Extended Fiducial Inference: Toward Automated Statistical Inference

Faming Liang

Purdue University

Background of the Research

Automating statistical inference is a central goal in the era of AI.

However, under the current frequentist and Bayesian frameworks, statistical inference still requires substantial expert intervention. Fiducial inference, as a method for making inferences about unknown parameters based solely on observed data, offers a promising pathway toward achieving this goal.

Outline

- Basic Theory of Extended Fiducial Inference (EFI)
- An Illustrative Example
- Expanding Boundaries of Statistical Inference
  - Complex hypothesis testing (test statistics having unknown distributions)
  - Large models: $d_\theta = O(n^\alpha)$ with $\frac{1}{2} < \alpha < 1$, including deep neural networks as special cases
  - Solving PDEs with Uncertainty Quantification (PINN)
  - Semi-supervised learning
- Conclusion

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 its fidelity.
- For high-dimensional problems, posterior consistency needs to be theoretically justified.

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, 1930), has continued to inspire statisticians, leading to the development of methods such as

- Dempster–Shafer Theory (Dempster, 1967; Shafer, 1976)
- Structural Inference (Fraser, 1968)
- Generalized Fiducial Inference (Hannig, 2009)
- Inferential Models (Martin and Liu, 2013)

Statistical Inference: Fiducial

Building on our early work 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.

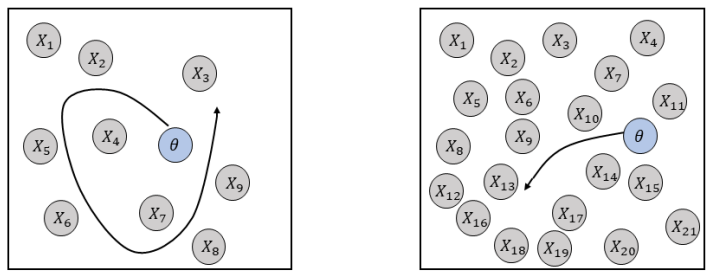


Figure 1: Illustration for the source of uncertainty of model parameters.

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: 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

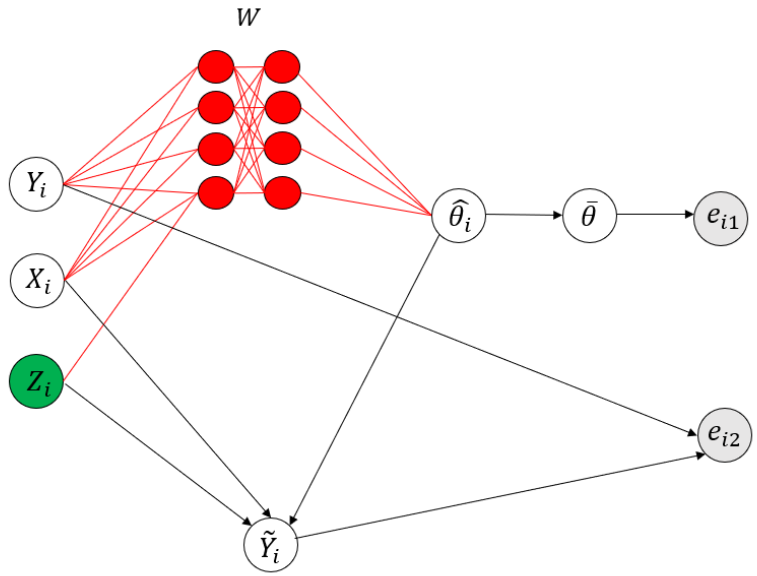


Figure 2: Illustration of the EFI network, where the red nodes and links form a DNN (parameterized by the weights $\boldsymbol{w}$) to learn, the green node represents latent variables to impute, and the black lines represent deterministic functions.

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:* $U_n(\boldsymbol{z}) \geq 0$, *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

Theorem 2

If $\Pi_n(\mathcal{Z}_n) > 0$, 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 regression, for which the resulting estimator of $\boldsymbol{\theta}$ is a set.

