Extended Fiducial Inference: Toward an Automated Process of Statistical Inference

Faming Liang

Purdue University

November 11, 2024

Statistical Inference

Statistical inference is a fundamental task in modern data science, which studies how to propagate the uncertainty embedded in data to model parameters.

Frameworks of Statistical Inference:

- Frequentist
- Bayes
- Fiducial

Statistical Inference: Frequentist Methods

The frequentist methods often estimate model parameters using the maximum likelihood approach and test hypotheses by comparing a test statistic with a known theoretical reference distribution.

- Parameter estimation: MLE can be significantly influenced by outliers, reducing the fidelity of parameter estimation.
- Hypothesis testing: the required theoretical reference distribution is test statistic-dependent, making statistical inference difficult to automate.
- Large sample theory: the sample size required to achieve asymptotic normality can be very large especially in high-dimensional scenarios.

Statistical Inference: Bayesian Methods

The dependence of Bayesian inference on the prior distribution has been a subject of criticism throughout the history of Bayesian statistics, often raising concerns about their fidelity.

Statistical Inference: Fiducial

Fiducial inference was generally regarded as a big blunder by Fisher. However, the goal he initially set, *making inference about unknown parameters on the basis of observations* (Fisher, 1956), has been continually pursued by many statisticians.

Building on our early works in sparse deep learning (Liang et al, 2018; Sun, Song, and Liang, 2022) and adaptive stochastic gradient Markov chain Monte Carlo (MCMC) (Liang, Sun, and Liang, 2022), we develop a new statistical inference framework called the *extended fiducial inference* (EFI), which achieves the initial goal of fiducial inference while possessing necessary features like *fidelity*, *automaticity*, and *scalability* that are essential for statistical inference in modern data science.

Data generating equations: Uncertainty

Consider a dataset of regression: $\{(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)\}$. In the view of structural inference, we can express the observations in the data generating equation:

$$y_i = f(x_i, z_i, \boldsymbol{\theta}), \quad i = 1, 2, \ldots, n, \tag{1}$$

where $\boldsymbol{\theta} \in \mathbb{R}^p$ and z_i represents a random error (latent variable). The system consists of $n + p$ unknowns: $\{\boldsymbol{\theta}, z_1, z_2, \ldots, z_n\}$, while there are only n equations. Therefore, the values of $\boldsymbol{\theta}$ cannot be uniquely determined by the data-generating equation, which gives the source of uncertainty of parameters.

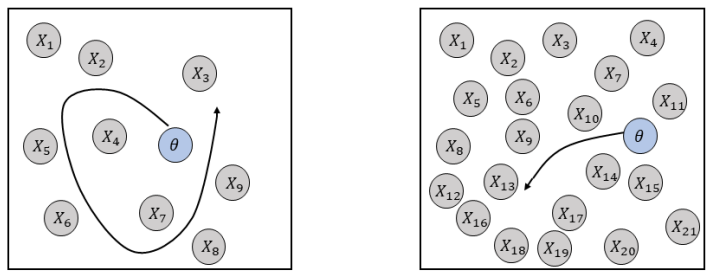

Figure 1: Illustration for the source of uncertainty of model parameters.

Data generating equations: Frequentist

To solve for $\boldsymbol{\theta}$ from the undetermined system (1), the frequentist methods often impose a constraint on the system such that the latent variables can be dismissed and $\boldsymbol{\theta}$ can be uniquely determined.

- The MLE method assumes the likelihood of $\{z_1, z_2, \ldots, z_n\}$ is maximized:

$$\prod_{i=1}^{n} \phi(z_i) = \max_{(\tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_n) \in \mathbb{R}^n} \prod_{i=1}^{n} \phi(\tilde{z}_i), \tag{2}$$

  which is equivalent to solving the optimization problem:

$$\max_{(\boldsymbol{\beta}, \sigma)} \sum_{i=1}^{n} \log \phi\Big(\frac{y_i - x_i^T \boldsymbol{\beta}}{\sigma}\Big). \tag{3}$$

- Moment estimation:
  $\sum_{i=1}^{n} y_i^k = \sum_{i=1}^{n} \int [f(x_i, z, \boldsymbol{\theta})]^k \pi_0(z) dz, \quad i = 1, 2, \ldots, p.$

Data generating equations: Bayesian

Bayesian methods treat $\boldsymbol{\theta}$ as random variables and circumvent the issue of latent variables by adopting a conditional approach:

$$\pi(\boldsymbol{\theta}|(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)) = \frac{\prod_{i=1}^{n} p(y_i|x_i, \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int \prod_{i=1}^{n} p(y_i|x_i, \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}, \tag{4}$$

The dependence of the inference on the prior distribution has been subject to criticism throughout the history of Bayesian statistics, as the prior distribution introduces subjective elements that may affect the fidelity of statistical inference.

Data generating equations: Generalized Fiducial Inference

GFI (Hannig, 2009; Hannig et al.,2016) provides an early method to solve the data generating equation:

(a) (Proposal) Generate $\tilde{\boldsymbol{Z}}_n = (\tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_n)^T$ from the Gaussian distribution $N(0, I_n)$.

(b) ($\boldsymbol{\theta}$-fitting) Find the best fitting parameters $\tilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \| \boldsymbol{Y}_n - \boldsymbol{X}_n\boldsymbol{\beta} - \sigma\tilde{\boldsymbol{Z}}_n \|$, and compute the fitted value $\tilde{\boldsymbol{Y}}_n = \boldsymbol{X}_n\tilde{\boldsymbol{\beta}} + \tilde{\sigma}\tilde{\boldsymbol{Z}}_n$.

(c) (Acceptance-rejection) Accept $\tilde{\boldsymbol{\theta}}$ if $\| \boldsymbol{Y}_n - \tilde{\boldsymbol{Y}}_n \| \leq \epsilon$ for some pre-specified small value ϵ, and reject otherwise.

