_2}{\sqrt{1+\eta^2}}$
- Fiducial density $r_{x,y}(\eta) = \frac{e^{-\frac{(x-\eta y)^2}{2(\eta^2+1)}} |y + x\eta|}{\sqrt{2\pi}(\eta^2+1)^{3/2}}$
  - Jacobian $J(x, y, \eta, \mu_y) = \frac{|y+x\eta|}{1+\eta^2}$
- Fieller picked the set of fiducial probability 90% corresponding to $\mathcal{U} = \{|u_1| < 1.96\}$
  - Pros: Fiducial sets are linked, exact coverage is guaranteed
  - Cons: The shape of the sets is strange (interval, complement of interval, whole real line)
- GFD1 $\approx$ GFD2 if $|y| >> 0$.

Ancillary Representation $(n > 1, p = 1)$

4. Let $(S(\mathbf{X}), \mathbf{A}(\mathbf{X}))$ be a smooth 1-1 transformation of $\boldsymbol{X} = \boldsymbol{G}(\boldsymbol{U}, \xi)$.
   - $S(\mathbf{X})$ is one dimensional satisfying 1, 2, 3.
   - $\mathbf{A}(\mathbf{X})$ is a vector of functional ancillary statistics $(\frac{\partial}{\partial \xi}\mathbf{A} \circ \mathbf{G}(\boldsymbol{U}, \xi) = \mathbf{0})$.

Theorem (Majumder, 2015)

If (4) is satisfied GFI derived from $(S, \mathbf{A})$ is exact.

Ancillary Representation $(n > 1, p = 1)$

4. Let $(S(\mathbf{X}), \mathbf{A}(\mathbf{X}))$ be a smooth 1-1 transformation of $\boldsymbol{X} = \boldsymbol{G}(\boldsymbol{U}, \xi)$.
   - $S(\mathbf{X})$ is one dimensional satisfying 1, 2, 3.
   - $\mathbf{A}(\mathbf{X})$ is a vector of functional ancillary statistics $(\frac{\partial}{\partial \xi} \mathbf{A} \circ \mathbf{G}(\boldsymbol{U}, \xi) = \mathbf{0})$.

Theorem (Majumder, 2015)

If (4) is satisfied GFI derived from $(S, \mathbf{A})$ is exact.

- Idea: The GFD is the same as FD based on $S = G_{S|\mathbf{a}}(\mathbf{U}_\mathbf{a}, \xi)$
  - $\mathbf{U}_\mathbf{a} \sim \mathbf{U} \mid \mathbf{a} = A(\mathbf{G}(\mathbf{U}, \xi))$
  - $G_{S|\mathbf{a}}$ is the restriction of G to the domain of $U_\mathbf{a}$.

Ancillary Representation $(n > 1, p = 1)$

4. Let $(S(\mathbf{X}), \mathbf{A}(\mathbf{X}))$ be a smooth 1-1 transformation of $\boldsymbol{X} = \boldsymbol{G}(\boldsymbol{U}, \xi)$.
   - $S(\mathbf{X})$ is one dimensional satisfying 1, 2, 3.
   - $\mathbf{A}(\mathbf{X})$ is a vector of functional ancillary statistics ($\frac{\partial}{\partial \xi}\mathbf{A} \circ \mathbf{G}(\boldsymbol{U}, \xi) = \mathbf{0}$).

Theorem (Majumder, 2015)

If (4) is satisfied GFI derived from $(S, \mathbf{A})$ is exact.

- Idea: The GFD is the same as FD based on $S = G_{S|\mathbf{a}}(\mathbf{U}_{\mathbf{a}}, \xi)$
  - $\mathbf{U}_{\mathbf{a}} \sim \mathbf{U} \mid \mathbf{a} = A(\mathbf{G}(\mathbf{U}, \xi))$
  - $G_{S|\mathbf{a}}$ is the restriction of G to the domain of $U_{\mathbf{a}}$.
- Same argument works for $p > 1$.

Take home

- Confidence sets need to be linked across potential data sets

Take home

- Confidence sets need to be linked across potential data sets
- Confidence Curves provide both confidence distribution and confidence sets

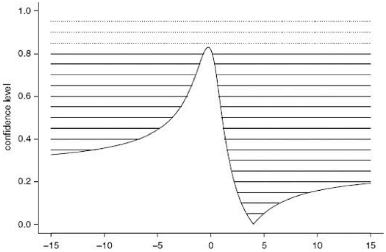


Figure 4.11 from Schweder & Hjort (2017) $x = 1.333, y = 0.333$

Another look

- What is needed about asymptoticly correct coverage?

Another look

- What is needed about asymptoticly correct coverage?
  - Convergence of the posteriors to something nice (Bernstein - von Mises)

Another look

- What is needed about asymptoticly correct coverage?
  - Convergence of the posteriors to something nice (Bernstein - von Mises)
  - Linkage of credible sets across all potential data (one sided CI)

Various Asymptotic Results (Frequentist)

$r_{\boldsymbol{x}}(\xi) \propto f_{\mathbf{X}}(\mathbf{x}|\xi) J(\mathbf{x}, \xi)$ where $J(\mathbf{x}, \xi) = D\left(\nabla_\xi \mathbf{G}(\mathbf{u}, \xi)|_{\mathbf{u}=\mathbf{G}^{-1}(\mathbf{x}, \xi)} \right)$

- Most start with $C_n^{-1} J(\mathbf{x}, \xi) \to J(\xi_0, \xi)$
- Bernstein-von Mises theorem for fiducial distributions provides asymptotic correctness of fiducial CIs H (2009, 2013), Sonderegger & H (2013) .
- Consistency of model selection H & Lee (2009), Lai, H & Lee (2015), H, Iyer, Lai & Lee (2016).
- Fiducial non-parametrics Cui & H (2019, 2020+, 2021+)

Outline

- Introduction
- Definition
- Theoretical Results
- Applications
  - High D Regression
  - Distributed Data
  - Fiducial Autoencoder
  - Likelihood ratio in Forensic Science
- Conclusions

Outline

Model Selection

- $\boldsymbol{Y} = \boldsymbol{G}(M, \boldsymbol{\xi}_M, \boldsymbol{U}), \qquad M \in \mathcal{M},\ \boldsymbol{\xi}_M \in \boldsymbol{\xi}_M$

Theorem: (H, Iyer, Lai, Lee 2016) Under assumptions

$$r_{\boldsymbol{y}}(M) \propto q^{|M|} \int_{\boldsymbol{\xi}_M} f_M(\boldsymbol{y}, \boldsymbol{\xi}_M) J_M(\boldsymbol{y}, \boldsymbol{\xi}_M)\, d\boldsymbol{\xi}_M$$

Model Selection

- $\boldsymbol{Y} = \boldsymbol{G}(M, \boldsymbol{\xi}_M, \boldsymbol{U}), \qquad M \in \mathcal{M},\ \boldsymbol{\xi}_M \in \boldsymbol{\xi}_M$

Theorem: (H, Iyer, Lai, Lee 2016) Under assumptions

$$r_{\boldsymbol{y}}(M) \propto q^{|M|} \int_{\boldsymbol{\xi}_M} f_M(\boldsymbol{y}, \boldsymbol{\xi}_M) J_M(\boldsymbol{y}, \boldsymbol{\xi}_M)\, d\boldsymbol{\xi}_M$$

- Need for penalty – in fiducial framework additional equations $0 = P_k, \quad k = 1, \ldots, \min(|M|, n)$

Model Selection

- $\boldsymbol{Y} = \boldsymbol{G}(M, \boldsymbol{\xi}_M, \boldsymbol{U}), \qquad M \in \mathcal{M},\ \boldsymbol{\xi}_M \in \boldsymbol{\xi}_M$

Theorem: (H, Iyer, Lai, Lee 2016) Under assumptions

$$r_{\boldsymbol{y}}(M) \propto q^{|M|} \int_{\boldsymbol{\xi}_M} f_M(\boldsymbol{y}, \boldsymbol{\xi}_M) J_M(\boldsymbol{y}, \boldsymbol{\xi}_M)\, d\boldsymbol{\xi}_M$$

- Need for penalty – in fiducial framework additional equations $0 = P_k, \quad k = 1, \ldots, \min(|M|, n)$
  - Default value $q = n^{-1/2}$ (motivated by MDL)

