Research Statement

Kejin Wu, Assistant Professor
Department of Mathematics and Statistics, Loyola University Chicago

My past and ongoing research focuses on two major topics: the prediction inference of dependent or independent data under various scenarios, and the uncertainty quantification of diverse estimators using bootstrap or subsampling techniques. With the recent information explosion, developing statistical inference in a computationally feasible way has also become one of my research interests. My goal is to develop systematic approaches to provide meaningful estimation/prediction inference of various statistical tasks with user-chosen methods while imposing minimal assumptions.

Research Overview

The diagram below summarizes the key points of my research:

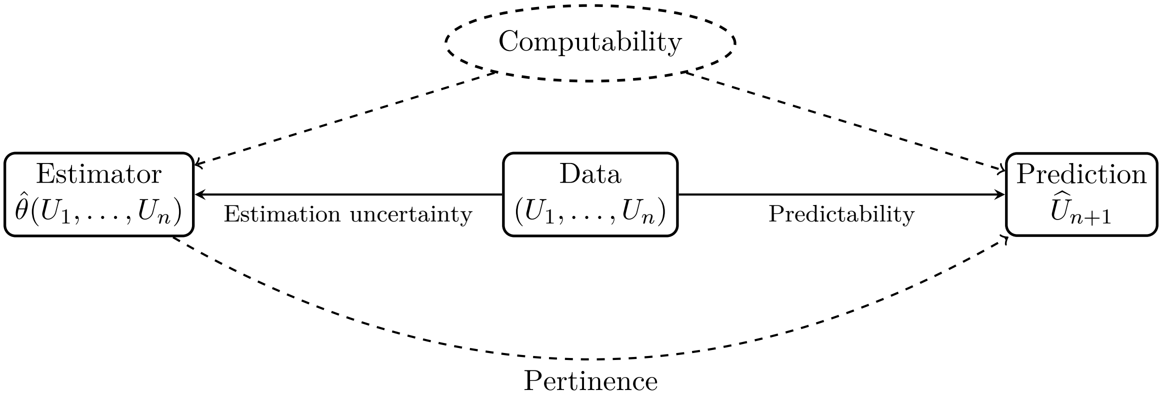

Figure 1: Research diagram.

In Figure 1, the two solid arrow lines stand for the direct application of data to do estimation and prediction where the *estimation uncertainty* and *predictability* should be considered, respectively. Dashed arrow lines indicate other complementary factors that require attention, namely the *computability* for both prediction and estimation processes and *pertinence* for the prediction inference.

The data $(U_1, \ldots, U_n)$ can represent independent or dependent data. For example, $(U_1, \ldots, U_n)$ can be time series data indexed by discrete time points from 1 to n. Alternatively, U_i can represent the independent data pair (Y_i, X_i) in the regression setting for $i = 1, \ldots, n$. For the estimation tasks, the estimator $\hat{\theta}(U_1, \ldots, U_n)$ based on data approximates the quantity $\theta(P)$ which relies on the underlying distribution P. Meanwhile, the estimation accuracy of $\hat{\theta}(U_1, \ldots, U_n)$ can be quantified by Confidence Interval (CI). This perspective is related to the concept of *estimation uncertainty.* In our work IV, we investigated the estimation inference of the trained DNN, denoted by $\hat{f}_{\text{DNN}}$, as a whole term in the regression context. More specifically, we constructed the CI with $\hat{f}_{\text{DNN}}(x)$ at any point of x in the domain of X with a subsampling technique; we also approximated the bias order of $\hat{f}_{\text{DNN}}(x)$.

