# Deep Limit Model-free Prediction

Kejin Wu [1] Dimitris N. Politis [1,2]

[1]Department of Mathematics, University of California, San Diego [2]Halicioğlu Data Science Institute, University of California, San Diego

## Intuition

Exploring the relationship between a predictor $X$ and a response $Y$ is a fundamental problem in statistics and machine learning.

Classically, people assume there is a model that **may** explain the relationship between $X$ and $Y$:

$$\boxed{X} \xleftrightarrow{\sim f} \boxed{Y}$$

where $\sim$ means that the association between $X$ and $Y$ is not exactly described by $f$ or there is a measurement error. A famous quote says *"Essentially, all models are wrong, but some are useful."*

Given a new $X_f$, people care about the corresponding $Y_f$, e.g., generating figures or texts given some inputs, i.e.,

$$\boxed{X_f} \xrightarrow{\approx \hat{f}} \boxed{Y_f = ?}$$

where $\approx$ involves additional error from estimating $f$ by $\hat{f}$ compared to $\sim$.

**Goal:** Make predictions without restrictive model assumptions and capture the estimation variability meanwhile.

## Model-free Prediction Principle

Instead of assuming there is a model $f$ that connects $X$ and $Y$, the Model-free prediction principle proposed by Politis (2015) relies on four steps:

1. Find an invertible transformation function $H_n$ which transforms non-$i.i.d.$ samples $(Y_1, \ldots, Y_n)$ to $i.i.d.$ vector $(e_1, \ldots, e_n) \overset{i.i.d.}{\sim} F_e$ with possible explanatory variables $(X_1, \ldots, X_n)$;
2. Solve for $Y_n$ in terms of $\mathbf{Y}_{n-1} := (Y_1, \ldots, Y_{n-1})$, $X_n$ and $e_n$, i.e., $Y_n = h_n(\mathbf{Y}_{n-1}, X_n, e_n)$;
3. Determine the future response $Y_f := h_n(\mathbf{Y}_n, X_f, e_f)$, where $e_f \sim F_e$ is independent with $Y_f$, $X_f$ and $(e_1, \ldots, e_n)$;
4. Evaluate the whole distribution of $Y_f$ by Monte Carlo ($F_e$ is known) or Bootstrap ($F_e$ is estimated).

## Limit Model-free Prediction

In practice, it is generally not easy to figure out $H_n$ and its inverse. A so-called Limit Model-free Prediction (LMF) method can circumvent some difficulties:

1. Determine $Y_n$ in terms of $\mathbf{Y}_{n-1}$, $X_n$ and $e_n$, i.e., $Y_n = g_n(\mathbf{Y}_{n-1}, X_n, e_n)$; $e_n \sim F_e$;
2. Same as Steps 3-4 of the Model-free Prediction Principle.

In short, the LMF prediction framework just needs the inverse of $H_n$.

**Noise outsourcing lemma (Kallenberg, 1997):**

Let $X$ and $Y$ be random variables with joint distribution $P_{X,Y}$. Then, there is a measurable function $G : [0,1] \times \mathcal{X} \to \mathcal{Y}$ such that

$$(X, Y) \overset{a.s.}{=} (X, G(X, Z)), \quad \text{where } Z \sim \text{Uniform}[0,1] \text{ and } Z \perp\!\!\!\perp X.$$

In particular, $Y \overset{a.s.}{=} G(X, Z)$. In other words, the randomness in the conditional $P_{Y|X=x}$ is outsourced to $Z$ through $G(x, Z)$ as $G$ is deterministic.

**Our extension (LMF via noise outsourcing lemma):**

Under our basic assumptions, there is a continuous $\widetilde{G}(\cdot, \cdot)$ which maps $A := \mathcal{X} \times \mathcal{Z}$ to $\mathcal{Y}$ such that $\widetilde{G}(x, z) = G(x, z)$ for all $(x, z) \in D \subseteq A$; here $\lambda(A \backslash D) < \epsilon$ for $\forall \epsilon > 0$; $\lambda$ denotes the Lebesgue measure; $\mathcal{Z}$ could be $\mathbb{R}^p$ or $[0,1]^p$ if we take $Z$ as $N(0, I_p)$ or Uniform$[0,1]^p$, respectively, for some positive integer $p$. $\widetilde{G}$ can be taken as the inverse transformation function in LMF prediction.

**Quantile Prediction Interval (QPI):**

The conditional distribution of $Y_f$ given $X_f = x_f$ can be approximated by the Monte Carlo method with $\widetilde{G}(x_f, Z)$, so the conditional QPI of $Y_f$ can be obtained, but it is not satisfied for finite samples in practice; see Wang and Politis (2021).