Alternative to penalty

- Penalty is used to discourage models with many parameters

Alternative to penalty

- Penalty is used to discourage models with many parameters
- Real issue: Not too many parameters but a smaller model can do almost the same job

Alternative to penalty

- Penalty is used to discourage models with many parameters
- Real issue: Not too many parameters but a smaller model can do almost the same job

$$r_{\boldsymbol{y}}(M) \propto \int_{\Xi_M} f_M(\boldsymbol{y}, \boldsymbol{\xi}_M) J_M(\boldsymbol{y}, \boldsymbol{\xi}_M) h_M(\boldsymbol{\xi}_M)\, d\boldsymbol{\xi}_M,$$

$$h_M(\boldsymbol{\xi}_M) = \begin{cases} 0 & \text{a smaller model predicts nearly as well} \\ 1 & \text{otherwise} \end{cases}$$

Alternative to penalty

- Penalty is used to discourage models with many parameters
- Real issue: Not too many parameters but a smaller model can do almost the same job

$$r_{\boldsymbol{y}}(M) \propto \int_{\Xi_M} f_M(\boldsymbol{y}, \boldsymbol{\xi}_M) J_M(\boldsymbol{y}, \boldsymbol{\xi}_M) h_M(\boldsymbol{\xi}_M)\, d\boldsymbol{\xi}_M,$$

$$h_M(\boldsymbol{\xi}_M) = \begin{cases} 0 & \text{a smaller model predicts nearly as well} \\ 1 & \text{otherwise} \end{cases}$$

- Motivated by non-local priors of Johnson & Rossell (2009)

Regression

- $\boldsymbol{Y} = X\beta + \sigma Z$
- First idea $h_M(\beta_M) = I_{\{|\beta_i|>\epsilon,\, i\in M\}}$ – issue: collinearity

Regression

- $\boldsymbol{Y} = X\beta + \sigma Z$
- First idea $h_M(\beta_M) = I_{\{|\beta_i| > \epsilon,\ i \in M\}}$ – issue: collinearity
- Better:

$$h_M(\beta_M) := I_{\left\{\frac{1}{2}\|X^T(X_M\beta_M - Xb_{min})\|_2^2 \geq \varepsilon_M\right\}}$$

where b_{min} solves

$$\min_{b \in R^p} \frac{1}{2}\|X^T(X_M\beta_M - Xb)\|_2^2 \quad \text{subject to} \quad \|b\|_0 \leq |M| - 1.$$

  - algorithm – Bertsimas et al (2016)

Regression

- $\boldsymbol{Y} = X\beta + \sigma Z$
- First idea $h_M(\beta_M) = I_{\{|\beta_i|>\epsilon,\, i\in M\}}$ – issue: collinearity
- Better:

$$h_M(\beta_M) := I_{\left\{\frac{1}{2}\|X^T(X_M\beta_M - Xb_{min})\|_2^2 \geq \varepsilon_M\right\}}$$

where b_{min} solves

$$\min_{b\in R^p} \frac{1}{2}\|X^T(X_M\beta_M - Xb)\|_2^2 \quad \text{subject to} \quad \|b\|_0 \leq |M| - 1.$$

  - algorithm – Bertsimas et al (2016)
  - similar to Dantzig selector Candes & Tao (2007)

Regression

- $\boldsymbol{Y} = X\beta + \sigma Z$
- First idea $h_M(\beta_M) = I_{\{|\beta_i|>\epsilon,\, i\in M\}}$ – issue: collinearity
- Better:

$$h_M(\beta_M) := I_{\left\{\frac{1}{2}\|X^T(X_M\beta_M - Xb_{min})\|_2^2 \geq \varepsilon_M\right\}}$$

where b_{min} solves

$$\min_{b\in R^p} \frac{1}{2}\|X^T(X_M\beta_M - Xb)\|_2^2 \quad \text{subject to} \quad \|b\|_0 \leq |M| - 1.$$

  - algorithm – Bertsimas et al (2016)
  - similar to Dantzig selector Candes & Tao (2007)
  - Call this: ε-admissible subset

GFD

$$r_{\boldsymbol{y}}(M) \propto \pi^{\frac{|M|}{2}} \Gamma\Big(\frac{n-|M|}{2}\Big) RSS_M^{-(\frac{n-|M|-1}{2})} E[h_M^{\varepsilon}(\beta_M^{\star})]$$

Observations:

- Expectation with respect to within model GFD (usual T)

GFD

$$r_{\boldsymbol{y}}(M) \propto \pi^{\frac{|M|}{2}} \Gamma\left(\frac{n-|M|}{2}\right) RSS_M^{-\left(\frac{n-|M|-1}{2}\right)} E[h_M^{\varepsilon}(\beta_M^{\star})]$$

Observations:

- Expectation with respect to within model GFD (usual T)
- $r_{\boldsymbol{y}}(M$ negligibly small for large models because of h, e.g., $|M| > n$ implies $r_{\boldsymbol{y}}(M) = 0$.

GFD

$$r_{\boldsymbol{y}}(M) \propto \pi^{\frac{|M|}{2}} \Gamma\left(\frac{n-|M|}{2}\right) RSS_M^{-\left(\frac{n-|M|-1}{2}\right)} E[h_M^{\varepsilon}(\beta_M^{\star})]$$

Observations:

- Expectation with respect to within model GFD (usual T)
- $r_{\boldsymbol{y}}(M$ negligibly small for large models because of h, e.g., $|M| > n$ implies $r_{\boldsymbol{y}}(M) = 0$.
- Implemented using Grouped Independence Metropolis Hastings (Andrieu & Roberts, 2009).

Main Result

Theorem Williams & H (2017+)
Suppose the true model is given by M_T. *Then under certain conditions, for a fixed positive constant* $\alpha < 1$,

$$r_{\boldsymbol{y}}(M_T) = \frac{r_{\boldsymbol{y}}(M_T)}{\sum_{j=1}^{n^\alpha} \sum_{M:|M|=j} r_{\boldsymbol{y}}(M)} \xrightarrow{P} 1 \text{ as } n, p \to \infty.$$

Some Conditions

- Number of Predictors: $\liminf_{\substack{n\to\infty \\ p\to\infty}} \frac{n^{1-\alpha}}{\log(p)} > 2$,

Some Conditions

- Number of Predictors: $\liminf_{\substack{n\to\infty \\ p\to\infty}} \frac{n^{1-\alpha}}{\log(p)} > 2$,
- For the true model/parameter $p_T < \log n^{\gamma}$

$$\varepsilon_{M_T} \leq \frac{1}{18} \|X^T(\mu_T - Xb_{min})\|_2^2$$

where b_{min} minimizes the norm subject to $\|b\|_0 \leq p_T - 1$.

Some Conditions

- Number of Predictors: $\liminf_{\substack{n\to\infty \\ p\to\infty}} \frac{n^{1-\alpha}}{\log(p)} > 2$,
- For the true model/parameter $p_T < \log n^\gamma$

$$\varepsilon_{M_T} \leq \frac{1}{18}\|X^T(\mu_T - Xb_{min})\|_2^2$$

where b_{min} minimizes the norm subject to $\|b\|_0 \leq p_T - 1$.

- For a large model $|M| > p_T$ and large enough n or p,

$$\frac{9}{2}\|X^T(H_M - H_{M(-1)})\mu_T\|_2^2 < \varepsilon_M,$$

where H_M and $H_{M(-1)}$ are the projection matrix for M and M with a covariate removed respectively.

Default ε

$$\varepsilon = \Lambda_M \widehat{\sigma}^2_M \left(\frac{n^{0.51}}{9} + |M| \frac{\log(p\pi)^{1.1}}{9} - p_T \right)_+,$$

- $\Lambda_M := \mathrm{tr}\left((H_M X)' H_M X\right)$ with $H_M := X_M (X_M' X_M)^{-1} X_M'$
- $\widehat{\sigma}^2_M := \mathsf{RSS}_M / (n - |M|)$

Default ε

$$\varepsilon = \Lambda_M \widehat{\sigma}^2_M \left(\frac{n^{0.51}}{9} + |M| \frac{\log(p\pi)^{1.1}}{9} - p_T \right)_+,$$

- $\Lambda_M := \operatorname{tr}\left((H_M X)' H_M X\right)$ with $H_M := X_M(X_M' X_M)^{-1} X_M'$
- $\widehat{\sigma}^2_M :=$ RSS$_M/(n - |M|)$
- Tuning parameter p_T represents belief about true $|M_T|$.

