Prediction-Powered Inference: Evaluating Efficiency in Genomic Data Analysis

Dowoo Kim
Department of Mathematics and Statistics, McGill University

Directed Reading Project from the source cited in [1] in the "References" section

Introduction

Prediction-powered inference is a framework for integrating machine learning predictions into statistical inference while ensuring valid confidence intervals and hypothesis tests. This project explores its application to protein folding prediction, leveraging machine learning models like AlphaFold. By combining high-accuracy predictions with gold-standard experimental data, we demonstrate that prediction-powered inference can yield more precise statistical conclusions in biological and healthcare-related research.

Fig 1. Q8W3K0: A potential plant disease resistance protein. Mean pLDDT 82.24. generated by Alphafold from Google Deepmind [2]

Methodology

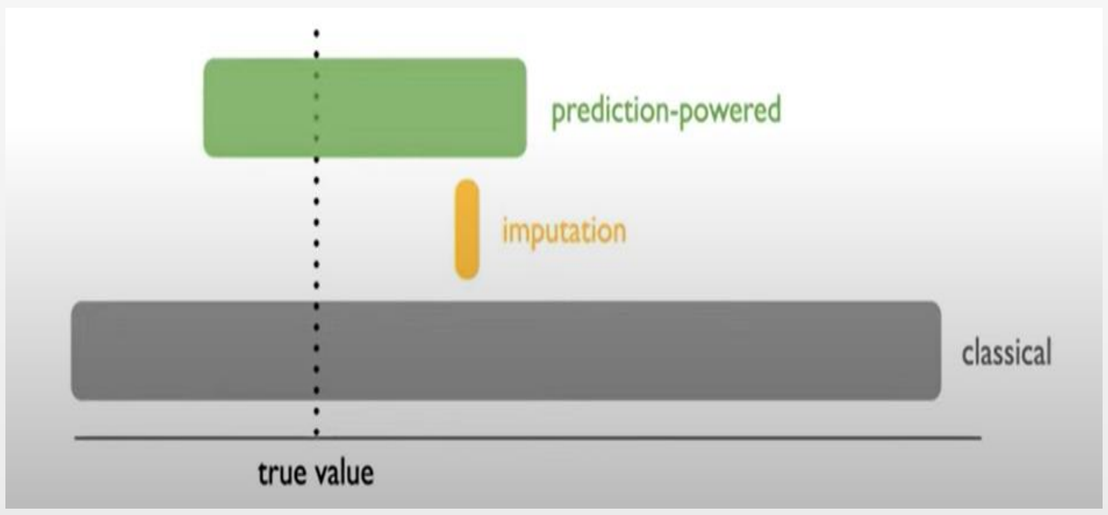

Fig. 2. Comparison of Confidence Intervals for the Odds Ratio of Structural Features.
This figure illustrates the confidence intervals for the odds ratio of structural features using three approaches: Classical, Imputation, and Prediction-Powered Inference. The dotted line represents the true odds ratio, computed from the gold-standard labeled data. [3]

Classical Approach ($\hat{\theta}_{class} = \frac{1}{n}\sum_{i=1}^{n} Y_i$):
Uses only labeled (experimental) data, leading to wider confidence intervals due to limited sample size and higher uncertainty.

Imputation Approach ($\hat{\theta}_f = \frac{1}{N}\sum_{i=1}^{N} f(X_i^e)$):
Uses only the predicted values from a machine learning model, ignoring bias in the predictions (as if they are true outcomes). This can lead to misleading or incorrect results if the model has systematic errors.

Prediction-Powered Inference:
Prediction-Powered inference combines the strengths of both methods by adjusting for systematic prediction errors.

1. Define the **rectifier** Δ to measure the prediction error: $\Delta = f(X) - Y$. This captures how much predictions deviate from true outcomes.

2. Construct a **confidence set** R for Δ, using labeled data to estimate the error range.

3. Adjust the implemented estimate using R:

$$C_{PPI} = \{\hat{\theta}_f + \Delta \mid \Delta \in R\}$$

This **corrects for prediction bias**, ensuring a valid confidence interval for θ^*.