## Approximate $\widetilde{G}$ by DNN

Define

$$\widehat{H} := \arg\min_{H_\theta \in \mathcal{F}_{\text{DNN}}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - H_\theta(X_i, Z_i))^2; \tag{1}$$

where $\mathcal{F}_{\text{DNN}}$ is an appropriate DNN class; we call $\{Z_i\}_{i=1}^n$ reference random variables which can be simulated from a simple distribution.

$\widehat{H}(X, Z)$ is an approximation to $H_0(X, Z) := \arg\min_H \mathbb{E}\,(Y - H(X, Z))^2$.

Intrinsically different with standard LS optimizer, $H_0(X, Z)$ can be thought as:

- A projection of $Y$ onto an extension of $\mathcal{S}_X$ by random variable $Z$; $\mathcal{S}_X$ is a closed subspace of $L^2$ space, which contains all functions of $X$;
- A $\mathcal{D}_{(X,Z)}$-measurable function; $\mathcal{D}_{(X,Z)}$ is the $\sigma$-algebra generated by $(X, Z)$.

## Capture DNN Estimation Variability

**Motivation:** LMF prediction framework with $\widetilde{G}$ can eliminate error in $\sim$. However, additional error in $\approx$ due to estimation still exists since we can only have $\widehat{H}$. As a result, the conditional Prediction Interval (PI) based on $\widehat{H}(x_f, Z)$ undercovers $Y_f$.

**Pertinent PI (PPI):** Politis (2015) proposed the concept of pertinence to capture the estimation variabilities based on re-sampling techniques.

In short, the fundamental idea of building PPI is approximating the predictive root $R_f$ by the variant $R_f^*$ in the bootstrap world, i.e., conditional on $\{(X_i, Y_i, Z_i)\}_{i=1}^n$:

$$R_f^* \xrightarrow[d]{Approximate} R_f;$$

where,

- $R_f$ could be $Y_f - \widehat{Y}_{f,L_2}$; $Y_f \sim P_{Y|x_f}$ and $\widehat{Y}_{f,L_2} := \mathbb{E}(\widehat{H}(x_f, Z))$ is the optimal $L_2$ point prediction; we approximate it by $\frac{1}{S}\sum_{s=1}^{S} \widehat{H}(x_f, Z_s)$;
- $R_f^*$ could be $Y_f^{(b)} - \widehat{Y}_{f,L_2}^{(b)}$; $Y_f^{(b)} \sim \widehat{H}(x_f, Z)$ and $\widehat{Y}_{f,L_2}^{(b)} := \mathbb{E}(\widehat{H}^{(b)}(x_f, Z))$ is the optimal $L_2$ point prediction conditional on pseudo training data generated by $\widehat{H}$; we approximate it by $\frac{1}{S}\sum_{s=1}^{S} \widehat{H}^{(b)}(x_f, Z_s)$; $\widehat{H}^{(b)}$ is the re-estimation of $\widetilde{G}$ based on the $b$-th pseudo training data.

Thus, an asymptotically pertinent PI with $1-\alpha$ coverage rate centered at $\widehat{Y}_{f,L_2}$ is:

$$\left[\widehat{Y}_{f,L_2} + Q_{\alpha/2}, \widehat{Y}_{f,L_2} + Q_{1-\alpha/2}\right];$$

$Q_{\alpha/2}$ and $Q_{1-\alpha/2}$ are $\alpha/2$ and $1-\alpha/2$ lower quantiles of $P_{R_f^*}$, the distribution of $R_f^*$. In practice, $P_{R_f^*}$ can be approximated by the empirical distribution of $\{Y_f^{(b)} - \widehat{Y}_{f,L_2}^{(b)}\}_{b=1}^B$.

## Simulation

**Data generating model:**

$$Y_i = X_{i,1}^2 + \exp(X_{i,2} + X_{i,3}/3) + X_{i,4} - X_{i,5} + \left(0.5 + X_{i,2}^2/2 + X_{i,5}^2/2\right) \cdot \varepsilon_i;$$

where $X_i$ and $\varepsilon_i$ are simulated from $N(0, I_5)$ and $N(0, 1)$.

**PI candidates:** Quantile PI (QPI) and PPI based on LMF prediction idea, PI-KL and PI-WA (based on deep generative method with adversarial training; see Zhou et al. (2023) and Liu et al. (2021)). All PIs are built with the same hyperparameters.

**Evaluation criterion:**

$$\text{CR} := P(Y_f \in \widehat{\mathcal{I}}),$$

approximated by $\frac{1}{T}\frac{1}{K}\sum_{k=1}^{K}\sum_{t=1}^{T} P(Y_f \in \widehat{\mathcal{I}} | x_f^t, \{(X_i^k, Y_i^k)\}_{i=1}^n)$; $x_f^t$ is the $t$-th test point; $\{(X_i^k, Y_i^k)\}_{i=1}^n$ is the $k$-th training set; $\widehat{\mathcal{I}}$ represents PI; $K = 200$; $T = 2000$.