Subsequently, statistical inference is made based on the accepted samples of $\tilde{\boldsymbol{\theta}}$.

Data generating equations: Extended Fiducial Inference

EFI treats $\boldsymbol{\theta}$ as fixed unknowns. Let $\boldsymbol{Z}_n := \{z_1, z_2, \ldots, z_n\}$ denote the collection of latent variables, and let $G(\boldsymbol{Y}_n, \boldsymbol{X}_n, \boldsymbol{Z}_n)$ denote an inverse function for the solution of $\boldsymbol{\theta}$ in the system (1).

- EFI jointly imputes $\boldsymbol{Z}_n$ and estimates $G(\boldsymbol{Y}_n, \boldsymbol{X}_n, \boldsymbol{Z}_n)$, and then quantifies the uncertainty of $\boldsymbol{\theta}$ based the estimated inverse function and the imputed values of $\boldsymbol{Z}_n$, where the estimated inverse function serves as an uncertainty propagator from $\boldsymbol{Z}_n$ to $\boldsymbol{\theta}$.
- Technically, EFI approximates $G(\boldsymbol{Y}_n, \boldsymbol{X}_n, \boldsymbol{Z}_n)$ using a sparse deep neural network (DNN), and employs an adaptive stochastic gradient MCMC algorithm to jointly impute $\boldsymbol{Z}_n$ and estimate the parameters of the sparse DNN.

Data generating equations: Extended Fiducial Inference

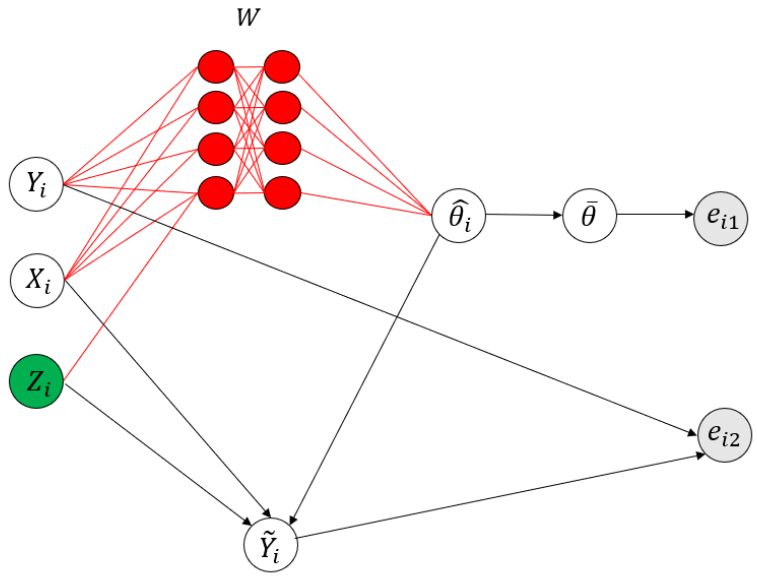

Figure 2: Illustration of the EFI network, where the red nodes and links form a DNN (parameterized by the weights $\boldsymbol{w}$) to learn, the green node represents latent variables to impute, and the black lines represent deterministic functions.

Extended Fiducial Inference: Theory-1

Assumption 1

There exists an inverse function $G : \mathbb{R}^n \times \mathbb{R}^{n \times d} \times \mathbb{R}^n \to \mathbb{R}^p$*:* $\boldsymbol{\theta} = G(\boldsymbol{Y}_n, \boldsymbol{X}_n, \boldsymbol{Z}_n)$.

Assumption 2

For a given inverse function, we define an energy function $U_n(\boldsymbol{z}) := U(\boldsymbol{Y}_n, \boldsymbol{X}_n, \boldsymbol{z}, G(\cdot))$ *satisfying the conditions:* $U_n(\cdot) \geq 0$, $\min_{\boldsymbol{z}} U_n(\boldsymbol{z})$ *exists and equals 0, and* $U_n(\boldsymbol{z}) = 0$ *if and only if* $\boldsymbol{Y}_n = f(\boldsymbol{X}_n, \boldsymbol{z}, G(\boldsymbol{Y}_n, \boldsymbol{X}_n, \boldsymbol{z}))$.

Let $\mathcal{Z}_n$ denote the zero-energy set $\mathcal{Z}_n = \left\{\boldsymbol{z} \in \mathbb{R}^n : U_n(\boldsymbol{z}) = 0\right\}$.

Lemma 1

If Assumptions 1-2 hold, then the zero-energy set $\mathcal{Z}_n$ *is invariant to the choice of* $G(\cdot)$.

Extended Fiducial Inference: Theory-2

Let $p_n^*(\boldsymbol{z}|\boldsymbol{Y}_n, \boldsymbol{X}_n)$ denote the extended fiducial density function of $\boldsymbol{Z}_n$ on $\mathcal{Z}_n$. We define

$$p_n^{\epsilon}(\boldsymbol{z}|\boldsymbol{X}_n, \boldsymbol{Y}_n) \propto \exp\left\{-\frac{U_n(\boldsymbol{z})}{\epsilon}\right\} \pi_0^{\otimes n}(\boldsymbol{z}), \tag{5}$$

where $\epsilon > 0$ represents the temperature, and $\pi_0^{\otimes n}(\boldsymbol{z}) = \pi_0(z_1) \times \pi_0(z_2) \times \cdots \times \pi_0(z_n)$.

Furthermore, we define $p_n^*(\boldsymbol{z}|\boldsymbol{Y}_n, \boldsymbol{X}_n)$ as the limit

$$p_n^* = \lim_{\epsilon \downarrow 0} p_n^{\epsilon}, \tag{6}$$

whose convergence can be studied as in Hwang (1980).

Extended Fiducial Inference: Theory-3

Let $\Pi_n(\cdot)$ denote a probability measure on $(\mathbb{R}^n, \mathscr{R})$ with $\mathscr{R}$ being the Borel σ-algebra and the corresponding density function given by $\pi_0^{\otimes n}$.