Simulation setup 1

- Generate 1000 data vectors y from linear model with $\beta^0_{M_o} = (-1.5, -1, -.8, -.6, .6, .8, 1, 1.5)'$, and $\sigma^0_{M_o} = 1$.

Simulation setup 1

- Generate 1000 data vectors y from linear model with $\beta^0_{M_o} = (-1.5, -1, -.8, -.6, .6, .8, 1, 1.5)'$, and $\sigma^0_{M_o} = 1$.
- The $n \times p$ design matrix X is generated with rows from the $\mathrm{N}_p(0, \Sigma)$ distribution, where the diagonal components $\Sigma_{ii} = 1$ and the off-diagonal components $\Sigma_{ij} = \rho$ for $i \neq j$.

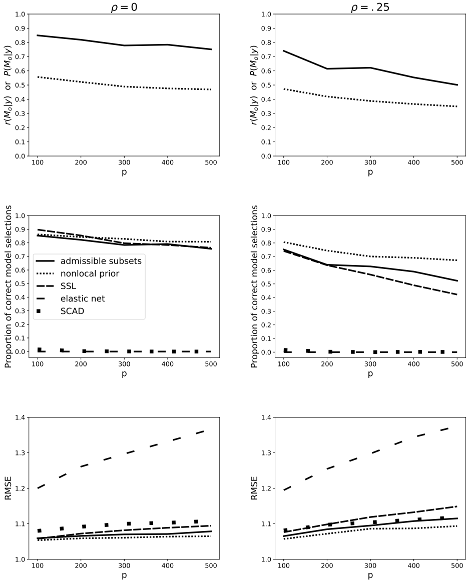
Simulation setup 1

- Generate 1000 data vectors y from linear model with $\beta^0_{M_o} = (-1.5, -1, -.8, -.6, .6, .8, 1, 1.5)'$, and $\sigma^0_{M_o} = 1$.
- The $n \times p$ design matrix X is generated with rows from the $\mathrm{N}_p(0, \Sigma)$ distribution, where the diagonal components $\Sigma_{ii} = 1$ and the off-diagonal components $\Sigma_{ij} = \rho$ for $i \neq j$.
- Implement 10-fold cross-validation scheme for choosing the tuning parameter p_o (prior to starting the algorithm).

Simulation setup 1

- Generate 1000 data vectors y from linear model with $\beta^0_{M_o} = (-1.5, -1, -.8, -.6, .6, .8, 1, 1.5)'$, and $\sigma^0_{M_o} = 1$.
- The $n \times p$ design matrix X is generated with rows from the $\mathsf{N}_p(0, \Sigma)$ distribution, where the diagonal components $\Sigma_{ii} = 1$ and the off-diagonal components $\Sigma_{ij} = \rho$ for $i \neq j$.
- Implement 10-fold cross-validation scheme for choosing the tuning parameter p_o (prior to starting the algorithm).
- Set $n = 100$, and consider $p = 100, 200, 300, 400, 500$.

Simulation results 1



Simulation setup 2

To illustrate the difference from the nonlocal prior approach, for $n = 30$, generate data from the following model.

$$Y \sim \mathsf{N}_n\Big(1 \cdot x^{(1)} + 1 \cdot x^{(2)} + \cdots + 1 \cdot x^{(9)}, I_n\Big),$$

where $x^{(1)}, x^{(2)}, x^{(3)} \overset{\text{iid}}{\sim} \mathsf{N}_n(0, I_n)$, and

$$\begin{array}{ccllllll}
x^{(4)} & \sim & \mathsf{N}_n\Big(& .25 \cdot x^{(1)} & & & & , .1^2 I_n\Big) \\
x^{(5)} & \sim & \mathsf{N}_n\Big(& & & .5 \cdot x^{(2)} & & , .1^2 I_n\Big) \\
x^{(6)} & \sim & \mathsf{N}_n\Big(& & & & - \; .75 \cdot x^{(3)} & , .1^2 I_n\Big) \\
x^{(7)} & \sim & \mathsf{N}_n\Big(& x^{(1)} & & & + \; x^{(3)} & , .1^2 I_n\Big) \\
x^{(8)} & \sim & \mathsf{N}_n\Big(& & & x^{(2)} & - \; x^{(3)} & , .1^2 I_n\Big) \\
x^{(9)} & \sim & \mathsf{N}_n\Big(& x^{(1)} & + & x^{(2)} & + \; x^{(3)} & , .1^2 I_n\Big)
\end{array}$$

Simulation results 2

	MAP size	RMSE	$P(M_{\text{MAP}}\|y)$
ε-*admissible* subsets	3.476	1.138	.365
nonlocal prior	8.997	1.197	.333

- RMSE of an out-of-sample test set of 30 observations
- Averaged over 1000 synthetic data sets

Simulation results 2

	MAP size	RMSE	$P(M_{\mathsf{MAP}}\|y)$
ε-*admissible* subsets	3.476	1.138	.365
nonlocal prior	8.997	1.197	.333

- RMSE of an out-of-sample test set of 30 observations
- Averaged over 1000 synthetic data sets
- Nonlocal prior procedure typically includes all 9 covariates even though the y can be mostly explained by 3.

Outline

Distributed Data

- DGE $\boldsymbol{X} = \boldsymbol{G}(U, \boldsymbol{\xi})$; do inference on $\boldsymbol{\xi}$

Distributed Data

- DGE $\boldsymbol{X} = \boldsymbol{G}(U, \boldsymbol{\xi})$; do inference on $\boldsymbol{\xi}$
- Issues:
  - n is so big that the $\boldsymbol{X}$'s cannot be loaded to one computer

Distributed Data

- DGE $\boldsymbol{X} = \boldsymbol{G}(U, \boldsymbol{\xi})$; do inference on $\boldsymbol{\xi}$
- Issues:
  - n is so big that the $\boldsymbol{X}$'s cannot be loaded to one computer
  - the data are at different sites

Distributed Data

- DGE $\boldsymbol{X} = \boldsymbol{G}(U, \boldsymbol{\xi})$; do inference on $\boldsymbol{\xi}$
- Issues:
  - n is so big that the $\boldsymbol{X}$'s cannot be loaded to one computer
  - the data are at different sites
  - data cannot be released off site for privacy concerns

Distributed Data

- DGE $\boldsymbol{X} = \boldsymbol{G}(U, \boldsymbol{\xi})$; do inference on $\boldsymbol{\xi}$
- Issues:
  - n is so big that the $\boldsymbol{X}$'s cannot be loaded to one computer
  - the data are at different sites
  - data cannot be released off site for privacy concerns
- partition $\boldsymbol{x}$ into K subsets, where each subsets can be analyzed

Distributed Data

- DGE $\boldsymbol{X} = \boldsymbol{G}(U, \boldsymbol{\xi})$; do inference on $\boldsymbol{\xi}$
- Issues:
  - n is so big that the $\boldsymbol{X}$'s cannot be loaded to one computer
  - the data are at different sites
  - data cannot be released off site for privacy concerns
- partition $\boldsymbol{x}$ into K subsets, where each subsets can be analyzed
  - e.g., use a computer cluster for parallel processing, where K is the number of nodes (or workers)

Distributed Data

- DGE $\boldsymbol{X} = \boldsymbol{G}(U, \boldsymbol{\xi})$; do inference on $\boldsymbol{\xi}$
- Issues:
  - n is so big that the $\boldsymbol{X}$'s cannot be loaded to one computer
  - the data are at different sites
  - data cannot be released off site for privacy concerns
- partition $\boldsymbol{x}$ into K subsets, where each subsets can be analyzed
  - e.g., use a computer cluster for parallel processing, where K is the number of nodes (or workers)
  - merge results from different nodes

Importance sampling for Massive Data

Importance sampling for Massive Data

- $\boldsymbol{x} = \boldsymbol{x}_1 \cup \boldsymbol{x}_2 \cup \ldots \cup \boldsymbol{x}_K$

Importance sampling for Massive Data

- $\boldsymbol{x} = \boldsymbol{x}_1 \cup \boldsymbol{x}_2 \cup \ldots \cup \boldsymbol{x}_K$
- $r_{\boldsymbol{x}}(\boldsymbol{\xi})$ – the generalized fiducial density of $\boldsymbol{x}$
- $r_{\boldsymbol{x}_k}(\boldsymbol{\xi})$ – the generalized fiducial density of $\boldsymbol{x}_k$

Importance sampling for Massive Data

- $\boldsymbol{x} = \boldsymbol{x}_1 \cup \boldsymbol{x}_2 \cup \ldots \cup \boldsymbol{x}_K$
- $r_{\boldsymbol{x}}(\boldsymbol{\xi})$ – the generalized fiducial density of $\boldsymbol{x}$
- $r_{\boldsymbol{x}_k}(\boldsymbol{\xi})$ – the generalized fiducial density of $\boldsymbol{x}_k$
- On each worker sample from $q_k(\boldsymbol{\xi})$

$$r_{\boldsymbol{x}}(\boldsymbol{\xi}) \propto \sum_k \text{'importance weight'} \times q_k(\boldsymbol{\xi})$$

Naive scheme

- generate a fiducial sample for data on each node (MCMC, SMC, ...)