Extended Fiducial Inference: Theory-5

Theorem 3

If If $\Pi_n(\mathcal{Z}_n) = 0$ *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 zero-energy set $\mathcal{Z}_n$, 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: 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 $\mu_n^*(b(\boldsymbol{\theta})|\boldsymbol{Y}_n, \boldsymbol{X}_n)$ 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 eliminates the requirement for the theoretical distributions of the estimators.

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

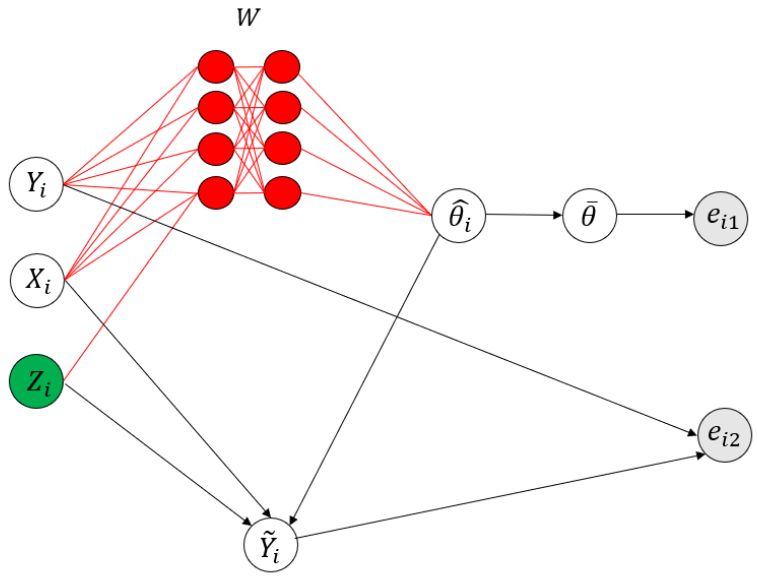


Figure 3: Illustration of the EFI network

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:

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

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.

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

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 5

(Informal version) If $\epsilon_k = O(k^{-\alpha})$ *and* $\gamma_k = O(k^{-\beta})$, $\alpha, \beta \in (0, 1]$, *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 6

(Informal version) Let μ_k denote the density function of $\boldsymbol{Z}_n^{(k)}$ drawn at iteration k. Then

$$\mathbb{W}_2(\mu_k, \pi^*) \to 0, \quad \text{as } k \to \infty,$$

where $\mathbb{W}_2(\cdot, \cdot)$ denotes the 2-Wasserstein distance.

Extended Fiducial Inference: Convergence of the Adaptive MCMC Algorithm

Theorem 7

(informal version) As $\epsilon \to 0$ and $n \to \infty$, the inverse mapping is consistent:

$$\| G_n^*(\boldsymbol{Y}_n, \boldsymbol{X}_n, \boldsymbol{Z}_n) - \boldsymbol{\theta}^* \| \xrightarrow{p} 0,$$

where $G_n^*(\boldsymbol{Y}_n, \boldsymbol{X}_n, \boldsymbol{Z}_n) = \frac{1}{n} \sum_{i=1}^{n} \hat{g}(y_i, x_i, z_i, \boldsymbol{w}_n^*)$, and $\hat{g}(\cdot)$ is the neural network function.

