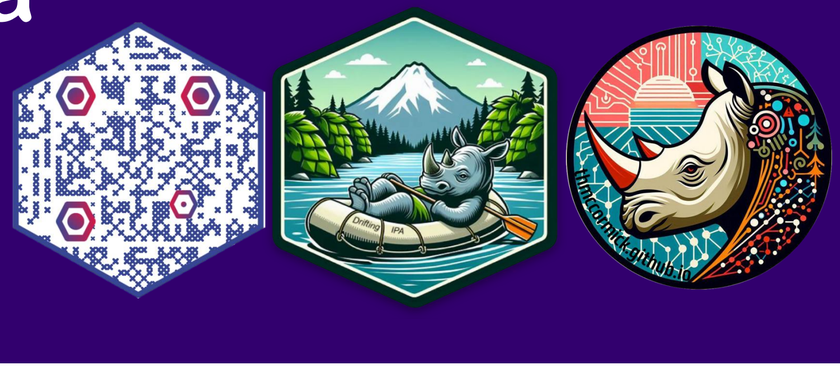
UNIVERSITY of WASHINGTON

Bayesian Optimal Experimental Design of Streaming Data Incorporating Machine Learning-Generated Synthetic Data



Kentaro Hoffman[1], Tyler McCormick[1]

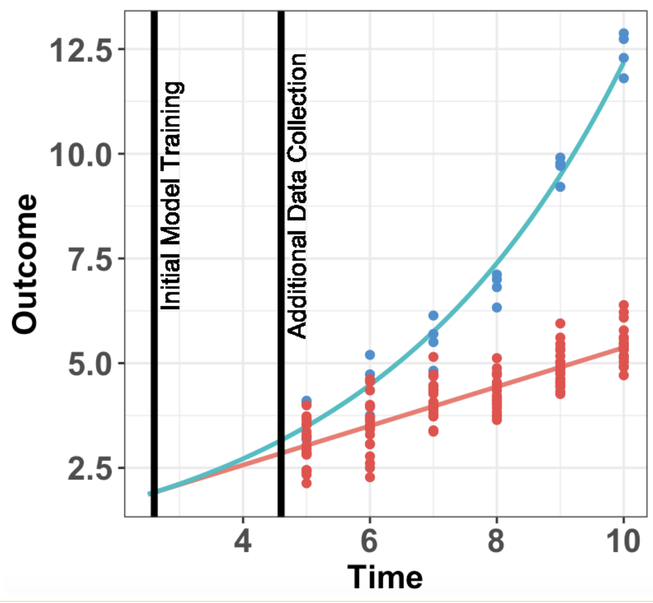
[1] University of Washington Department of Statistics

How do we improve our experimental design when we can get samples from a black box prediction model?

AI models can be used to generate **hard-to-measure** outcomes (health events) but due to high training costs, they can't always be kept **up to date**

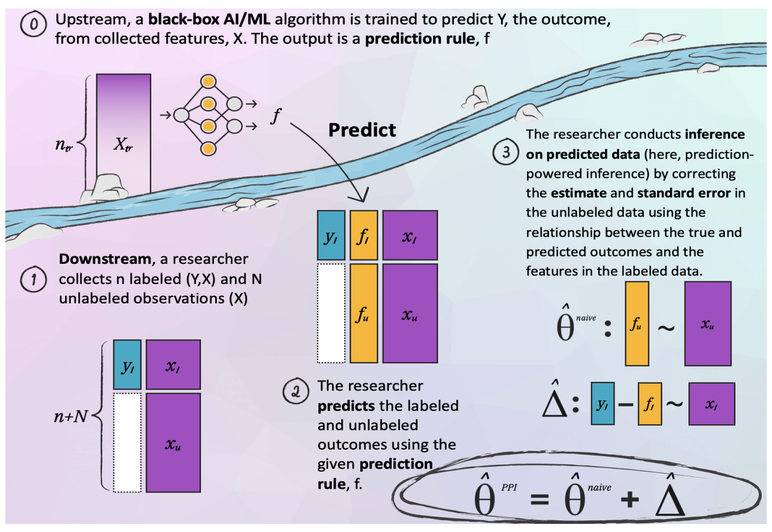
Goal: Use Bayesian Optimal Experimental design to determine the ratio of real-to-synthetic data for a dynamic linear model using AI-imputed data

Cheap or Expensive Data?