The convergence in (6) can be studied in two cases:

(a) $\Pi_n(\mathcal{Z}_n) > 0$;

(b) $\Pi_n(\mathcal{Z}_n) = 0$.

Extended Fiducial Inference: Theory-4

Assumption 3

$\Pi_n(U_n(\boldsymbol{z}) < a) > 0$ *for any* $a > 0$.

Theorem 2

If Assumptions 1-3 hold and $\Pi_n(\mathcal{Z}_n) > 0$, *then* $p_n^*(\boldsymbol{z}|\boldsymbol{X}_n, \boldsymbol{Y}_n)$ *is invariant to the choice of the inverse function* $G(\cdot)$ *and the energy function* $U_n(\cdot)$, *and it is given by*

$$\frac{dP_n^*(\boldsymbol{z}|\boldsymbol{X}_n, \boldsymbol{Y}_n)}{d\boldsymbol{z}} = \frac{1}{\Pi_n(\mathcal{Z}_n)} \pi_0^{\otimes n}(\boldsymbol{z}), \quad \boldsymbol{z} \in \mathcal{Z}_n, \tag{7}$$

where P_n^* *represents the cumulative distribution function (CDF) corresponding to* p_n^*.

An example for this case is logistic or multinomial logistic regression, for which the resulting estimator of $\boldsymbol{\theta}$ is a set.

Extended Fiducial Inference: Theory-5

Assumption 4

(i) There exists $a > 0$ such that $\{U_n(\boldsymbol{z}) \leq a\}$ is compact; (ii) $\pi_0^{\otimes n}(\boldsymbol{z})$ is continuous, and $U_n(\boldsymbol{z}) \in C^3(\mathbb{R}^n)$ is three-time continuously differentiable; (iii) $\mathcal{Z}_n$ has finitely many components and each component is a compact smooth manifold with the highest dimension p; (iv) $\pi_0^{\otimes n}(\boldsymbol{z})$ is not identically zero on the p-dimensional manifold, and $\det(\frac{\partial^2 U}{\partial \boldsymbol{t}^2}(\boldsymbol{z})) \neq 0$ for $\boldsymbol{z} \in \mathcal{Z}_n$.

Lemma 3

If Assumptions 1-4 hold, then the limiting probability measure p_n^* concentrates on the highest dimensional manifold and is given by

$$\frac{dP_n^*(\boldsymbol{z}|\boldsymbol{X}_n, \boldsymbol{Y}_n)}{d\nu}(\boldsymbol{z}) = \frac{\pi_0^{\otimes n}(\boldsymbol{z}) \left(\det(\nabla_{\boldsymbol{t}}^2 U_n(\boldsymbol{z}))\right)^{-1/2}}{\int_{\mathcal{Z}_n} \pi_0^{\otimes n}(\boldsymbol{z}) \left(\det(\nabla_{\boldsymbol{t}}^2 U_n(\boldsymbol{z}))\right)^{-1/2} d\nu}, \quad \boldsymbol{z} \in \mathcal{Z}_n, \quad (8)$$

where ν is the sum of intrinsic measures on the p-dimensional manifold in $\mathcal{Z}_n$.

Extended Fiducial Inference: Extended fiducial density (EFD)

Given the inverse function $G(\cdot)$, we define the parameter space

$$\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^p : \boldsymbol{\theta} = G(\boldsymbol{Y}_n, \boldsymbol{X}_n, \boldsymbol{z}), \boldsymbol{z} \in \mathcal{Z}_n\}.$$

Definition 4

For any function $b(\boldsymbol{\theta})$ of interest, its EFD associated with the inverse function $G(\cdot)$ is defined as

$$\mu_n^*(B|\boldsymbol{Y}_n, \boldsymbol{X}_n) = \int_{\mathcal{Z}_n(B)} dP_n^*(\boldsymbol{z}|\boldsymbol{Y}_n, \boldsymbol{X}_n), \quad \text{for any measurable set } B \subset \Theta, \tag{9}$$

where $\mathcal{Z}_n(B) = \{\boldsymbol{z} \in \mathcal{Z}_n : b(G(\boldsymbol{Y}_n, \boldsymbol{X}_n, \boldsymbol{z})) \in B\}$.

Extended Fiducial Inference: Simulation Procedure

(a) (Manifold Sampling) For any given inverse function $\tilde{G}(\cdot)$, simulate M samples, denoted by $\mathcal{S}_M = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_M\}$, from $\pi_0^{\otimes n}(\boldsymbol{z})$ subject to the constraint $U_n(\boldsymbol{z}) = 0$. This can be done using a constrained Monte Carlo algorithm such as constrained Hamiltonian Monte Carlo.

(b) (Weighting) Calculate the importance weight $\omega_i = \left(\det(\nabla^2_{\boldsymbol{t}} U_n(\boldsymbol{z}_i))\right)^{-1/2}$ for each sample $\boldsymbol{z}_i \in \mathcal{S}$ using an inverse function $G(\cdot)$ of interest.

(c) (Resampling) Draw m samples from $\mathcal{S}_M$ with replacement according to the probabilities: $\frac{\omega_i}{\sum_{j=1}^{M} \omega_j}$ for $i = 1, 2, \ldots, M$.

(d) (Inference) For $b(\boldsymbol{\theta})$, find the EFD associated with $G(\cdot)$ according to (9) based on the m samples obtained in step (c).

Extended Fiducial Inference: Flexibility

Remark: EFI provides a flexible framework of statistical inference. One can adjust the inverse function $G(\cdot)$ and the energy function $U_n(\cdot)$ to ensure that the resulting fiducial distribution estimator of $b(\boldsymbol{\theta})$ satisfies desired properties, such as *efficiency, unbiasedness*, and *robustness*.