Table 1. Simulation results of CR with varying $n$ and $p$ for different PIs.

| | CR | AL | CR | AL | CR | AL |
|---|---|---|---|---|---|---|
| p = 5 | n = 200 | | n = 500 | | n = 2000 | |
| QPI | 0.861(0.170) | 5.487(1.054) | 0.927(0.110) | 6.734(1.463) | 0.787(0.177) | 3.621(0.855) |
| PPI | 0.893(0.139) | 6.208(1.384) | 0.941(0.095) | 7.258(1.808) | 0.789(0.173) | 3.728(0.959) |
| PI-KL | 0.842(0.193) | 5.496(0.861) | 0.869(0.157) | 5.434(1.218) | 0.913(0.104) | 5.670(2.282) |
| PI-WA | 0.852(0.181) | 5.439(0.907) | 0.882(0.150) | 5.970(2.030) | 0.899(0.105) | 5.365(1.996) |
| p = 10 | | | | | | |
| QPI | 0.928(0.129) | 7.497(0.720) | 0.949(0.094) | 8.194(0.950) | 0.855(0.157) | 4.474(0.817) |
| PPI | **0.944(0.105)** | 8.103(1.072) | 0.961(0.076) | 8.623(1.325) | 0.855(0.154) | 4.546(0.953) |
| PI-KL | 0.900(0.133) | 6.701(0.835) | 0.925(0.119) | 6.806(0.933) | 0.928(0.099) | 5.882(1.403) |
| PI-WA | 0.898(0.146) | 6.757(0.719) | 0.933(0.116) | 7.545(1.340) | 0.934(0.100) | 6.199(1.880) |
| p = 15 | | | | | | |
| QPI | 0.915(0.137) | 7.408(0.669) | 0.945(0.097) | 7.430(0.949) | 0.915(0.123) | 5.895(0.647) |
| PPI | 0.930(0.119) | 7.760(0.936) | **0.953(0.085)** | 7.749(1.172) | 0.916(0.121) | 5.971(0.807) |
| PI-KL | 0.909(0.136) | 7.427(0.817) | 0.949(0.095) | 8.082(1.068) | 0.943(0.089) | 6.556(1.491) |
| PI-WA | 0.901(0.137) | 6.797(0.687) | 0.950(0.095) | 7.972(1.312) | 0.947(0.088) | 6.778(1.541) |
| p = 20 | | | | | | |
| QPI | 0.879(0.172) | 6.726(0.485) | 0.959(0.085) | 8.830(0.683) | 0.940(0.102) | 6.849(0.562) |
| PPI | 0.893(0.154) | 6.941(0.702) | 0.966(0.073) | 9.100(0.950) | 0.942(0.097) | 6.925(0.759) |
| PI-KL | 0.923(0.126) | 7.799(0.842) | 0.954(0.087) | 8.311(0.861) | 0.946(0.093) | 6.806(1.097) |
| PI-WA | 0.910(0.140) | 7.402(0.698) | 0.945(0.099) | 8.011(0.800) | 0.946(0.092) | 6.804(1.534) |
| p = 25 | | | | | | |
| QPI | 0.871(0.172) | 7.020(0.287) | 0.961(0.088) | 9.633(0.645) | 0.946(0.099) | 7.296(0.475) |
| PPI | 0.884(0.160) | 7.189(0.548) | 0.967(0.078) | 9.881(0.938) | **0.948(0.095)** | 7.370(0.695) |
| PI-KL | 0.907(0.142) | 7.370(0.618) | 0.954(0.090) | 8.670(0.813) | 0.945(0.093) | 6.915(1.009) |
| PI-WA | 0.897(0.151) | 7.071(0.510) | 0.960(0.081) | 8.514(0.942) | 0.944(0.097) | 7.117(1.491) |

## Future Work

- Explore the possibility of applying the Model-free prediction idea on other machine learning tasks, e.g., classification;
- Combine the LMF prediction idea with LLM.

## References

Kallenberg, O. (1997). *Foundations of modern probability Second Edition*. Springer.

Liu, S., Zhou, X., Jiao, Y., and Huang, J. (2021). Wasserstein generative learning of conditional distribution. *arXiv preprint arXiv:2112.10039*.

Politis, D. N. (2015). *Model-free prediction in regression: A transformation-based approach to inference*. Springer.

Wang, Y. and Politis, D. N. (2021). Model-free bootstrap and conformal prediction in regression: Conditionality, conjecture testing, and pertinent prediction intervals. *arXiv preprint arXiv:2109.12156*.

Zhou, X., Jiao, Y., Liu, J., and Huang, J. (2023). A deep generative approach to conditional sampling. *Journal of the American Statistical Association*, 118(543):1837–1848.