In contrast, the prediction $\widehat{U}_{n+1}$ of the future occupies another important application in the real world. Classically, people rely on some model assumption to depict the data at hand, and then make predictions based on the estimated model variant. Following this traditional approach, we developed the methodology to make multi-step ahead point predictions and asymptotically valid Prediction Interval (PI) for non-linear time series in work I. Asymptotic validity means the coverage rate of PI equals the nominal confidence level asymptotically. However, as the famous saying goes, ‘‘all models are wrong, but some are useful”. I am also interested in exploring the *predictability* of the future without any restricted model assumption, which is known as Model-free prediction. In work II, we put forward several Model-free/Model-based methods to forecast the volatility of stock returns. The goal is to find an invertible transformation function that maps $\{U_t\}_{t=1}^n$ to an i.i.d. sequence $\{W_t\}_{t=1}^n$ so that the prediction of U_{n+1} can be derived from the straightforward prediction of W_{n+1}. Also, we explored the predictability of the purely MF approach in the regression context with the help of DNN in work III. It turns out that the future distribution of Y_{n+1} can be mimicked by a DNN function f_{DNN} which takes X_{n+1} and auxiliary variable Z_{n+1} as input, i.e., $Y_{n+1} \overset{a.s.}{=} f_{\text{DNN}}(X_{n+1}, Z_{n+1})$.

The prediction and estimation stand for their distinct interests separately, but they should be considered simultaneously, especially when the prediction inference is desired with finite samples. In essence, the error involved in prediction comes from two sources: one is the inherent randomness from the future value, and the other is all estimation errors generated in the prediction process. An effective PI needs to capture both variabilities even with finite samples, a property known as *pertinence*. Guided by this spirit, we designed specific algorithms to capture the estimation variabilities for predictions with time series and regression data in work I and III.

In addition to the methodologies of making estimation and prediction inferences, another crucial factor that should be kept in mind is the *computability*, especially for the statistical analysis process fueled by big data or complicated estimators. Inspired by the divide-and-conquer approach, which was originally applied in the algorithm to reduce the computational complexity, we demonstrated that Deep Neural Networks (DNN) can be trained more quickly and accurately in work IV with a scalable subsampling framework proposed recently. Extensions of this approach to speed up the running time of other complex estimators are also possible.

Research Summary

In this section, I will provide a detailed introduction to my past work I to IV.

I Model-based prediction of a special type of time series

In the domain of univariate time series analysis, single- or multi-step ahead predictions are crucial for forecasting crop yields, stock prices, traffic volume, etc. Conventionally, people usually take a Model-based (MB) approach, where some specific underlying model is assumed to depict the data and then predictions are made based on this model. A flexible choice is the Autoregressive (AR) model, expressed as $U_n = f(\boldsymbol{U}_{n-1}) + g(\boldsymbol{U}_{n-1}) \cdot \epsilon_n$; $\boldsymbol{U}_{n-1}$ represents the vector $(U_{n-1}, \ldots, U_{n-p}) \in \mathbb{R}^p$; $f(\cdot)$ and $g(\cdot)$ can be any appropriate functions as long as some mixing or weak dependence property of time series is satisfied; ϵ_n is the *i.i.d.* innovation term of the model. For Linear AR (LAR) models, the k-step ahead L_2 optimal point prediction (with respect to L_2 risk) can be obtained by iterating the one-step ahead predictor with a known or estimated model; $k \in \mathbb{Z}^+$. However, the LAR model may not be enough to analyze complicated data in the real world, necessitating the use of Non-linear AR (NLAR) models. Unfortunately, the optimal multi-step ahead prediction (in terms of L_2 or L_1 risk) of NLAR can not be achieved through the iterative procedure used for LAR models. This challenge also extends to constructing PI.