Naive scheme

- generate a fiducial sample for data on each node (MCMC, SMC, ...)
- compute the weight $w_k(\boldsymbol{\xi}) = \frac{r_{\boldsymbol{x}}(\boldsymbol{\xi})}{r_{\boldsymbol{x}_k}(\boldsymbol{\xi})}$

Naive scheme

- generate a fiducial sample for data on each node (MCMC, SMC, ...)
- compute the weight $w_k(\boldsymbol{\xi}) = \frac{r_{\boldsymbol{x}}(\boldsymbol{\xi})}{r_{\boldsymbol{x}_k}(\boldsymbol{\xi})}$
- Not feasible and very inefficient!

Naive scheme

- generate a fiducial sample for data on each node (MCMC, SMC, ...)
- compute the weight $w_k(\boldsymbol{\xi}) = \frac{r_{\boldsymbol{x}}(\boldsymbol{\xi})}{r_{\boldsymbol{x}_k}(\boldsymbol{\xi})}$
- Not feasible and very inefficient!
  - Target of order $n^{-1/2}$
  - Fiducial sample on each worker of order $n_k^{-1/2}$.
  - Most realizations get extremely small weights.

Improved scheme

Improved scheme

- Each worker computes MLE $\hat{\theta}_k$ and empirical Fisher Information $\hat{I}_k$ and passes it to other workers

Improved scheme

- Each worker computes MLE $\hat{\theta}_k$ and empirical Fisher Information $\hat{I}_k$ and passes it to other workers
- Each worker simulates a sample from $q(\mathbf{x}_k) \propto r_{\boldsymbol{x}_k}(\boldsymbol{\xi}) \times \prod_{j \neq k} g(\boldsymbol{\xi} | \hat{\theta}_j, \hat{I}_j)$

Improved scheme

- Each worker computes MLE $\hat{\theta}_k$ and empirical Fisher Information $\hat{I}_k$ and passes it to other workers
- Each worker simulates a sample from $q(\mathbf{x}_k) \propto r_{\boldsymbol{x}_k}(\boldsymbol{\xi}) \times \prod_{j \neq k} g(\boldsymbol{\xi} | \hat{\theta}_j, \hat{I}_j)$
  - Practical choice $g \sim$ Normal$(\hat{\theta}_j, \gamma \hat{I}_k^{-1})$.

Improved scheme

- Each worker computes MLE $\hat{\theta}_k$ and empirical Fisher Information $\hat{I}_k$ and passes it to other workers
- Each worker simulates a sample from $q(\mathbf{x}_k) \propto r_{\boldsymbol{x}_k}(\boldsymbol{\xi}) \times \prod_{j \neq k} g(\boldsymbol{\xi} | \hat{\theta}_j, \hat{I}_j)$
  - Practical choice $g \sim$ Normal$(\hat{\theta}_j, \gamma \hat{I}_k^{-1})$.
- Weight $w_k(\boldsymbol{\xi}) \approx \prod_{j \neq k} \frac{f(\mathbf{x}_k, \boldsymbol{\xi})}{g(\boldsymbol{\xi} | \theta_j, I_j)}$
  (Computed and thinned in parallel.)

Improved scheme

- Each worker computes MLE $\hat{\theta}_k$ and empirical Fisher Information $\hat{I}_k$ and passes it to other workers
- Each worker simulates a sample from $q(\mathbf{x}_k) \propto r_{\boldsymbol{x}_k}(\boldsymbol{\xi}) \times \prod_{j \neq k} g(\boldsymbol{\xi}|\hat{\theta}_j, \hat{I}_j)$
  - Practical choice $g \sim$ Normal$(\hat{\theta}_j, \gamma \hat{I}_k^{-1})$.
- Weight $w_k(\boldsymbol{\xi}) \approx \prod_{j \neq k} \frac{f(\mathbf{x}_k, \boldsymbol{\xi})}{g(\boldsymbol{\xi}|\theta_j, I_j)}$
  (Computed and thinned in parallel.)

- We have shown consistency and asymptotic normality of the error of our importance sampling scheme.

Experiments

- All good performance
  - linear regression with Cauchy errors ($n = 10^4$, $K = 5$, $N_{\text{rep}} = 200$)

Experiments

- All good performance
  - linear regression with Cauchy errors ($n = 10^4$, $K = 5$, $N_{\mathrm{rep}} = 200$)
  - nonlinear regression $Y = (\beta_0 + \beta_1 X)^{-1} + \sigma Z$ ($n = 10^4$, $K = 5$, $N_{\mathrm{rep}} = 200$)

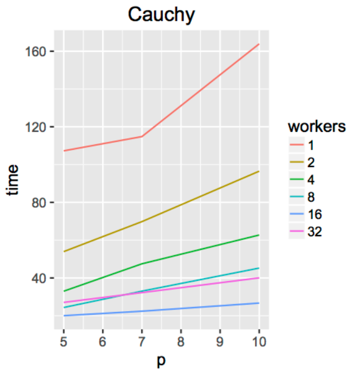
Experiments

- All good performance
  - linear regression with Cauchy errors ($n = 10^4$, $K = 5$, $N_{\text{rep}} = 200$)
  - nonlinear regression $Y = (\beta_0 + \beta_1 X)^{-1} + \sigma Z$ ($n = 10^4$, $K = 5$, $N_{\text{rep}} = 200$)
  - Gaussian mixture: $0.6N(\mu_0, \sigma) + 0.4N(\mu_1, \sigma)$ ($n = 10^4$, $K = 5$, $N_{\text{rep}} = 200$)

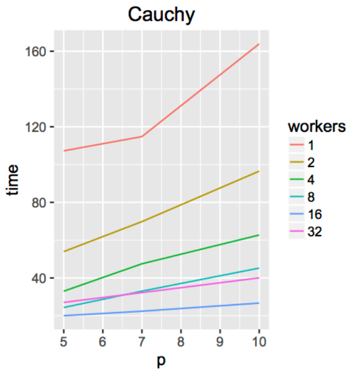
Experiments

- All good performance
  - linear regression with Cauchy errors ($n = 10^4$, $K = 5$, $N_{\text{rep}} = 200$)
  - nonlinear regression $Y = (\beta_0 + \beta_1 X)^{-1} + \sigma Z$ ($n = 10^4$, $K = 5$, $N_{\text{rep}} = 200$)
  - Gaussian mixture: $0.6N(\mu_0, \sigma) + 0.4N(\mu_1, \sigma)$ ($n = 10^4$, $K = 5$, $N_{\text{rep}} = 200$)
  - Generalized pareto, prediction of out of sample quantiles ($n = 10^6$, $K = 10$, $N_{\text{rep}} = 100$)

Computational time

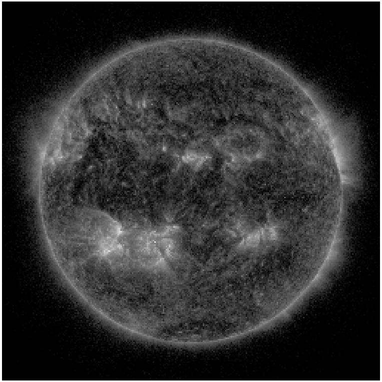


Computational time

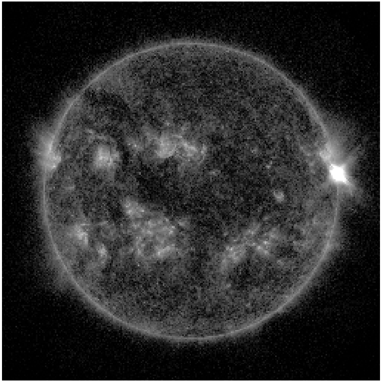


- Speed improves until $K = 16$ then deteriorates.
  (Cheng & Shang, 2015)