Data Type	Properties
Real Data	• Accurate • Expensive
Synthetic Data	• Cheaper • Biased
Inference on Predicted Data	• Gets valid inference • Improved Data Efficiency

Inference on Predicted Data (IPD)



Inference on predicted data **(IPD)** is a problem where one is interested in doing statistical inference when part of the data is **partially black-box imputed**

One procedure for IPD is **Prediction-Powered Inference** which estimates $\widehat{\boldsymbol{\beta}}$ from:

$$\widehat{\boldsymbol{\beta}}^{\boldsymbol{PPI}} = \boldsymbol{argmin}_{\boldsymbol{\beta}}\, \boldsymbol{L_n(\beta)} + (\widetilde{\boldsymbol{L}}_{\boldsymbol{n}}(\boldsymbol{\beta}) - \boldsymbol{L}_{\boldsymbol{n}}^{\boldsymbol{f}}(\boldsymbol{\beta}))$$

$[\boldsymbol{L_n(\beta)}, \widetilde{\boldsymbol{L}}_{\boldsymbol{n}}(\boldsymbol{\beta}), \boldsymbol{L}_{\boldsymbol{n}}^{\boldsymbol{f}}(\boldsymbol{\beta})]$ are the loss functions using the black-box predicted values, the experimentally observed values, and the black-box predictions of the experimental data.

In the case of linear models, this simplifies to:

$$\hat{\beta}^{naive} = argmin_{\beta} \left|\left|Y_{syn} - X_{syn}\,\beta\right|\right|$$
$$\hat{\Delta} = argmin_{\beta} ||(f(X_{real}) - Y_{real}) - X_{real}\beta||$$
$$\widehat{\beta^{PPI}} = \hat{\beta}^{naive} + \hat{\Delta}$$

This estimator has multiple useful properties:

- It is **unbiased** for the true $\boldsymbol{\beta}$
- $\widehat{\boldsymbol{\Delta}}$ describes how biased **your prediction rule is from $\boldsymbol{Y}$ when it comes to estimating $\boldsymbol{\beta}$**
- $\widehat{\boldsymbol{\beta}}^{\boldsymbol{naive}}$ and $\widehat{\boldsymbol{\Delta}}$ are estimated on separate datasets so the width of the CI of $\widehat{\boldsymbol{\beta}}^{\boldsymbol{naive}}$ is the sum of the CI of $\widehat{\boldsymbol{\beta}}^{\boldsymbol{naive}}$ and $\widehat{\boldsymbol{\Delta}}$
- Both are parametric estimators

Estimation the standard errors under model misspecification is done via **sandwich variance estimator**

Bayesian Sandwich estimation

Finding a Bayesian analog of the sandwich variance estimator is still an open problem

$$y(t) \sim N(\phi(x,t), \sigma^2(x,t))$$

$$\phi(x,t) = \beta(t)x + \Phi^T B_\phi(x)$$

$$log(\sigma(x,t)) = \gamma(t)x + \Psi^T B_\sigma(x).$$

One practical approach developed [1] is to use a model flexible model where the mean and variance use B-splines

Computationally fast and surprisingly accurate in real data problems

Dynamic Linear Model

We use a common approach to Dynamic Linear models and posit

$$y_t = X_t^T \beta_t + \epsilon_t$$

Where β_t evolves according to the state equation:

$$\beta_t = \beta_{t-1} + \delta_t$$

And $\delta_t \sim N(0, W_t)$.