This mirrors the flexibility of frequentist methods, where different estimators of $b(\boldsymbol{\theta})$ can be designed for different purposes. However, its conditional inference nature makes EFI even more attractive than frequentist methods, as it circumvents the need for derivations of theoretical distributions of the estimators.

Extended Fiducial Inference: Additive noise models

Assumption 5

(i) $U_n(\cdot)$ *is specified in the form:* $U_n(\boldsymbol{z}) = h(J(\boldsymbol{z})) = \sum_{i=1}^{n} h(e_i)$ *for some function* $h(\cdot)$ *satisfying* $\frac{\partial h(J)}{\partial J}(\boldsymbol{z}) = 0$ *for any* $\boldsymbol{z} \in \mathcal{Z}_n$, *where* $J(\boldsymbol{z}) = \boldsymbol{Y}_n - f(\boldsymbol{X}_n, \boldsymbol{z}, G(\boldsymbol{Y}_n, \boldsymbol{X}_n, \boldsymbol{z})) = (e_1, e_2, \ldots, e_n)^T$, *and* $e_i = y_i - f(x_i, z_i, \boldsymbol{\theta})$ *for* $i = 1, 2, \ldots, n$; *and (ii) the model noise is additive; i.e., the function* $f(X, Z, \boldsymbol{\theta})$ *in model (1) is a linear function of* Z.

Theorem 5

If Assumptions 1-5 hold, then $P_n^*(\boldsymbol{z}|\boldsymbol{Y}_n, \boldsymbol{X}_n)$ *given in (8) is invariant to the choices of* $G(\cdot)$ *and* $U_n(\cdot)$. *Furthermore,* P_n^* *reduces to a truncated distribution of* $\pi_0^{\otimes n}$ *on the manifold* $\mathcal{Z}_n$.

Extended Fiducial Inference: Linear Regression Example

Consider the linear regression model $\boldsymbol{Y}_n = \boldsymbol{X}_n\boldsymbol{\beta} + \sigma\boldsymbol{Z}_n$.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$. To conduct EFI for $\boldsymbol{\theta}$, we set $G(\boldsymbol{Y}_n, \boldsymbol{X}_n, \boldsymbol{z}) = \hat{\boldsymbol{\theta}}(z_1, z_2, \ldots, z_p) := (\hat{\boldsymbol{\beta}}^T, \hat{\sigma})^T$, a solver for the first p equations in (1), and set the energy function

$$U_n(\boldsymbol{z}) = \|\boldsymbol{Y}_n - f(\boldsymbol{X}_n, \boldsymbol{z}, G(\boldsymbol{Y}_n, \boldsymbol{X}_n, \boldsymbol{z}))\|^2. \tag{10}$$

The resulting EFD of σ^2 is given by

$$\mu_n^*(\sigma^2|\boldsymbol{Y}_n, \boldsymbol{X}_n) = \pi_{\chi^{-2}_{n-p+1}}\left(\frac{\sigma^2}{A}\right)\frac{1}{A}, \tag{11}$$

where A serves as an unbiased estimator of $(n-p+1)\sigma^2$.

If we use the mean of $\mu_n^*(\sigma^2|\boldsymbol{Y}_n, \boldsymbol{X}_n)$ as an estimator of σ^2, it has a bias of $\frac{2}{n-p-3}\sigma^2$. In contrast, the MLE of σ^2 has a bias of $-\frac{p-1}{n}\sigma^2$. Therefore, the EFD results in a smaller bias than the MLE when $n > (p+3)(p-1)/(p-3)$.

Extended Fiducial Inference: Linear Regression Example

The EFD of $\boldsymbol{\beta}$ can be obtained by completing the integration:

$$\mu_n^*(\boldsymbol{\beta}|\boldsymbol{Y}_n, \boldsymbol{X}_n) = \int \mu_n^*(\boldsymbol{\beta}|\boldsymbol{Y}_n, \boldsymbol{X}_n, \sigma^2)\mu_n^*(\sigma^2|\boldsymbol{Y}_n, \boldsymbol{X}_n)d\sigma^2, \quad (12)$$

which is a multivariate non-central t-distribution $t(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}, \nu_{\boldsymbol{\beta}})$ with the parameters given by

$$\boldsymbol{\mu}_{\boldsymbol{\beta}} = (\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T\boldsymbol{Y}_n, \quad \boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \frac{A}{n-p+1}(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}, \quad \nu_{\boldsymbol{\beta}} = n-p+1.$$

The mean and covariance matrix of the EFD is given by $(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^T\boldsymbol{Y}_n$ and $\frac{A}{n-p-1}(\boldsymbol{X}_n^T\boldsymbol{X}_n)^{-1}$.

EFI with a sparse DNN inverse function

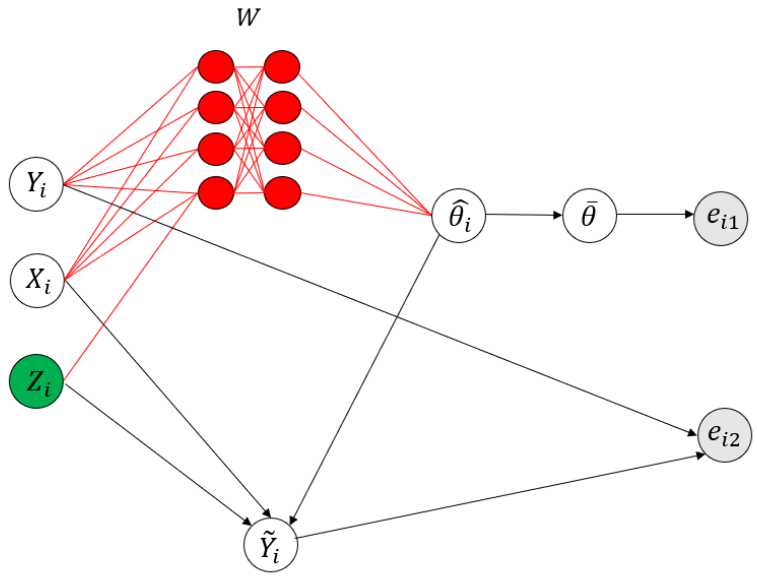

Figure 3: Illustration of the EFI network

A neural network estimator of the inverse function $G(\boldsymbol{Y}_n, \boldsymbol{X}_n, \boldsymbol{Z}_n)$:

$$\bar{\boldsymbol{\theta}}_n := \frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\theta}}_i = \frac{1}{n}\sum_{i=1}^{n}\hat{g}(y_i, x_i, z_i, \boldsymbol{w}), \tag{13}$$

where $\boldsymbol{w}$ denotes the parameters of the neural network.