Sun Spots

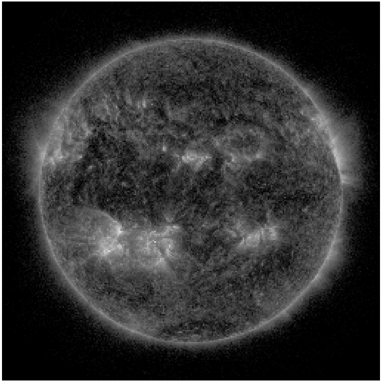


low activity

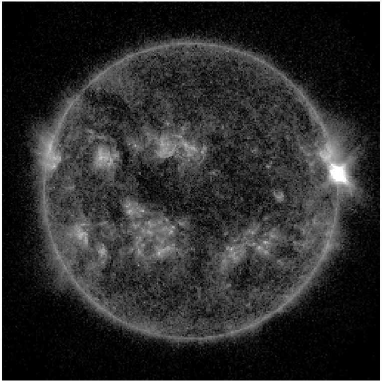


high activity

Sun Spots



low activity



high activity

- The bright flare on the right has value 253. Is this high?

Data

- Solar Dynamics Observatory (SDO), launched on 2010

Data

- Solar Dynamics Observatory (SDO), launched on 2010
- one instrument is Atmospheric Imaging Assembly (AIA) (Schuh et al. 2013)
  - photographs the sun in 8 wavelengths every 12s
  - image size: 4096×4096
  - 1.5 TB compressed data per day
  - same as 3 TB raw (i.e., uncompressed) data per day

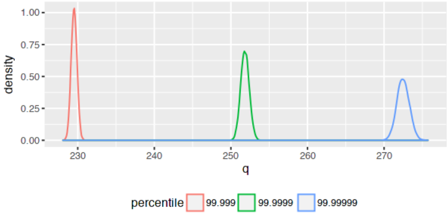
Data

- Solar Dynamics Observatory (SDO), launched on 2010
- one instrument is Atmospheric Imaging Assembly (AIA) (Schuh et al. 2013)
  - photographs the sun in 8 wavelengths every 12s
  - image size: 4096×4096
  - 1.5 TB compressed data per day
  - same as 3 TB raw (i.e., uncompressed) data per day
- ultimate goal: detect and predict solar flares

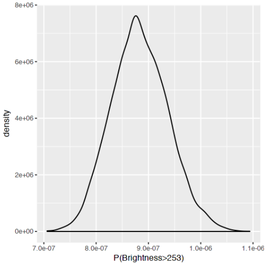
Data

- Solar Dynamics Observatory (SDO), launched on 2010
- one instrument is Atmospheric Imaging Assembly (AIA) (Schuh et al. 2013)
  - photographs the sun in 8 wavelengths every 12s
  - image size: 4096×4096
  - 1.5 TB compressed data per day
  - same as 3 TB raw (i.e., uncompressed) data per day
- ultimate goal: detect and predict solar flares
- Tool: GFD for Generalized Pareto (Wandler & H, 2012)

GFD for extreme quantiles



Large Quantiles



Fiducial probability of exceeding 253

Outline

Deep Neural Network (DNN)

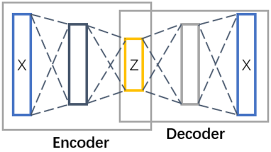
- Idea: Use deep neural network in fiducial computations

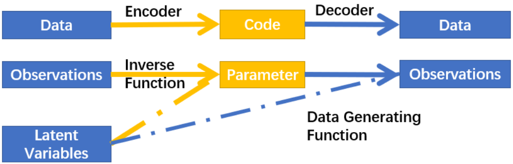
Deep Neural Network (DNN)

- Idea: Use deep neural network in fiducial computations
  - **Universal approximation theorem:** A large enough network with a linear output layer and at least one hidden layer can approximate any Borel measurable function.

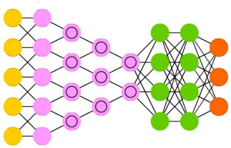
Deep Neural Network (DNN)

- Idea: Use deep neural network in fiducial computations
  - **Universal approximation theorem:** A large enough network with a linear output layer and at least one hidden layer can approximate any Borel measurable function.
  - Idea: Use Auto-encoder to approximate fiducial inverse



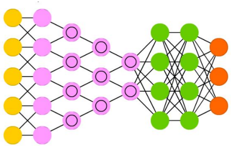


Challenges



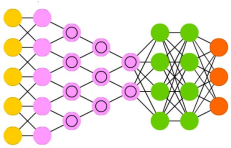
- A large number of choices
  - DNN architecture (fully connected, convolution, auto-encoder, adversarial + combination ...)

Challenges



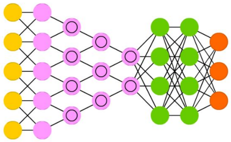
- A large number of choices
  - DNN architecture (fully connected, convolution, auto-encoder, adversarial + combination ...)
  - Number of layers, number of nodes per layers, activation function (RELU, sigmoid, softmax,...)

Challenges



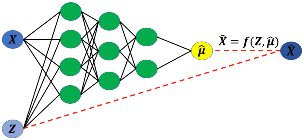
- A large number of choices
  - DNN architecture (fully connected, convolution, auto-encoder, adversarial + combination ...)
  - Number of layers, number of nodes per layers, activation function (RELU, sigmoid, softmax,...)
  - Optimization algorithm (stochastic gradient descent, Adaptive Subgradient Methods, ADAM (Kingma & Ba 2014), ...)

Challenges



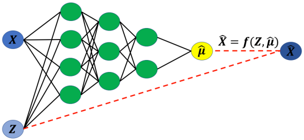
- A large number of choices
  - DNN architecture (fully connected, convolution, auto-encoder, adversarial + combination ...)
  - Number of layers, number of nodes per layers, activation function (RELU, sigmoid, softmax,...)
  - Optimization algorithm (stochastic gradient descent, Adaptive Subgradient Methods, ADAM (Kingma & Ba 2014), ...)
  - Host of other sensitivities (data generation, stopping rules, anti-over fitting measures,...)

Fiducial Auto Encoder



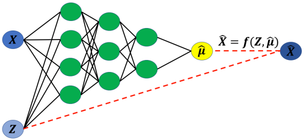
- Encoder: Fully connected layers,
- Decoder: DGE $\boldsymbol{X} = \boldsymbol{G}(\boldsymbol{Z}, \boldsymbol{\xi})$

Fiducial Auto Encoder



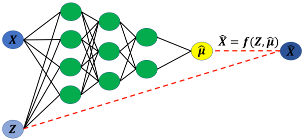
- Encoder: Fully connected layers,
- Decoder: DGE $\boldsymbol{X} = \boldsymbol{G}(\boldsymbol{Z}, \boldsymbol{\xi})$
- Training data: Generated from DGE with different values of $\boldsymbol{\xi}_t, \boldsymbol{Z}_t$.