Our works [WP24a] and [PW23] provide two model options: (a) $f(\cdot)$ and $g(\cdot)$ are assumed to have a known non-linear form, but the corresponding parameters are unknown; (b) $f(\cdot)$ and $g(\cdot)$ are assumed to possess some smoothness property and estimated by non-parametric methods. Regardless of which option is applied, we propose a general algorithm to solve the forecasting difficulty with the NLAR model. The key step involves simulating pseudo values of innovation $(\epsilon^*_{n+1},\ldots,\epsilon^*_{n+k})$ from the distribution P_ϵ if it is known, or bootstrapping $(\hat{\epsilon}^*_{n+1},\ldots,\hat{\epsilon}^*_{n+k})$ from the empirical distribution of the residuals $\widehat{P}_\epsilon$. Then, a pseudo value U^*_{n+k} can be attained by iterating the model equation based on $\boldsymbol{U}_n$ and $\{\epsilon^*_{n+i}\}_{i=1}^k$ or $\{\hat{\epsilon}^*_{n+i}\}_{i=1}^k$. By repeating the simulation or the bootstrap, the conditional distribution of future variable $P_{U_{n+k}|\boldsymbol{U}_n}$ can be estimated by the empirical distribution of $\{U^{(b)}_{n+k}\}_{b=1}^B$, denoted as $\widehat{P}_{U^*_{n+k}|\boldsymbol{U}_n}$; each $U^{(b)}_{n+k}$ is a pseudo value of U_{n+k}. As $n\to\infty$ and $B\to\infty$, we show that the mean and median of $\widehat{P}_{U^*_{n+k}|\boldsymbol{U}_n}$ are consistent with L_2 and L_1 optimal point prediction of U_{n+k}, respectively. Also, the naive quantile PI based on $\widehat{P}_{U^*_{n+k}|\boldsymbol{U}_n}$ is asymptotically valid. Moreover, when estimation error is involved, e.g., P_ϵ and the model are replaced by their corresponding estimators, we build the Pertinent PI (PPI) by leveraging the following uniform consistency result under standard assumptions:

$$\sup_{|x|\leq c_n}\left|\mathbb{P}\left(U^*_{n+k}-\widehat{U}^*_{n+k}\leq x|U_n,\ldots,U_{n-p+1}\right)-\mathbb{P}\left(U_{n+k}-\widehat{U}_{n+k}\leq x|U_n,\ldots,U_{n-p+1}\right)\right|\xrightarrow{p}0;$$

$\mathbb{P}\left(\cdot|\cdot\right)$ is the conditional probability; $c_n\to\infty$ in an appropriate rate as $n\to\infty$; $\widehat{U}^*_{n+k}$ and $\widehat{U}_{n+k}$ represent meaningful point predictions in the bootstrap- and real-world, respectively, e.g., the optimal L_2 point prediction. The PPI is centered around $\widehat{U}_{n+k}$ and its two endpoints are determined by the quantile of the conditional distribution of $U^*_{n+k}-\widehat{U}^*_{n+k}$. In other words, the estimation variabilities embedded in the prediction can be captured by approximating the conditional distribution of the predictive root $U_{n+k}-\widehat{U}_{n+k}$ as a whole term by the bootstrap.

II Model-free/Model-based forecasting of volatility in financial returns

To alleviate the model limitation in the prediction process, the ultimate goal is to develop a Model-free (MF) prediction framework without any restricted model assumptions. Remarkably, it is worth exploring the performance of the intermediate stage between MB and MF approaches, namely the Model-free/Model-based (MFMB) method. For instance, in forecasting the volatility of univariate financial returns, the existing MFMB method builds on the prior Autoregressive Conditional Heteroskedasticity (ARCH) model as follows:

$$U_t=W_t(a+\sum_{i=1}^p a_iU_{t-i}^2)^{1/2};\ \text{for } t=p+1,\ldots,n;$$

$a\geq 0$ and $a_i\geq 0$ for all $i=1,\ldots,p$; $W_t\sim i.i.d.\ N(0,1)$; series $\{U_t\}$ represents financial log-returns. Subsequently, a *transformation* function can be modified from the ARCH model to connect two probability spaces, i.e.,

$$W_t=U_t/(\alpha s_{t-1}^2+b_0U_t^2+\sum_{i=1}^p b_iU_{t-i}^2)^{1/2}\ ;\ \text{for } t=p+1,\ldots,n;$$

$\{W_t\}$ should be understood as the transformed vector here; α is a fixed, scale-invariant constant; s_{t-1}^2 is an estimator of the variance of $\{U_1,\ldots,U_{t-1}\}$; $\{b_i\}_{i=0}^p$ is no longer the coefficient vector of a model, but instead has an exponentially decayed form, i.e., $b_0=c'$, $b_i=c'e^{-ci}$, for all