Final PPI Estimator
By integrating prediction error correction, the **prediction-powered estimator** is:

$$\hat{\theta}_{PPI} = \frac{1}{N}\sum_{i=1}^{N} f(X_i^e) - \frac{1}{n}\sum_{i=1}^{n}(f(X_i^e) - Y_i)$$

- The **first term** leverages the large predicted dataset.
- The **second term** corrects for bias using labeled data.
- This estimator has **lower variance** than classical inference, leading to **more efficient estimation**.

Confidence intervals:
The **95% confidence interval** for θ^* is:

$$\hat{\theta}_{PPI} \pm 1.96\sqrt{\frac{\hat{\sigma}^2_{\hat{Y}_f}}{N} + \frac{\hat{\sigma}^2_Y}{n} + \frac{\hat{\sigma}^2_f}{N}}$$

This **reduces variance** compared to classical confidence intervals:

$$\hat{\theta}_{\text{class}} = 1.96\sqrt{\frac{\hat{\sigma}^2_Y}{n}}$$

The variance is reduced due to three key factors:
- **Large sample size for predictions** ($N \gg n$) lowers variance in estimation
- **Small correction from labeled data** ensures the adjustment term has low variance.
- **Combining both sources of information** results in a tighter and more efficient estimator.

Results

Random genotype data (10 SNPs per individual) are simulated and a continuous phenotype are generated via a linear model with added noise.

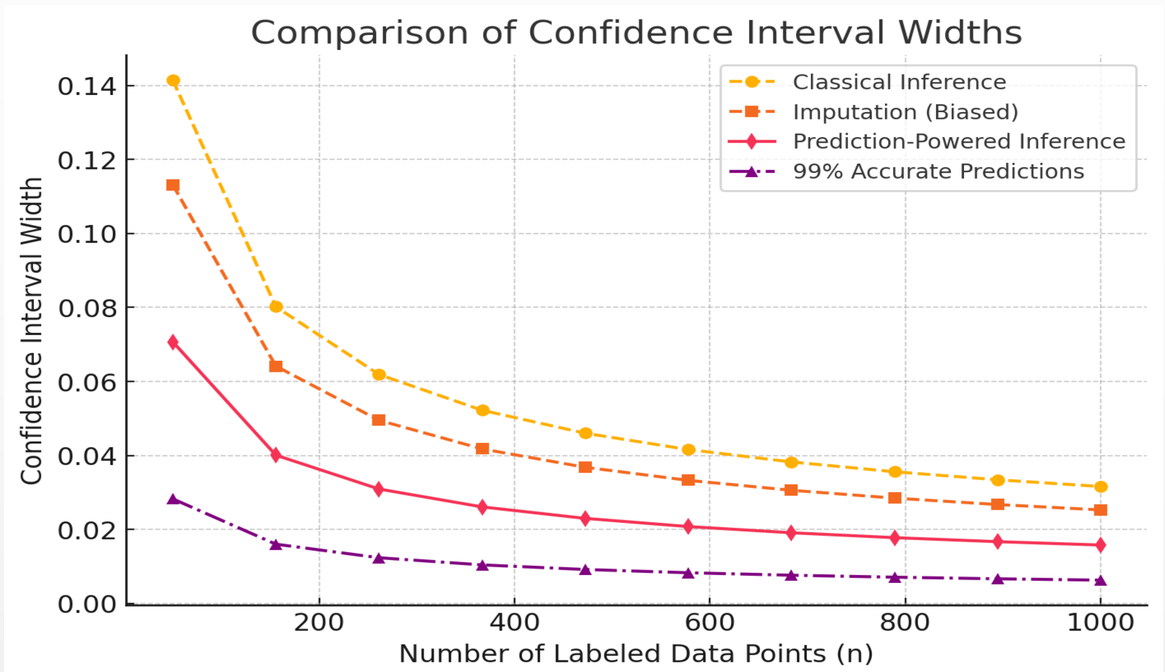

Fig. 2, showing how confidence interval widths vary for four inference methods, each plotted against the number of labeled data points n. The yellow dashed line represents the “Classical Inference” approach, which relies solely on limited labeled data and thus yields wide intervals. The orange dashed line shows “Imputation (Biased),” which underestimates uncertainty by treating predictions as exact labels. The green red line corresponds to “Prediction-Powered Inference,” combining a large unlabeled dataset with a small labeled set to correct bias and achieve narrower, valid intervals. Finally, the purple dash-dot line highlights a “99% Accurate Predictions” scenario, illustrating near-perfect predictions that drastically reduce interval widths, the confidence intervals become extremely tight. The figure demonstrates how leveraging unlabeled data effectively can significantly tighten intervals while maintaining statistical rigor.