Fiducial Auto Encoder



- Encoder: Fully connected layers,
- Decoder: DGE $\boldsymbol{X} = \boldsymbol{G}(\boldsymbol{Z}, \boldsymbol{\xi})$
- Training data: Generated from DGE with different values of $\boldsymbol{\xi}_t, \boldsymbol{Z}_t$.
- Loss function: $L = w_1\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2 + w_2\|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}\|^2$

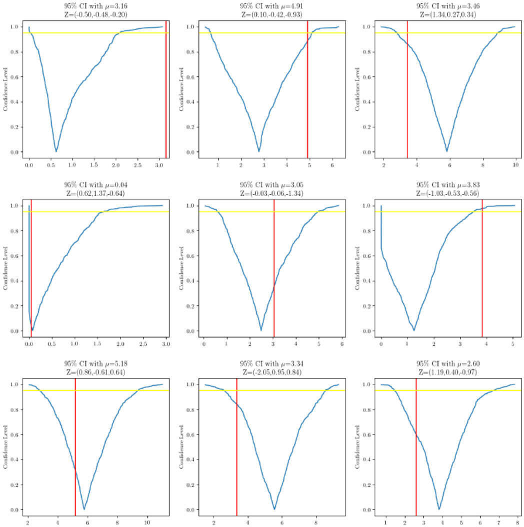
Fiducial Auto Encoder



- Encoder: Fully connected layers,
- Decoder: DGE $\boldsymbol{X} = \boldsymbol{G}(\boldsymbol{Z}, \boldsymbol{\xi})$
- Training data: Generated from DGE with different values of $\boldsymbol{\xi}_t, \boldsymbol{Z}_t$.
- Loss function: $L = w_1 \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2 + w_2 \|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}\|^2$
- Trained encoder used for inference

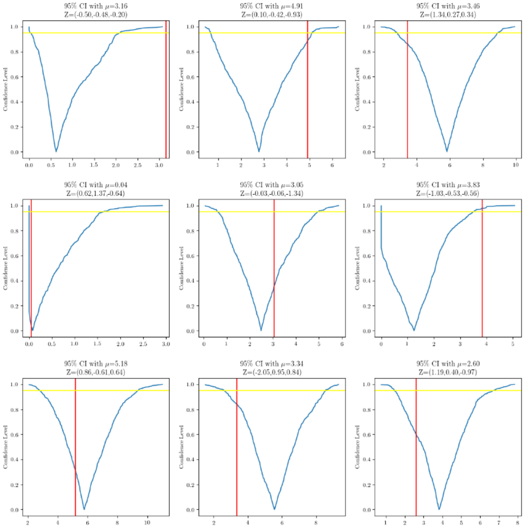
Inference

- Model: $\boldsymbol{X} = \mu + \mu^{q/2}\boldsymbol{Z}$
- Use encoder repeatedly
- Inputs: Observed $\boldsymbol{X}$, multiple independent $\boldsymbol{Z}^*$
- Output: Approximate fiducial sample μ^*



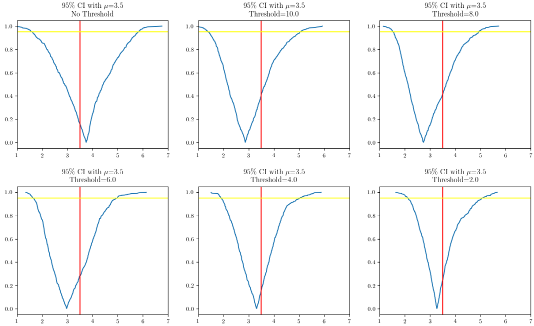
Inference

- Model:
  $\boldsymbol{X} = \mu + \mu^{q/2}\boldsymbol{Z}$
- Use encoder repeatedly
- Inputs: Observed $\boldsymbol{X}$, multiple independent $\boldsymbol{Z}^*$
- Output: Approximate fiducial sample μ^*
- Issues: conservative, biased



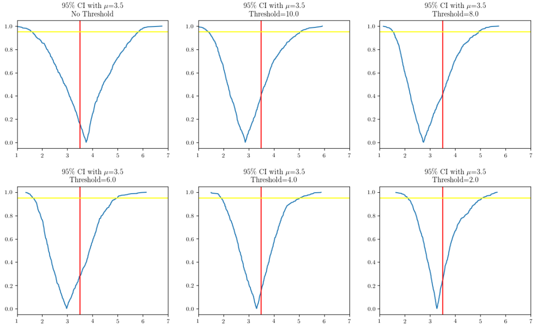
Approximate Fiducial Calculations

- Following AE: $\boldsymbol{X}^\star = \boldsymbol{G}(\boldsymbol{Z}^\star, \mu^\star)$ needs to replicate $\boldsymbol{X}$.
- Keep μ^* when $\|\boldsymbol{X}^* - \boldsymbol{X}\| \leq \epsilon$.
- Big improvement in coverage and length



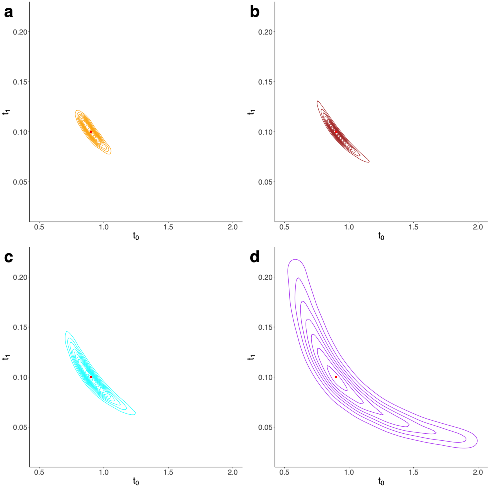
Approximate Fiducial Calculations

- Following AE: $\boldsymbol{X}^{\star} = \boldsymbol{G}(\boldsymbol{Z}^{\star}, \mu^{\star})$ needs to replicate $\boldsymbol{X}$.
- Keep μ^* when $\|\boldsymbol{X}^* - \boldsymbol{X}\| \leq \epsilon$.
- Big improvement in coverage and length
- Future work: GAN improve efficiency?



Biological Oxygen Demand

- $\boldsymbol{Y} = \xi_1(1 - e^{-\xi_2 \boldsymbol{X}}) + \boldsymbol{Z}$
- $\boldsymbol{x} = (2, 4, 6, 8, 10)$, $\boldsymbol{y} = (0.15, 0.30, 0.41, 0.48, 0.57)$, $\boldsymbol{Z} \sim N(0, 0.015I)$
- Methods:
  - a FAE
  - b GFD-HMC
  - c bootstrap
  - d Bayes ROT (Bardsley et al, 2014)



Outline

Forensic Science

- In criminal cases, experts encouraged to summarize evidence using LR (e.g., ENFSI guidelines)

Forensic Science

- In criminal cases, experts encouraged to summarize evidence using LR (e.g., ENFSI guidelines)

US FRE Rule 702

A witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if:

1. the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue;
2. the testimony is based on sufficient facts or data;
3. the testimony is the product of reliable principles and methods; and
4. the expert has reliably applied the principles and methods to the facts of the case.

https://www.law.cornell.edu/rules/fre/rule_702

Forensic Science

- In criminal cases, experts encouraged to summarize evidence using LR (e.g., ENFSI guidelines)

US FRE Rule 702

A witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if:

1. the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue;
2. the testimony is based on sufficient facts or data;
3. the testimony is the product of reliable principles and methods; and
4. the expert has reliably applied the principles and methods to the facts of the case.

https://www.law.cornell.edu/rules/fre/rule_702

- Reliable = Can be trusted

Examples of data

Likelihood ratio	Hotelling T^2/univariate kernel, equations (6)/(7)	Normal, equations (11)/(12)	MVK kernel, equations (14)/(15)
$>10^7$	0	1	0
10^7–10^6	0	3	0
10^6–10^5	0	8	0
10^5–10^4	0	9	0
10^4–10^3	22	11	16
10^3–10^2	10	17	19
10^2–10^1	12	5	13
10^1–1	5	8	5
1–10^{-1}	3	10	10
10^{-1}–10^{-2}	5	7	9
10^{-2}–10^{-3}	3	3	8
10^{-3}–10^{-4}	6	6	6
10^{-4}–10^{-5}	4	6	3
$<10^{-5}$	1821	1797	1802
Total	1891	1891	1891

Likelihood ratio	Hotelling T^2/univariate kernel, equations (6)/(7)	Normal, equations (11)/(12)	MVK kernel, equations (14)/(15)
<1	0	0	0
1–10^2	1	0	0
10^2–10^3	18	8	13
10^3–10^4	35	17	48
10^4–10^5	8	16	1
10^5–10^6	0	11	0
10^6–10^7	0	5	0
10^7–10^8	0	0	0
10^8–10^9	0	1	0
10^9–10^{10}	0	1	0
$>10^{10}$	0	3	0
Total	62	62	62

Figure: Glass evidence from Aitken & Lucy (2004)

not mated

mated

Examples of data

Likelihood ratio	*Hotelling T^2/univariate kernel, equations (6)/(7)*	*Normal, equations (11)/(12)*	*MVK kernel, equations (14)/(15)*
$>10^7$	0	1	0
10^7–10^6	0	3	0
10^6–10^5	0	8	0
10^5–10^4	0	9	0
10^4–10^3	22	11	16
10^3–10^2	10	17	19
10^2–10^1	12	5	13
10^1–1	5	8	5
1–10^{-1}	3	10	10
10^{-1}–10^{-2}	5	7	9
10^{-2}–10^{-3}	3	3	8
10^{-3}–10^{-4}	6	6	6
10^{-4}–10^{-5}	4	6	3
$<10^{-5}$	1821	1797	1802
Total	1891	1891	1891