EFI with a sparse DNN inverse function

We estimate $\boldsymbol{w}$ by solving the equation

$$\nabla_{\boldsymbol{w}} \log \pi(\boldsymbol{w}|\boldsymbol{X}_n, \boldsymbol{Y}_n) = 0, \tag{14}$$

which, under the Bayesian context, can be expressed as

$$\nabla_{\boldsymbol{w}} \log \pi(\boldsymbol{w}|\boldsymbol{X}_n, \boldsymbol{Y}_n) = \int \nabla_{\boldsymbol{w}} \log \pi(\boldsymbol{w}|\boldsymbol{X}_n, \boldsymbol{Y}_n, \boldsymbol{Z}_n) \pi(\boldsymbol{Z}_n|\boldsymbol{X}_n, \boldsymbol{Y}_n, \boldsymbol{w}) d\boldsymbol{w} = 0. \tag{15}$$

It can be solved using an adaptive stochastic gradient MCMC algorithm (Liang, Sun, Liang, 2022).

EFI with a sparse DNN inverse function

We define an energy function:

$$\tilde{U}_n(\boldsymbol{Z}_n, \boldsymbol{w}; \boldsymbol{X}_n, \boldsymbol{Y}_n) = \eta \sum_{i=1}^{n} \|\hat{\boldsymbol{\theta}}_i - \bar{\boldsymbol{\theta}}_n\|^2 + \sum_{i=1}^{n} d(y_i, x_i, z_i, \hat{\boldsymbol{\theta}}_i), \quad (16)$$

where the first term serves as a penalty function enforcing $\hat{\boldsymbol{\theta}}_i$'s to converge to the same value, and $\eta > 0$ is a regularization parameter.

$$\begin{aligned} \pi(\boldsymbol{w}|\boldsymbol{X}_n, \boldsymbol{Y}_n, \boldsymbol{Z}_n) &\propto \pi(\boldsymbol{w}) e^{-\lambda \tilde{U}_n(\boldsymbol{Z}_n, \boldsymbol{w}; \boldsymbol{X}_n, \boldsymbol{Y}_n)}, \\ \pi(\boldsymbol{Z}_n|\boldsymbol{X}_n, \boldsymbol{Y}_n, \boldsymbol{w}) &\propto \pi_0^{\otimes n}(\boldsymbol{Z}_n) e^{-\lambda \tilde{U}_n(\boldsymbol{Z}_n, \boldsymbol{w}; \boldsymbol{X}_n, \boldsymbol{Y}_n)}, \end{aligned} \quad (17)$$

where λ is a tuning parameter resembling the inverse of the temperature in (5).

Extended Fiducial Inference: Algorithm

(a) (*Latent variable sampling*) Generate $\boldsymbol{Z}_n^{(k+1)}$ from a transition kernel induced by a stochastic gradient MCMC algorithm. For example, we can simulate $\boldsymbol{Z}_n^{(k+1)}$ using the stochastic gradient Langevin dynamics (SGLD) algorithm:

$$\boldsymbol{Z}_n^{(k+1)} = \boldsymbol{Z}_n^{(k)} + \epsilon_{k+1}\widehat{\nabla}_{\boldsymbol{z}_n} \log \pi(\boldsymbol{Z}_n^{(k)}|\boldsymbol{X}_n, \boldsymbol{Y}_n, \boldsymbol{w}^{(k)}) + \sqrt{2\tau\epsilon_{k+1}}\boldsymbol{e}^{(k+1)}, \tag{18}$$

where $\boldsymbol{e}^{(k+1)} \sim N(\boldsymbol{0}, I_{d_{\boldsymbol{z}}})$, ϵ_{k+1} is the learning rate, and τ is the temperature that is generally set to 1 in simulations.

(b) (*Parameter updating*) Update the estimate of $\boldsymbol{w}$ by SGD:

$$\boldsymbol{w}^{(k+1)} = \boldsymbol{w}^{(k)} + \frac{\gamma_{k+1}}{n}\widehat{\nabla}_{\boldsymbol{w}} \log \pi(\boldsymbol{w}^{(k)}|\boldsymbol{X}_n, \boldsymbol{Y}_n, \boldsymbol{Z}_n^{(k+1)}), \tag{19}$$

where γ_{k+1} denotes the step size of stochastic approximation.

Extended Fiducial Inference: Convergence of the Adaptive MCMC Algorithm

Theorem 6

(Informal version) If we set $\epsilon_k = \frac{C_\epsilon}{c_\epsilon + k^\alpha}$ *and* $\gamma_k = \frac{C_\gamma}{c_\gamma + k^\beta}$ *for some constants* $C_\epsilon > 0$, $c_\epsilon > 0$, $C_\gamma > 0$ *and* $c_\gamma > 0$, $\alpha, \beta \in (0, 1]$, *and* $\beta \leq \alpha \leq \min\{1, 2\beta\}$, *then there exists a root* $\boldsymbol{w}_n^* \in \{\boldsymbol{w} : \nabla_{\boldsymbol{w}} \log \pi(\boldsymbol{w}|\boldsymbol{X}_n, \boldsymbol{Y}_n) = 0\}$ *such that*

$$\mathbb{E}\|\boldsymbol{w}_n^{(k)} - \boldsymbol{w}_n^*\|^2 \leq \xi \gamma_k, \quad k \geq k_0,$$

for some constant $\xi > 0$ *and iteration number* $k_0 > 0$.