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

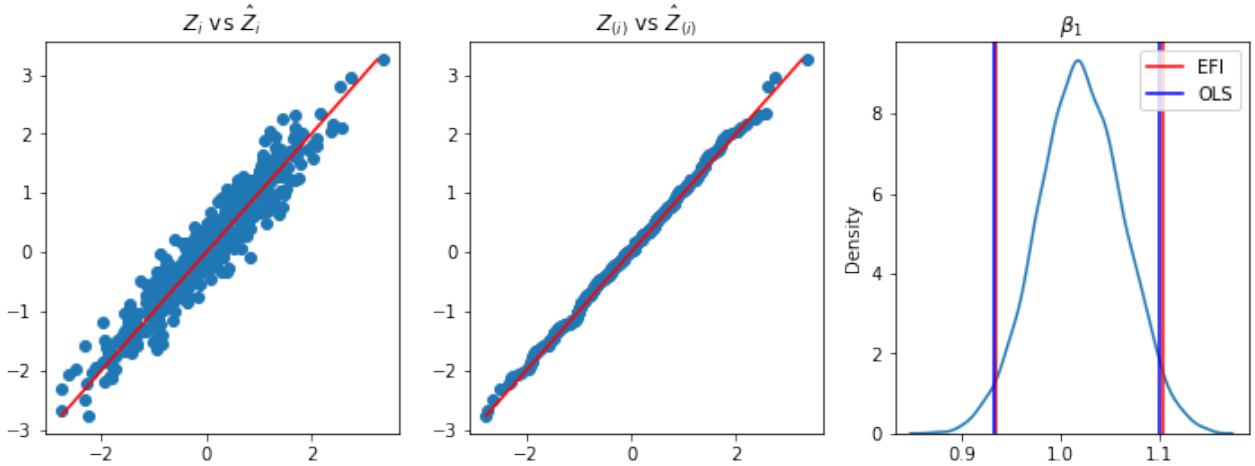


Figure 4: Results of EFI (with the ReLU activation function) for one dataset simulated from (21) with $n = 500$: (left) scatter plot of $\hat{\boldsymbol{z}}_n$ (y-axis) versus $\boldsymbol{z}_n$ (x-axis), (middle) Q-Q plot of $\hat{\boldsymbol{z}}_n$ and $\boldsymbol{z}_n$, (right) confidence intervals 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
EFI	Softplus	0.95	0.177	0.95	0.176
EFI	ReLU	0.95	0.176	0.95	0.176

EFI for Complex Hypothesis Tests

Consider the following mediation analysis model:

$$\begin{aligned} Y &= \beta_T T + \beta M + \boldsymbol{\beta}_x^T X + \epsilon_Y, \quad \epsilon_Y \sim N(0, \sigma_Y^2), \\ M &= \gamma T + \boldsymbol{\gamma}_x^T X + \epsilon_M, \quad \epsilon_M \sim N(0, \sigma_M^2), \end{aligned} \tag{22}$$

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 2: 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 3: 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**

EFI for Large Models

Theorem 8

(Kim and Liang, 2025) Suppose that ϵ is sufficiently small, and

$$\sum_{l=1}^{H} d_l \prec n, \tag{23}$$

where d_l denotes the width of layer l, $d_H = dim(\boldsymbol{\theta})$, and H denotes the depth of the $\boldsymbol{w}$-network. Then $G^*(\boldsymbol{Y}_n, \boldsymbol{X}_n, \boldsymbol{Z}_n) = \frac{1}{n}\sum_{i=1}^{n} \hat{g}(y_i, \boldsymbol{x}_i, z_i, \boldsymbol{w}_n^*)$ constitutes a consistent estimator of the inverse function.

EFI for Large Models: Individual Treatment Effect

Consider a causal inference problem:

$$y_i = \mu + \boldsymbol{x}_i^\top \boldsymbol{\beta} + (\eta_0 + \eta(\boldsymbol{x}_i))T_i + \sigma z_i, \quad i = 1, 2, \ldots, n, \tag{24}$$

where $\boldsymbol{x}_i = (x_{i,1}, x_{i,2})^\top$ with each element drawn independently from $Unif(0,1)$, $\mu = 1$, $\boldsymbol{\beta} = (1,1)^\top$, $\eta_0 = 1$, $\sigma = 1$, $z_i \sim N(0,1)$,

$$\eta(\boldsymbol{x}_i) = s(x_{i1})s(x_{i2}) - E(s(x_{i1})s(x_{i2})).$$

where $s(a) = \frac{2}{1+exp(-12(a-0.5))}$, and generate the treatment variable T_i according to the propensity score model:

$$e(\boldsymbol{x}_i) = \frac{1}{4}(1 + \beta_{2,4}(x_{i,1})), \tag{25}$$

where $\beta_{2,4}$ is the CDF of the beta distribution with parameters (2,4), ensuring $e(\boldsymbol{x}_i) \in [0.25, 0.5]$ and thereby sufficient overlap between the treatment and control groups.