Likelihood ratio	*Hotelling T^2/univariate kernel, equations (6)/(7)*	*Normal, equations (11)/(12)*	*MVK kernel, equations (14)/(15)*
<1	0	0	0
1–10^2	1	0	0
10^2–10^3	18	8	13
10^3–10^4	35	17	48
10^4–10^5	8	16	1
10^5–10^6	0	11	0
10^6–10^7	0	5	0
10^7–10^8	0	0	0
10^8–10^9	0	1	0
10^9–10^{10}	0	1	0
$>10^{10}$	0	3	0
Total	62	62	62

Figure: Glass evidence from Aitken & Lucy (2004)

not mated | mated

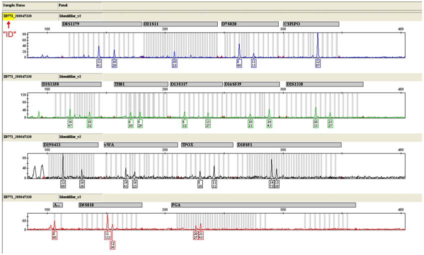
- Our mathematical abstraction:
  - Two streams of data (mated/non mated). Algorithms produce LR-like measure.

Well-calibrated?

- When is $1,000,000 : 1$ more like $100 : 1$? Does it matter?

Well-calibrated?

- When is $1,000,000 : 1$ more like $100 : 1$? Does it matter?
- The LR value may have an effect on verdict
  - Barrie et al (2018) report LR values across different labs of 172 to 3.2×10^{14} starting from the same EPG!



LR of LR = LR

- H_P : Defendant a contributor to the sample
  H_D : Defendant not a contributor to the sample

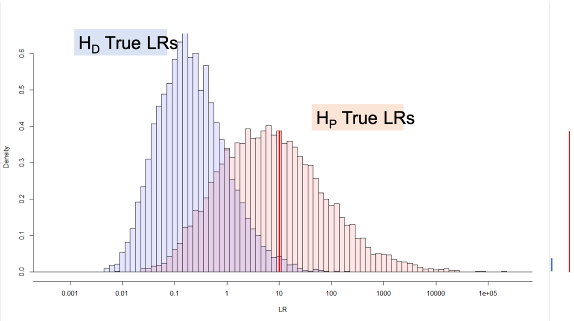
LR of LR = LR

- H_P : Defendant a contributor to the sample
  H_D : Defendant not a contributor to the sample
- Reported LR density: $g(l)$ under H_P, $f(l)$ under H_D.

LR of LR = LR

- H_P : Defendant a contributor to the sample
  H_D : Defendant not a contributor to the sample
- Reported LR density: $g(l)$ under H_P, $f(l)$ under H_D.
- Key observation: well-calibrated if and only if

$$g(l) = l f(l) \tag{2}$$



LR of LR = LR

- H_P : Defendant a contributor to the sample
  H_D : Defendant not a contributor to the sample
- Reported LR density: $g(l)$ under H_P, $f(l)$ under H_D.
- Key observation: well-calibrated if and only if

$$g(l) = l f(l) \tag{2}$$

- Integrating (2)

$$G(b) - G(a) = bF(b) - aF(a) - \int_a^b F(l)dl, \quad 0 < a < b < \infty.$$

Calibration Statistic

- Select grid a_i covering “mated data” (usually powers of 10)

Calibration Statistic

- Select grid a_i covering "mated data" (usually powers of 10)
- Define

$$d(G,F) = \left(\log_{10} \left(\frac{G(a_i) - G(a_{i-1})}{a_i F(a_i) - a_{i-1} F(a_{i-1}) - \int_{a_{i-1}}^{a_i} F(l)dl} \right), \quad i = 2, \ldots k \right)^\top$$

Calibration Statistic

- Select grid a_i covering "mated data" (usually powers of 10)
- Define

$$d(G,F) = \left(\log_{10}\left(\frac{G(a_i)-G(a_{i-1})}{a_iF(a_i)-a_{i-1}F(a_{i-1})-\int_{a_{i-1}}^{a_i}F(l)dl}\right), \quad i=2,\dots k\right)^\top$$

- Estimation and uncertainty quantification via GFD

Fiducial Non-Parametric

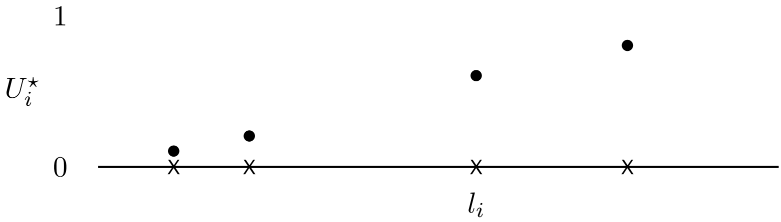
- Data generating equation $L_i = F^{-1}(U_i)$

Fiducial Non-Parametric

- Data generating equation $L_i = F^{-1}(U_i)$
- inverts to $\{F^* : F^*(l_i - \epsilon) < U_i^\star \leq F^*(l_i)\}$

Fiducial Non-Parametric

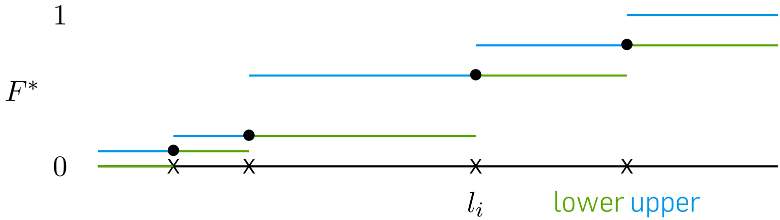
- Data generating equation $L_i = F^{-1}(U_i)$
- inverts to $\{F^* : F^*(l_i - \epsilon) < U_i^\star \leq F^*(l_i)\}$



U_i^* ordered uniforms

Fiducial Non-Parametric

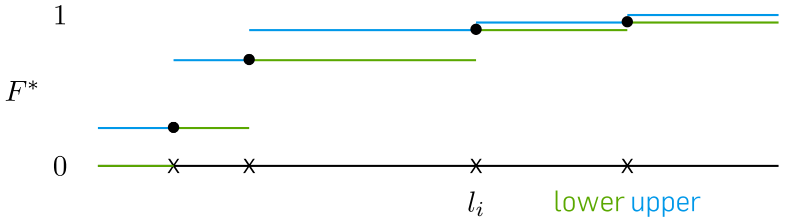
- Data generating equation $L_i = F^{-1}(U_i)$
- inverts to $\{F^* : F^*(l_i - \epsilon) < U_i^\star \leq F^*(l_i)\}$



F^* is any cdf between bounds

Fiducial Non-Parametric

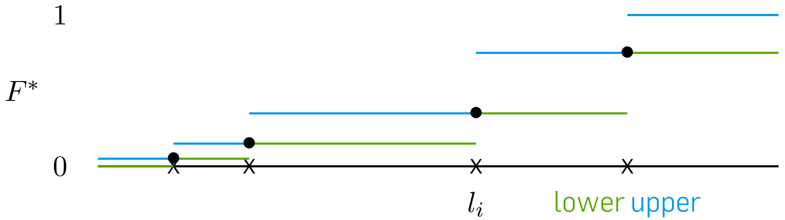
- Data generating equation $L_i = F^{-1}(U_i)$
- inverts to $\{F^* : F^*(l_i - \epsilon) < U_i^\star \leq F^*(l_i)\}$



F^* is any cdf between bounds

Fiducial Non-Parametric

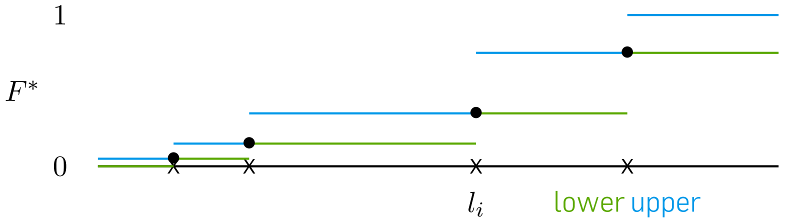
- Data generating equation $L_i = F^{-1}(U_i)$
- inverts to $\{F^* : F^*(l_i - \epsilon) < U_i^\star \leq F^*(l_i)\}$