Extended Fiducial Inference: Convergence of the Adaptive MCMC Algorithm

Theorem 7

(Informal version) Under the conditions of Theorem 6, for any $k \in \mathbb{N}$,

$$\mathbb{W}_2(\mu_{T_k}, \pi^*) \leq (\hat{C}_0 \delta_g^{1/4} + \tilde{C}_1 \gamma_1^{1/4}) T_k + \hat{C}_2 e^{-T_k/c_{LS}},$$

for some positive constants $\hat{C}_0$, $\hat{C}_1$, *and* $\hat{C}_2$, *where* $\mathbb{W}_2(\cdot, \cdot)$ *denotes the 2-Wasserstein distance,* c_{LS} *denotes the logarithmic Sobolev constant of* π^*, *and* δ_g *reflects the variation of the stochastic gradient* $\widehat{\nabla}_{\boldsymbol{Z}_n} \log \pi(\boldsymbol{Z}_n^{(k)} | \boldsymbol{X}_n, \boldsymbol{Y}_n, \boldsymbol{w}^{(k)})$.

Extended Fiducial Inference: Convergence of the Adaptive MCMC Algorithm

Theorem 8

(informal version) Under the limit $\lambda \to \infty$, the posterior consistency holds for $\pi(\boldsymbol{w}_n | \boldsymbol{Y}_n, \boldsymbol{X}_n, \boldsymbol{Z}_n)$ and the inverse mapping estimator $\hat{g}(\cdot)$ constitutes a consistent estimator for the model parameters, i.e.,

$$\|\hat{g}(y, x, z, \boldsymbol{w}_n^*) - \boldsymbol{\theta}^*\| \xrightarrow{p} 0, \quad \text{as } n \to \infty,$$

where $\boldsymbol{\theta}^*$ denotes the fixed unknown parameter values, and (y, x, z) denotes a generic element of $(\boldsymbol{Y}_n, \boldsymbol{X}_n, \boldsymbol{Z}_n)$.

Extended Fiducial Inference: Convergence of the Adaptive MCMC Algorithm

We can approximate the fiducial distribution of $\boldsymbol{\theta}$ by

$$\hat{\mu}_n(d\boldsymbol{\theta}) = \frac{1}{\mathcal{M}} \sum_{k=1}^{\mathcal{M}} \delta_{\bar{\boldsymbol{\theta}}_n^k}(d\boldsymbol{\theta}), \quad \text{as } \mathcal{M} \to \infty, \tag{20}$$

where $\bar{\boldsymbol{\theta}}_n^k := \frac{1}{n} \sum_{i=1}^{n} \hat{g}(x_i, y_i, z_i^k, \boldsymbol{w}_n^{(k)})$, $\boldsymbol{Z}_n^{(k)} := (z_1^k, z_2^k, \ldots, z_n^k)$, and $(\boldsymbol{Z}_n^{(k)}, \boldsymbol{w}_n^{(k)})$ denotes the sample and parameter estimate produced by the adaptive MCMC algorithm at iteration k.

Extended Fiducial Inference: Linear regression

Considering a linear regression model given by

$$y_i = x_i^T \boldsymbol{\theta} + \sigma z_i, \quad i = 1, 2, \ldots, n, \tag{21}$$

where $z_i \sim N(0,1)$, $x_i = (x_{i,0}, \ldots, x_{i,9})^T$, $x_{i,0} = 1$, $x_{i,k} \sim N(0,1)$ for $k = 1, \ldots, 9$, $\sigma = 1$, and the regression coefficient $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_9)^T = (1,1,1,1,1,0,0,0,0,0)^T$. We simulated 100 datasets from this model, each with a sample size of $n = 500$.

Extended Fiducial Inference: Linear regression

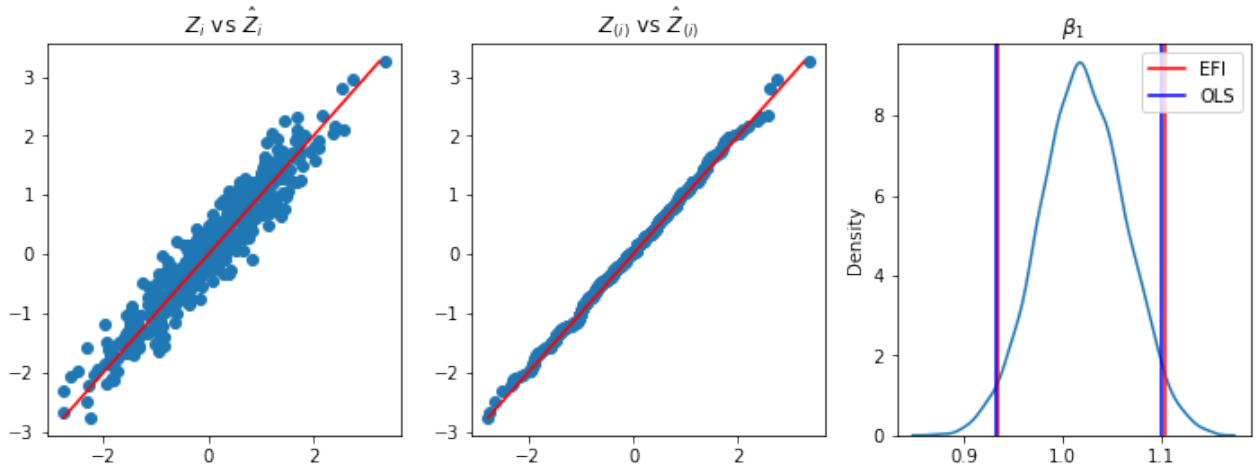

Figure 4: Results of EFI (with the ReLU activation function) for one dataset simulated from (21) with $n = 500$: (left) scatter plot of $\hat{\boldsymbol{z}}_n$ (y-axis) versus $\boldsymbol{z}_n$ (x-axis), (middle) Q-Q plot of $\hat{\boldsymbol{z}}_n$ and $\boldsymbol{z}_n$, (right) confidence intervals of β_1 produced by EFI and OLS.