$1 \leq i \leq p$, $c' = \frac{1-\alpha}{\sum_{j=0}^{p} e^{-cj}}$. Considering the inverse form the transformation function, we derive the equation: $U_{n+1} = \sqrt{W_{n+1}^2(\alpha s_n^2 + \sum_{i=1}^{p} b_i U_{n+1-i}^2)/(1 - b_0 W_{n+1}^2)}$. Ideally, if $\{W_t\}_{t=p+1}^{n}$ are *i.i.d.* with a simple distribution, we can execute bootstrap on $\{W_t\}_{t=p+1}^{n}$ to generate pseudo values of U_{n+1}. Subsequently, the point prediction and PI can be decided based on the resulting empirical distribution. Similarly, k-step ahead prediction inference can be conducted.

In short, rather than estimating a model, the MFMB method aims to find an optimal transformation such that $\{W_t\}_{t=p+1}^{n}$ are as close to *i.i.d.* as possible. Moreover, the transformation function depends on only two unknown parameters α and c, which offers significant advantages when the available sample is short compared to the demand of estimating $p + 1$ coefficients in the standard ARCH model. Our works [WK21], [WK23] and [WKG23] evaluate the current MFMB method for long-term time aggregated forecasting and explore other transformation forms based on GARCH and GARCHX models.

III Model-free prediction of regression based on DNN

An MF method neither makes restricted model assumptions nor arises from any prior model. Take the regression task as an example, our purpose is the prediction inference of $Y_f \in \mathbb{R}$ given a new independent variable $X_f \in \mathbb{R}^d$ with only assumptions about the joint distribution of $P_{X,Y}$. To realize the essence of MF prediction, the objective is to find invertible transformation functions H_{X_i} that transform each Y_i to W_i conditional on X_i for $i = 1, \ldots, n$; W_i should have a simple distribution P_W conditional on X_i. In other words, the collection of functions $\{H_{X_i}\}_{i=1}^{n}$ connects two equivalent probability spaces: one is the space of $\{(X_i, Y_i)\}_{i=1}^{n}$ and the other is the space of $\{(X_i, Z_i)\}_{i=1}^{n}$; here X_i and Z_i are mutually independent.

We consider the MF prediction with DNN estimators in [WP24b]. This combination kills two birds with one stone: (1) According to folk wisdom and empirical/theoretical evidence, DNN suffer less from the curse of dimensionality compared to the kernel estimators used in the current MF method; (2) Although DNN is an unstable estimator that may vary widely across different samples, the MF prediction gives a method to capture the estimation variability of DNN when predictions are needed. Consequently, our work develops a so-called *Deep Limit Model-free* (DLMF) prediction method. Moreover, we establish the theoretical foundation of DLMF starting from a concept known as the noise-outsourcing lemma. Under mild conditions, it turns out that there is a continuous $\Xi(\cdot,\cdot) : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ such that $\Xi(x, z) = H_0(x, z)$ for all $(x, z) \in D \subseteq A$; here $\lambda(A \backslash D) < \epsilon$ for $\forall \epsilon > 0$; A is space $\mathcal{X} \times \mathcal{Z}$; λ denotes the Lebesgue measure; $\mathcal{X}, \mathcal{Z}$ and $\mathcal{Y}$ are domains of X, Z and Y, respectively; H_0 is a measurable function such that $Y \overset{a.s.}{=} H_0(X, Z)$. $\mathcal{Z}$ 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. Therefore, $\Xi(\cdot,\cdot)$ can be a candidate for the transformation function. By the universal approximation property of DNN, $\Xi(\cdot,\cdot)$ can be estimated arbitrarily well as long as the sample size is large. Likewise, PPI can be considered in conjunction with the DLMF method. As revealed by simulation and empirical studies, the PPI of DLMF outperforms the PI generated by other deep generative counterparts.