F^* is any cdf between bounds

Fiducial Non-Parametric

- Data generating equation $L_i = F^{-1}(U_i)$
- inverts to $\{F^* : F^*(l_i - \epsilon) < U_i^\star \leq F^*(l_i)\}$



F^* is any cdf between bounds

- Facts (Cui & H, 2019)
  - $EF^\star_{lower}(l) < \hat{F}(l) < EF^\star_{upper}(l)$
  - Bernstein-von Mises theorem, good small sample properties

Calibration Confidence Intervals

- Recall

$$d(G,F) = \left(\log_{10}\left(\frac{G(a_i)-G(a_{i-1})}{a_iF(a_i)-a_{i-1}F(a_{i-1})-\int_{a_{i-1}}^{a_i}F(l)dl}\right), \quad i=2,\dots k\right)^\top$$

Theorem (H, Iyer, 2020+)

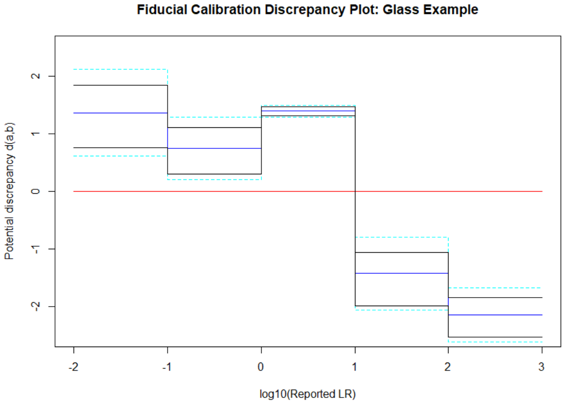
Assume obs LRs independent; $0 < F(a_1) < \cdots < F(a_k) < 1$, $0 < G(a_1) < \cdots < G(a_k) < 1$, $n = \min(n_g, n_f)$, $n/n_f \to p_f$, $n/n_g \to p_g$. *Then*

$$\sqrt{n}(d(\hat{G},\hat{F}) - d(G,F)) \xrightarrow{\mathcal{D}} N(0,\Sigma_{g,f}),$$

and conditionally on the observed LRs

$$\sqrt{n}(d(G^\star,F^\star) - d(\hat{G},\hat{F})) \xrightarrow{\mathcal{D}} N(0,\Sigma_{g,f}) \quad a.s.$$

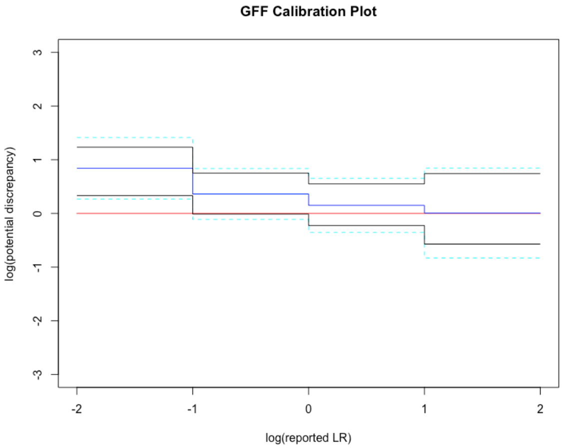
Calibration -- glass LR



Glass (classical)

target
median
fiducial CI
fid cband

Calibration -- glass LR

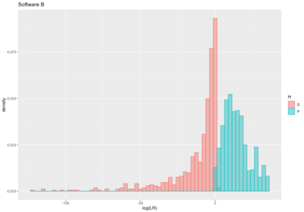


Glass (Williams, H., Omen)

target
median
fiducial CI
fid cband

Extrapolation via Generalized Pareto Distribution (GPD)

- DNA: Little overlap between mated and non-mated LR



Extrapolation via Generalized Pareto Distribution (GPD)

- DNA: Little overlap between mated and non-mated LR
- Data above large threshold follow GPD
  - GPD interpolates bounded, exponential and Pareto tails

$$f(x) = \frac{1}{\sigma}\left(1 + \frac{kx}{\sigma}\right)^{-1-1/k}, \quad x > 0 \quad \text{and if } k < 0 \text{ then also } x < -\frac{\sigma}{k}$$

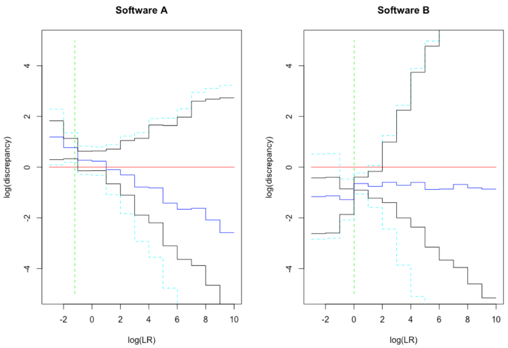
Extrapolation via Generalized Pareto Distribution (GPD)

- DNA: Little overlap between mated and non-mated LR
- Data above large threshold follow GPD
  - GPD interpolates bounded, exponential and Pareto tails

$$f(x) = \frac{1}{\sigma}\left(1 + \frac{kx}{\sigma}\right)^{-1-1/k}, \quad x > 0 \quad \text{and if } k < 0 \text{ then also } x < -\frac{\sigma}{k}$$

- Above threshold use GFD for GPD (Wandler & H, 2012)

DNA calibration



target
median
fiducial CI
threshold

Outline

BFF

- Many great minds contributed to foundations of statistics in the past – Fisher, Neyman, de Finetti, Lindley, Savage, LeCam, Cox, Efron, Berger, Fraser, Reid, Dempster, Dawid, ...

BFF

- Many great minds contributed to foundations of statistics in the past – Fisher, Neyman, de Finetti, Lindley, Savage, LeCam, Cox, Efron, Berger, Fraser, Reid, Dempster, Dawid, ...
  - Area was not known for harmonious relationships and respectful discourse

the "protracted battle" among leading statistics founding fathers "has left statistics without a philosophy that matches contemporary attitudes." *(Kass, 2011)*

BFF

- Many great minds contributed to foundations of statistics in the past – Fisher, Neyman, de Finetti, Lindley, Savage, LeCam, Cox, Efron, Berger, Fraser, Reid, Dempster, Dawid, ...
    - Area was not known for harmonious relationships and respectful discourse

the "protracted battle" among leading statistics founding fathers "has left statistics without a philosophy that matches contemporary attitudes." (Kass, 2011)

Can Bayesian, Fiducial and Frequentist become Best Friends Forever?

BFF Future

BFF Future

- Can BFF collaboration solve bring unique breakthroughs?
  - Understanding of Deep Learning?

BFF Future

- Can BFF collaboration solve bring unique breakthroughs?
  - Understanding of Deep Learning?
- Computational convenience and efficiency
  - Presence in Probabilistic Programing Languages
  - Scalability

BFF Future

- Can BFF collaboration solve bring unique breakthroughs?
  - Understanding of Deep Learning?
- Computational convenience and efficiency
  - Presence in Probabilistic Programing Languages
  - Scalability
- New kind of theoretical guarantees
  - How safely and efficiently marginalize Bayes/fiducial/confidence distributions in complex models?
  - Going beyond probability theory?

BFF Future

- Can BFF collaboration solve bring unique breakthroughs?
  - Understanding of Deep Learning?
- Computational convenience and efficiency
  - Presence in Probabilistic Programing Languages
  - Scalability
- New kind of theoretical guarantees
  - How safely and efficiently marginalize Bayes/fiducial/confidence distributions in complex models?
  - Going beyond probability theory?
- Applications
  - The proof is in the pudding!

I have a dream …

I have a dream ...

- One famous statistician said (I paraphrase)

  "I use Bayes because there is no need to prove asymptotic theorem; it is correct."

I have a dream …

- One famous statistician said (I paraphrase)

  "I use Bayes because there is no need to prove asymptotic theorem; it is correct."

- I have a dream that people will one day soon gain similar trust in fiducial inspired approaches.

I have a dream ...

- One famous statistician said (I paraphrase)

  "I use Bayes because there is no need to prove asymptotic theorem; it is correct."

- I have a dream that people will one day soon gain similar trust in fiducial inspired approaches.

Thank you!