Extended Fiducial Inference: Linear regression

Table 1: Statistical inference results for the model (21) with known σ^2, where "Coverage" refers to the averaged coverage rate over 100 datasets and respective parameters, and "CI-width" refers to the average width of respective confidence intervals.

		Signal parameters		Noise parameters	
Method	Activation	Coverage rate	CI-width	Coverage rate	CI-width
OLS	—	0.95	0.177	0.956	0.177
GFI	—	0.95	0.177	0.952	0.177
EFI-a	ReLU	0.948	0.176	0.95	0.171
EFI	Sigmoid	0.948	0.176	0.956	0.176
EFI	Tanh	0.948	0.176	0.956	0.176
EFI	Softplus	0.95	0.177	0.95	0.176
EFI	ReLU	0.95	0.176	0.95	0.176

Extended Fiducial Inference: Behrens-Fisher Problem

Consider two Gaussian distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$. Suppose that two independent random samples of sizes n_1 and n_2 are drawn from them, respectively. The structural equations are given by

$$\begin{aligned} y_{1i} &= \mu_1 + \sigma_1 z_{1i}, \quad i = 1, \ldots, n_1, \\ y_{2i} &= \mu_2 + \sigma_2 z_{2i}, \quad i = 1, \ldots, n_2, \end{aligned} \tag{22}$$

where $z_{i1}, z_{i2} \sim N(0, 1)$ independently. The Behrens-Fisher problem pertains to the inference for the difference $\mu_1 - \mu_2$ when the ratio σ_1/σ_2 is unknown.

Extended Fiducial Inference: Behrens-Fisher Problem

Table 2: Statistical inference results for the Behrens-Fisher problem, where "Coverage" refers to the coverage rate of $\mu_1 - \mu_2$ calculated by averaging over 200 datasets, and "CI-width" refers to the average width of respective confidence intervals.

Method	$(\sigma_1^2, \sigma_2^2) = (0.25, 1)$			$(\sigma_1^2, \sigma_2^2) = (1, 1)$		
	Coverage	CI-width	std CI	Coverage	CI-width	std CI
$n_1 = n_2 = 50$						
Behrens-Fisher	0.95	0.634	0.0040	0.955	0.802	0.0043
Welch	0.95	0.630	0.0040	0.95	0.794	0.0042
Hsu-Scheffé	0.95	0.635	0.0040	0.955	0.804	0.0043
EFI	0.95	0.609	0.0058	0.955	0.788	0.0047
$n_1 = n_2 = 500$						
Behrens-Fisher	0.95	0.196	0.0004	0.95	0.248	0.0004
Welch	0.95	0.196	0.0004	0.95	0.247	0.0004
Hsu-Scheffé	0.95	0.196	0.0004	0.95	0.248	0.0004
EFI	0.95	0.198	0.0006	0.95	0.245	0.0014

EFI: Fidelity in Parameter Estimation

Consider the linear regression again. We set $n = 600$ and generated random errors from a mixture Gaussian distributions: $z_1, z_2, \ldots, z_{540} \sim N(0, 1)$ and $z_{541}, z_{542}, \ldots, z_{600} \sim N(4, 1)$. The latter cases were considered as outliers, although some of them might be indistinguishable from the former ones.

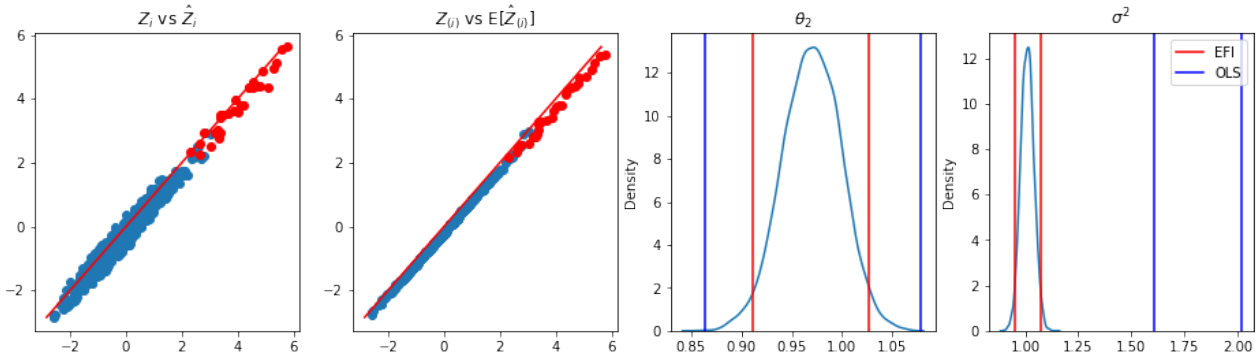

Figure 5: Fidelity of EFI in parameter estimation: (left) scatter plot of residuals: z_i versus $\hat{z}_i$; (middle left) scatter plot of ordered residuals: $z_{(i)}$ versus $\hat{z}_{(i)}$; (middle right) EFI and OLS confidence intervals for β_1; (right) EFI and OLS confidence intervals for σ^2.

EFI: Fidelity in Parameter Estimation

EFI essentially estimates $\boldsymbol{\theta}$ by maximizing the predictive likelihood:

$$\pi(\boldsymbol{Z}_n|\boldsymbol{X}_n, \boldsymbol{Y}_n, \boldsymbol{\theta}) \propto \pi_0^{\otimes n}(\boldsymbol{Z}_n) e^{-\lambda \sum_{i=1}^{n} d(y_i, x_i, z_i, \boldsymbol{\theta})},$$

which balances the fitting errors and the likelihood of random errors.