IV Scalable subsampling inference of DNN

The computational challenges in statistical analysis have raised growing concerns as sample sizes continue to increase. Computationally intensive methods, such as bootstrap, place a heavy load on the CPU when dealing with large datasets because the estimation processes are repeated on many resamples, each as large as the original sample. Even with the subagging method, in which the estimation is performed with all b-size subsamples from an n-size dataset, the computational cost remains substantial when n and b are large since choosing a single random subsample needs $O(b)$

time and space which corresponds to optimal time and space complexity for this task.

Recently, a scalable subsampling (SS) technique was proposed by [Pol24]. SS involves creating $q = \lfloor (n-b)/h \rfloor + 1$ non-stochastic subsamples from the dataset $\{U_1, \ldots, U_n\}$; $\lfloor \cdot \rfloor$ denotes the floor function; $U_i := (X_i, Y_i)$; $X_i \in \mathbb{R}^d, Y \in \mathbb{R}$. These subsamples, denoted as $B_1, \ldots, B_q$, are defined by $B_j = \{U_{(j-1)h+1}, \ldots, U_{(j-1)h+b}\}$; the parameter h controls the amount of overlap (or separation) between B_j and B_{j+1}; the subsample size b and the overlap h are functions of n, but these dependencies will not be explicitly denoted. For a given estimator, if its bias is comparatively negligible to its standard deviation, the overall mean square error bound can be improved by applying the SS technique, i.e., the scalable subsampling estimator being the average of estimators built by all subsamples. Moreover, if $O(n^{\tau})$ number of operations are required to perform an estimation with the full sample, only $O(nb^{\tau-1})$ number of operations are needed for estimations on all subsamples when $b = h$. When τ is large, such as in the implementation of LASSO regression where d increases with n, SS can offer notable computational savings.

Under mild conditions, our work [WP24c] explores the estimation inference of the DNN estimator in the regression context. First, we show that the error bound for a scalable subsampling estimator $\overline{f}_{\text{DNN}}$ satisfies $\|\overline{f}_{\text{DNN}} - f\|^2_{L_2(X)} \leq n^{\frac{-\Lambda}{\Lambda + \frac{d}{\xi+d}}} \mathcal{L}(n)$ with high probability; $\mathcal{L}(n)$ is a slowly varying function involving a constant and all $\log(n)$ terms; $\xi > 0$ is a real number that measures the smoothness of the underlying true regression function; Λ is assumed to be larger than $\frac{\xi}{\xi+d}$. This error bound is superior to existing general and attainable variants when no assumption about the model structure or intrinsic dimension of data is made. Secondly, via a scaling-down estimation technique, we attempt to figure out the bias order of $\overline{f}_{\text{DNN}}$ by finding a rate τ such that $\mathbb{E}(\overline{f}_{\text{DNN}}(x) - f(x)) = O(n^{-\tau/2})$. Thirdly, several algorithms are set up to construct the pointwise CI. In addition, the PI based on the SS estimator is discussed.

Future Research Agenda

So far, my research work has been motivated by thoughts of estimation uncertainty, predictability and computability. In the same manner, there are ample potential research extensions. I summarize some of them below. Beyond these individual projects, I am open to collaborating with other researchers in areas where my expertise can contribute meaningfully.

I Model-free prediction of dependent data

A direct extension of current work is applying the MF prediction philosophy to dependent data. For example, with time series data, it is possible to model it by a general equation: $U_n = G(\boldsymbol{U}_{n-1}, \epsilon_n)$. Instead of thinking $G(\cdot, \cdot)$ as a model to explain the underlying time series generating process, it is more appropriate to view it as a *generator* that maps $\boldsymbol{U}_{n-1}$ and ϵ_n together to U_n. This viewpoint parallels the transformation function used in the MFMB prediction, but no prior model is used. As a result, this prediction approach is not constrained by any model assumption. Ideally, it could encompass a broad range of true underlying models if they exist.