EFI for Large Models: Individual Treatment Effect

- We generated 20 datasets from the model (24) independently, each consisting of $n_{train} = 500$ training samples and $n_{test} = 1000$ test samples.
- We modeled $\tau(\boldsymbol{x})$ by a DNN. The DNN has two hidden layers, each consisting of 10 hidden neurons. The number of parameters of the DNN is $|\boldsymbol{\theta}_\tau| = 151$, and the total dimension of $\boldsymbol{\theta} = (\mu, \eta_0, \boldsymbol{\beta}, \boldsymbol{\theta}_\tau^\top, \log(\sigma))^\top$ is 156 ($\approx n_{train}^{0.81}$), which falls into the class of large models.
- In Kim and Liang (2025), more extreme examples have been considered.

EFI for Large Models: Individual Treatment Effect

Table 4: Comparison of Double-NN and CQR for inference of the predictive ITE for Example (24), where the coverage and length of the prediction intervals were calculated by averaging over 20 datasets with the standard deviation given in the parentheses.

Method	Case $\mathcal{I}_c$		Case $\mathcal{I}_t$		Method	Case $\mathcal{I}_m$	
	Coverage	Length	Coverage	Length		Coverage	Length
Double-NN	0.9549 (0.0095)	4.2004 (0.1567)	0.9581 (0.0098)	4.1812 (0.1541)	Double-NN	0.9583 (0.0103)	5.6056 (0.2207)
CQR-BART	0.9472 (0.0342)	4.2702 (0.5225)	0.9533 (0.0341)	4.4024 (0.8972)	CQR(inexact)	0.9530 (0.0198)	6.3244 (0.5426)
CQR-Boosting	0.9556 (0.0294)	5.5199 (0.5866)	0.9548 (0.0259)	4.4493 (0.5097)	CQR(exact)	1.0000 (0.0002)	13.4005 (2.4936)
CQR-RF	0.9529 (0.0233)	5.4609 (0.5172)	0.9652 (0.0171)	4.6428 (0.5408)	CQR(naive)	0.9998 (0.0004)	12.8861 (1.5275)
CQR-NN	0.9570 (0.0195)	6.4072 (0.8087)	0.9755 (0.0199)	5.8125 (1.4332)			

Solving PDEs with Uncertainty Quantification

Consider a one-dimensional Poisson equation:

$$\beta \frac{\partial^2 u}{\partial x^2} = f, \quad x \in \Omega, \tag{26}$$

where $\Omega = [-0.7, 0.7]$, $\beta = 0.01$, $u = \sin^3(6x)$, and f can be derived from (26).