Compared to the MLE:

$$\hat{\boldsymbol{\theta}}_{MLE} = \arg\max_{\boldsymbol{\theta}} \pi_0^{\otimes n}(\boldsymbol{Z}_n),$$

where $\boldsymbol{Z}_n$ can be expressed as a function of $(\boldsymbol{Y}_n, \boldsymbol{X}_n, \boldsymbol{\theta})$, the EFI estimator tends to be more robust to outliers and provides higher fidelity in parameter estimation.

If the model is correctly specified, no outliers exist, and the sample size is reasonably large, the two methods yield similar estimates.

EFI for Semi-Supervised Learning

Table 3: Comparison of EFI with supervised learning and semi-supervised learning (50% labels were removed) algorithms for some classification problems, where $\mu \pm se$ represents the mean prediction accuracy of the 5-fold cross validation runs and the standard deviation of the mean value.

		Supervised Learning		Semi-Supervised Learning		
Dataset	size	Full	Labeled Only	Self-training	Label-propagation	EFI
Divorce	170	98.82±1.05	96.47±1.29	92.94±3.87	96.47±1.29	98.82±1.05
Diabetes	520	89.62±1.29	87.69±1.69	87.31±2.01	85.77±2.01	88.08± 0.64
Breast Cancer	699	96.52±0.66	95.36±0.26	94.39±0.76	95.07±0.52	96.23±0.52
Raisin	900	82.89±1.16	83.78±0.24	58.67±1.35	50.22±0.20	85.56± 0.99

EFI for Complex Hypothesis Tests

Consider the following mediation analysis model:

$$\begin{aligned} Y &= \beta_T T + \beta M + \boldsymbol{\beta}_x^T X + \epsilon_Y, \quad \epsilon_Y \sim N(0, \sigma_Y^2), \\ M &= \gamma T + \boldsymbol{\gamma}_x^T X + \epsilon_M, \quad \epsilon_M \sim N(0, \sigma_M^2), \end{aligned} \tag{23}$$

where Y, T, M and X denote the outcome, treatment, mediator and design matrix, respectively.

The mediator effect can be inferred by testing the hypothesis $H_0 : \beta\gamma = 0$ against $H_A : \beta\gamma \neq 0$ with the natural test statistic $\hat{\beta}\hat{\gamma}$.

EFI for Complex Hypothesis Tests

This is a challenging inferential task due to the non-uniform asymptotics of the univariate test statistic. Specifically, the null hypothesis consists of three cases:

(i) $\beta = 0, \gamma \neq 0$,

(ii) $\beta \neq 0, \gamma = 0$,

(iii) $\beta = \gamma = 0$,

while the theoretical reference distribution of $\hat{\beta}\hat{\gamma}$ under case (iii) is different from that under cases (i) and (ii). It is known that traditional statistical tests such as Sobel's test and Max-P test are conservative under case (iii).

EFI for Complex Hypothesis Tests

Table 4: Type-I errors of the Sobel, MaxP, minimax optimal (mm-opt), bootstrap, and EFI tests for the mediator effect, where the significance level of each test is $\alpha = 0.05$.

	$n = 500$			$n = 1000$			$n = 2000$		
(β, γ)	(0.2,0)	(0,0.2)	(0,0)	(0.2,0)	(0,0.2)	(0,0)	(0.2,0)	(0,0.2)	(0,0)
Sobel	0.01	0.00	0.00	0.05	0.02	0.00	0.04	0.06	0.00
MaxP	0.04	0.03	0.00	0.06	0.05	0.00	0.07	0.07	0.00
mm-opt	0.05	0.04	0.03	0.06	0.05	0.07	0.07	0.07	0.07
Bootstrap	0.06	0.05	0.01	0.04	0.07	0.00	0.13	0.04	0.00
EFI	0.05	0.06	0.04	0.06	0.04	0.04	0.05	0.04	0.05

EFI for Complex Hypothesis Tests

Table 5: Powers of the Sobel, MaxP, minimax optimal (mm-opt), bootstrap, and EFI tests for the mediator effect, where the significance level of each test is $\alpha = 0.05$. Part of the results of mm-opt are not available (NA), as the test is inefficient for the alternative hypothesis settings of (β, γ) when the sample size becomes large.

	$n = 500$			$n = 1000$			$n = 2000$		
(β, γ)	(0.1,0.4)	(-0.1,0.4)	(0.2,0.2)	(0.1,0.4)	(-0.1,0.4)	(0.2,0.2)	(0.1,0.4)	(-0.1,0.4)	(0.2,0.2)
Sobel	0.29	0.31	0.67	0.65	0.57	0.96	0.78	0.89	**1.00**
MaxP	0.34	0.37	0.79	0.66	0.59	**0.98**	0.78	0.89	**1.00**
mm-opt	0.34	0.37	0.79	NA	NA	NA	NA	NA	NA
Bootstrap	0.33	0.42	0.52	0.59	0.51	0.93	**0.93**	0.92	**1.00**
EFI	**0.48**	**0.64**	**0.84**	**0.70**	**0.74**	0.97	0.86	**0.95**	**1.00**

Conclusion

- EFI is a novel and flexible framework for statistical inference, applicable to general statistical models regardless of the type of noise, whether additive or non-additive.
- The EFI-DNN algorithm provides an effective implementation for EFI. It jointly imputes the realized random errors in observations using MCMC and approximates the inverse function using a sparse DNN. The consistency of the sparse DNN estimator ensures the uncertainty embedded in the observations is properly propagated to the model parameters through the estimated inverse function, thereby validating downstream statistical inference.
- The EFI-DNN algorithm has demonstrated appealing properties in parameter estimation (fidelity), hypothesis testing (automated process), and semi-supervised learning (missing data).