II Volatility forecast sparked by other transformation functions

A general consensus in the volatility forecasting literature is that the realized variance is a more accurate measure of volatility than squared returns. Typically, the realized variance can be computed from high-frequency auxiliary information, such as intraday transaction prices, bid/ask quotes, and trading volume. Motivated by this agreement, the realized variance can enhance existing volatility models resulting in parallel GARCH and realized GARCH models. In addition, the INGARCH model has wide applications in areas involving count time series such as epidemiology, finance, and sports. It is interesting to see the performance of the MFMB method sparked by these GARCH variants.

III Estimation uncertainty in other statistical tasks

DNN has also been applied to other statistical tasks. For example, a DNN that takes quantile values and independent variables as inputs can return the value of dependent variables, enabling the estimation of a quantile process. Additionally, a DNN can serve as an estimator for the conditional treatment effects in the field of causal inference. Therefore, it is appealing to quantify the estimation uncertainties within these problems by scalable subsampling technique. We anticipate a decrease in the computational aspect and an improvement in the estimation accuracy. Besides, the quantification of the causal structure uncertainty of time series can be another interesting topic.

IV Calibration prediction interval with DNN

Apart from capturing the estimation uncertainties using bootstrap/subsampling, a calibration approach based on DNN can be operated to correct the coverage level of PI. The main idea is that the conditional CDF values at a grid of points spanned within a user-determined region can be estimated by DNNs. Then, the PI for new dependent variables can be determined in a calibration manner based on these grid points. It is well known that it is impossible to make a PI with a finite length that guarantees at least $1-\alpha$ coverage for Y conditional on $X_f = x_f$ for any distribution of $P_{X,Y}$ and all $x_f \in N(P)$, which is the so-called non-atoms. With some compromises on the estimation of DNN, we make an effort to show that it is possible to build a calibration PI that guarantees at least $1-\alpha$ converge rate even for finite samples when some specific distribution $P_{X,Y}$ is assumed.

Mentoring Plan

I firmly believe that every student possesses innate talent and potential. As a faculty member, I will design research plans that ignite students curiosity and help them build confidence step by step. My research spans both theoretical and empirical areas, including forecasting volatility of financial series, time series analysis, the theoretical foundations of model-free prediction, and the analysis of various subsampling estimators. For undergraduate students, application-based projects can broaden their understanding of statistics and trigger their interest in the field. For graduate students, I will guide them through theoretical projects that enhance their research skills. Whether mentoring undergraduates or graduate students, I am committed to providing the support and guidance they need to solve problems independently. Additionally, I will actively mentor and advise students from underrepresented groups.

References

[Pol24] Dimitris N Politis. Scalable subsampling: computation, aggregation and inference. *Biometrika*, 111(1):347--354, 2024.

[PW23] Dimitris N Politis and **Kejin Wu**. Multi-step-ahead prediction intervals for nonparametric autoregressions via bootstrap: Consistency, debiasing, and pertinence. *Stats*, 6(3):839--867, 2023.

[WK21] **Kejin Wu** and Sayar Karmakar. Model-free time-aggregated predictions for econometric datasets. *Forecasting*, 3(4):920--933, 2021.

[WK23] **Kejin Wu** and Sayar Karmakar. A model-free approach to do long-term volatility forecasting and its variants. *Financial Innovation*, 9(1):59, 2023.

[WKG23] **Kejin Wu**, Sayar Karmakar, and Rangan Gupta. Garchx-novas: A model-free approach to incorporate exogenous variables. *arXiv preprint arXiv:2308.13346*, 2023. *Submitted to Journal of Forecasting.*

[WP24a] **Kejin Wu** and Dimitris N Politis. Bootstrap prediction inference of nonlinear autoregressive models. *Journal of Time Series Analysis*, 2024.

[WP24b] **Kejin Wu** and Dimitris N Politis. Deep limit model-free prediction in regression. *arXiv preprint arXiv:2408.09532*, 2024. *Submitted to ACM/IMS Journal of Data Science.*

[WP24c] **Kejin Wu** and Dimitris N Politis. Scalable subsampling inference for deep neural networks. *arXiv preprint arXiv:2405.08276*, 2024. *Accepted by ACM/IMS Journal of Data Science.*