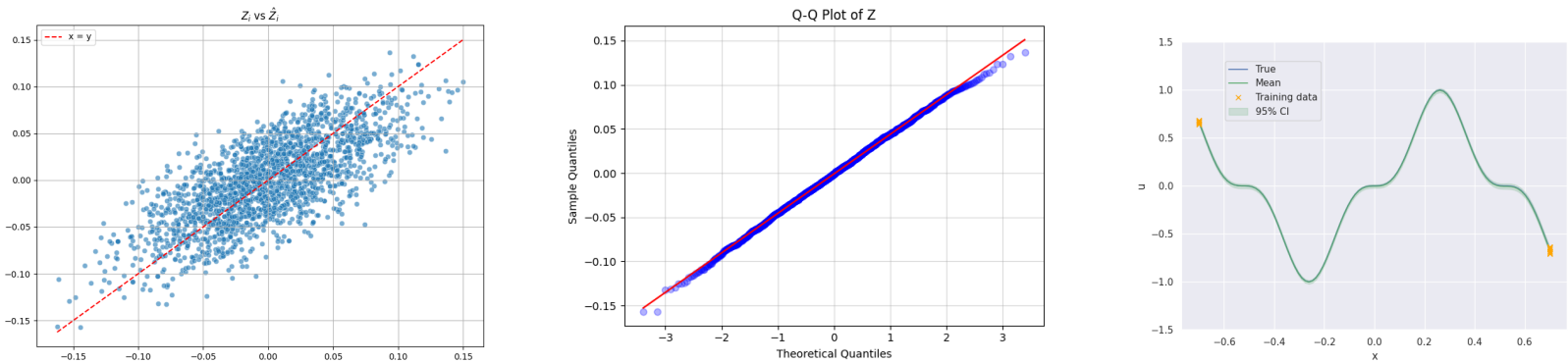


Figure 5: EFI-PINN diagnostic for 1D-Poisson

Solving PDEs with Uncertainty Quantification

We consider solving a nonlinear Poisson with an unknown parameter:

$$\beta \frac{\partial^2 u}{\partial x^2} + k \tanh(u) = f, \quad x \in \Omega, \tag{27}$$

where $\Omega = [-0.7, 0.7]$, $\beta = 0.01$, $k = 0.7$, $u = \sin^3(6x)$.

Table 5: Metrics for nonlinear 1D-Poisson with parameter estimation, averaged over 100 runs.

Method	hidden layers	MSE($\times 10^{-4}$)	Coverage Rate	CI-Width	k-Mean	k-Coverage Rate	k-CI-Width
PINN	[50, 50]	1.79 (0.13)	0.1464 (0.0097)	0.0045 (0.0001)	0.6998 (0.0006)	0.00 (0.0000)	7.9e-5 (5e-6)
Dropout (0.5%)	[50, 50]	1.84 (0.10)	1.0000 (0.0000)	0.2393 (0.0035)	0.6912 (0.0007)	0.09 (0.0288)	0.0045 (0.0002)
Dropout (1.0%)	[50, 50]	2.72 (0.24)	1.0000 (0.0000)	0.3032 (0.0071)	0.6874 (0.0008)	0.11 (0.0314)	0.0091 (0.0007)
Dropout (5.0%)	[50, 50]	27.08 (2.12)	1.0000 (0.0000)	0.5363 (0.0131)	0.6493 (0.0023)	0.06 (0.0239)	0.0252 (0.0014)
Bayesian	[50, 50]	1.31 (0.08)	0.9752 (0.0055)	0.0517 (3e-5)	0.6994 (0.0005)	1.00 (0.0000)	0.0411 (0.0002)
EFI	**[50, 50]**	**1.03 (0.08)**	**0.9473 (0.0099)**	**0.0396 (0.0002)**	**0.6985 (0.0004)**	**0.94 (0.0239)**	**0.0179 (0.0002)**
EFI	**[100, 100]**	**0.98 (0.07)**	**0.9560 (0.0099)**	**0.0395 (0.0002)**	**0.6995 (0.0004)**	**0.96 (0.0197)**	**0.0168 (0.0002)**

EFI for Semi-Supervised Learning

Table 6: 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

Conclusion

- We have developed extended fiducial inference (EFI) as a novel and flexible framework for statistical inference, applicable to general statistical models regardless of the noise structure, whether additive or non-additive.
- We have introduced the EFI-DNN algorithm for effective implementation of EFI, which jointly imputes the realized random errors in observations using MCMC and approximates the inverse mapping through deep neural networks.
- EFI exhibits strong theoretical and practical properties. It significantly expands the boundaries of statistical inference and holds great potential for advancing the automation of statistical inference.

References

1. Liang, F., Kim, S., & Sun, Y. (2025) Extended fiducial inference: toward an automated process of statistical inference. *Journal of the Royal Statistical Society Series B*, 87, 98-131.
2. Kim, S. & Liang, F. (2025). Extended Fiducial Inference for Individual Treatment Effects via deep neural networks. *Statistics and Computing*, **35**:97.
3. Shi, F., Jiang, Z., & Liang, F. (2025). Uncertainty quantification for physics-informed neural networks with extended fiducial inference. NeurIPS 2025.