This yields **estimation equation:**

$$\beta_{t-1}|Y^{t-1} \sim N(\hat{\beta}_{t-1}, \hat{\Sigma}_{t-1})$$

And **prediction equation:**

$$\beta_t|\, Y^{t-1} \sim N(\hat{\beta}_{t-1}, R_t)$$

Where**:**

$$R_t = \hat{\Sigma}_{t-1}/\lambda_t$$

With $\lambda_t < 1$ ensuring the model **"forgets"** past observations

When using an IPD estimator, this yields:

$$\beta_t|\hat{Y}^{t-1}, Y^{t-1} \sim N(\hat{\beta}_{t-1}^{PPI}, \hat{\Sigma}_{t-1}^{PPI})$$

$$\beta_t|\hat{Y}^{t-1}, Y^{t-1} \sim N(\hat{\beta}_{t-1}^{Naive} + \hat{\beta}_{t-1}^{\Delta}, R_{t-1}^{Naive} + R_{t-1}^{\Delta})$$

And:

$$R_{t-1}^{Naive} = \frac{\hat{\Sigma}_{t-1}^{Naive}}{\lambda^{Naive}}, R_{t-1}^{\Delta} = \frac{\hat{\Sigma}_{t-1}^{\Delta}}{\lambda^{\Delta}}$$

Bayesian Optimal Experimental Design

Let $\zeta_t \in (0,1)$ be the fraction of the budget spent on real and synthetic data. One can optimize for **Expected Information Gain:**

$$EIG(\zeta_t) = E_{\boldsymbol{Y}^t(\zeta_t),\beta}[H(p(\beta|\boldsymbol{Y^{t-1}}(\zeta_t))) - H(p(\beta|\boldsymbol{Y^t}(\zeta_t)))]$$
$$= E_{\boldsymbol{Y}^t(\zeta_t),\beta}[\log \frac{p(\boldsymbol{Y}^t|\beta)}{p(\boldsymbol{Y}^t)}].$$

Which [4] showed can be well approximated by the nested-Monte-Carlo approach:

$$EIG(\zeta_t) \approx \frac{1}{N}\sum_{i=1}^{N} \frac{p(\boldsymbol{Y}_n^t|\beta_{n,0})}{\frac{1}{M}\sum_{i=1}^{M} p(\boldsymbol{Y}_n^t|\beta_{n,m})}, \beta_{n,*} \sim p(\beta|\boldsymbol{Y}^{t-1}), \boldsymbol{Y}_n^t \sim p(\boldsymbol{Y}_n^t|\beta_{n,0})$$

Intuition: As we get further from the initial model training, EIG is maximized by increasing the ratio of real to synthetic data

Future Work

Trial Design for global mortality estimation

Globally, **1/3** of deaths are not given a reliable cause of death, and autopsies are often unavailable due to a **lack of local resources**. Alternative forms of data collection such as **verbal autopsies**, **phone surveys**, or **model-based imputation** exist but face a transportability bias issue. This procedure can help design more effective ways to combine on-site data collection with cheaper forms of data collection.

Bayesian Model Averaging over Rashomon sets to improve robustness

Here, we only consider the effectiveness of using one black-box imputation model, but different models can be good at different time periods (such as dynamic switching model), by combining Rashomon sets with Bayesian model averaging, we can do our inference over the set of **reasonably-accurate** models and improve the applicability

References

[1] Adam A. Szpiro. Kenneth M. Rice. Thomas Lumley. "Model-robust regression and a Bayesian "sandwich" estimator." Ann. Appl. Stat. 4 (4) 2099 - 2113, December 2010.

[2] Anastasios N. Angelopoulos *et al.* Prediction-powered inference.*Science* **382**,669-674(2023).

[3] Kentaro Hoffman, Stephen Salerno, Awan Afiaz, Jeffrey T. Leek, and Tyler H. McCormick. Do we really even need data?, 2024a. *ARXIV*

[4] Adam Foster, Martin Jankowiak, Eli Bingham, Paul Horsfall, Yee Whye Teh, Tom Rainforth, and Noah D. Goodman. Variational bayesian optimal experimental design. In Neural Information Processing Systems, 2019.