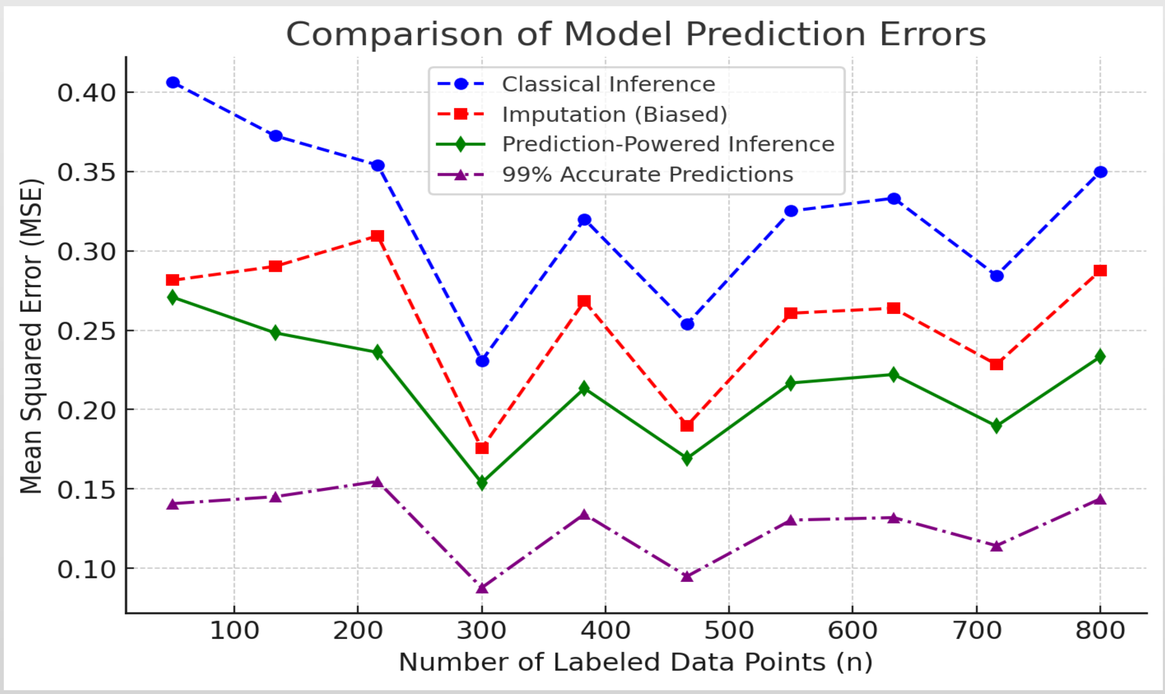

Fig. 3, displaying how the mean squared error (MSE) behaves across four inference methods, each plotted against the number of labeled data points n. The blue dashed line (“Classical Inference”) shows relatively high MSE due to reliance on limited labeled data. The orange dashed line (“Imputation (Biased)”) achieves lower MSE but is overconfident, as it assumes model predictions are perfect. The green solid line (“Prediction-Powered Inference”) outperforms classical inference by combining labeled and unlabeled data with a bias correction, thus lowering MSE further. Finally, the purple dash-dot line (“99% Accurate Predictions”) illustrates an almost ideal model, yielding the lowest MSE overall. This figure demonstrates how more sophisticated methods can exploit unlabeled data or high prediction accuracy to substantially reduce error.

Future Work

Apply prediction-powered inference to image data, such as MRI images and microscopic data, to evaluate the model's accuracy and identify scenarios where it is less efficient than alternative models and suggest more efficient inference to decrease the error of prediction.

References

[1] Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., & Zrnic, T. (2023). Prediction-Powered Inference. ArXiv. https://arxiv.org/abs/2301.09633
[2] European Bioinformatics Institute. (n.d.). AlphaFold protein structure prediction for Q8W3K0 [Image]. AlphaFold EBI. https://alphafold.ebi.ac.uk/assets/img/Q8W3K0.png
[3] [Clara Wong-Fannjiang]. (2023, May 30). Prediction-powered inference (Clara Wong-Fannjiang, MSAC 2023) [Video]. Youtube. https://www.youtube.com/watch?v=FW5I5xEYETY