Research Statement

Guoyi Zhang

07/29/2023

My primary research interests revolve around nonparametric function estimation and computational statistics. Much of my work involves developing new statistical theory, methods, and algorithms within these domains. With an educational background in engineering, management, and statistics, my research extends to machine learning, survey sampling, mixed models, financial engineering, and applications in healthcare. Among my 37 research articles, 30 have been published in high-quality refereed statistical journals or journals from other applied areas. Seven of these papers are single-authored, and I have served as the first author on 23 of them. Additionally, 12 of the articles are co-authored with graduate students and 23 of them have been done since I became an Associate in August 2015. Throughout my research career, I have encountered and overcome various challenges while exploring new questions and ideas. These experiences have deepened my appreciation for the beauty of statistics, and I thoroughly enjoy engaging in research.

In the subsequent sections, my aim is to present a summary of my research accomplishments and shed light on the future avenues I plan to pursue in four significant domains since 2015: (1) nonparametric function estimation, (2) computational and applied statistics, (3) applications, and (4) other ongoing research. The intent behind this discussion is to make the content accessible to nonspecialists, focusing on providing a broad perspective rather than delving into intricate historical and technical details that can be found in the referenced papers.

1 Nonparametric Function Estimation

Nonparametric function estimation is a powerful statistical tool that has been a focal point of my research. My exploration in this field began with studying confidence bands in nonparametric regression (Zhang & Lu, 2008). During my dissertation, I delved into smoothing spline estimators for multivariate regression (Zhang, 2011, 2012a). Subsequently, my investigations expanded to encompass generalized additive partially linear models, extensions for complex survey data, and advancements in technical analysis. Currently, my research focuses on statistical learning for multiple tasks using kernel methods, as well as developing nonparametric regression tests specifically tailored for survey data.

1.1 Smoothing splines and generalized additive partially linear models

Spline smoothing serves as a crucial statistical tool for nonparametric function estimation. Among various approaches, smoothing splines stand out due to their computational efficiency. However, extending smoothing splines to higher dimensional settings has posed challenges. Even the widely used thin plate splines have a complexity of $O(n^3)$ for a sample of size n, which renders them computationally slow and impractical for big data applications. In my dissertation research (Zhang, 2011, 2012a), I tackled this issue by developing a high-dimensional smoothing spline that achieves comparable estimation accuracy while maintaining computational efficiency.

Generalized additive models (GAMs) are an effective regression tool for analyzing high-dimensional data. A more flexible extension of GAM is the generalized additive partially linear model (GAPLM), which incorporates both parametric and nonparametric components in the modeling process. Let Y be a response random variable, and let the predictors be divided into two groups: $\mathbf{T}$ and $\mathbf{X}$. The group $\mathbf{T}$ is used in a linear model, while the predictors X_h, where $h = 1, 2, \cdots, d_2$, are used in a generalized additive model. We define $\mathbf{T} = (1, T_1, \cdots, T_{d_1})^T$, $\mathbf{X} = (X_1, \cdots, X_{d_2})^T$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_{d_1})^T$ as the parameter vector, where the superscript T denotes the transpose. Assuming a fixed σ-finite measure, the probability density function of Y_i conditional on $\mathbf{X}_i$ and $\mathbf{T}_i$ is given by the exponential family as follows:

$$f(Y_i|\mathbf{X}_i, \mathbf{T}_i, \phi) = \exp\left\{Y_i \cdot m(\mathbf{X}_i, \mathbf{T}_i) - b\{m(\mathbf{X}_i, \mathbf{T}_i)\}/a(\phi) + h(Y_i, \phi)\right\}.$$

The expected value of Y given $\mathbf{T}$ and $\mathbf{X}$ can be expressed as a function of the mean $m(\mathbf{T}, \mathbf{X})$:

$$E(Y|\mathbf{T}, \mathbf{X}) = b'\{m(\mathbf{T}, \mathbf{X})\}, \tag{1}$$

where $m(\mathbf{T}, \mathbf{X}) = \boldsymbol{\beta}^T\mathbf{T} + \sum_{h=1}^{d_2} m_h(X_h)$ and $m_h(X_h)$ represents the nonparametric component function of the GAPLM. The functions b' and b'' are the first and second derivatives of a function b that implicitly relates $m(\mathbf{t}, \mathbf{x})$ to the conditional variance function $\sigma^2(\mathbf{t}, \mathbf{x}) = \mathrm{Var}(Y|\mathbf{T} = \mathbf{t}, \mathbf{X} = \mathbf{x})$. In our research (Liu, Härdle, & Zhang, 2017), we assume that $\mathrm{Var}(Y|\mathbf{T} = \mathbf{t}, \mathbf{X} = \mathbf{x}) = a(\phi)b''[(b')^{-1}\{E(Y|\mathbf{T} = \mathbf{t}, \mathbf{X} = \mathbf{x})\}]$, where $a(\phi)$ is a nuisance parameter that quantifies overdispersion.

Several estimation methods for Model (1) have been proposed, but they either lack theoretical justification (i.e., missing asymptotic properties and confidence bands for the estimators of nonparametric components) or are computationally expensive. To improve upon the

current methods, we extend the hybrid spline-backfitted kernel (SBK) methods from GAM to GAPLM. SBK combines the favorable characteristics of spline and kernel methods, offering fast, efficient, and reliable inference on component functions. The estimation procedure involves two steps. In the first step, the parameters $\boldsymbol{\beta}$ and the spline estimation of nonparametric component functions $m_h(X_h)$ are obtained using quasi-likelihood. In the second step, kernel smoothing is employed to enhance the accuracy and reliability of the component function estimation. The asymptotic properties are derived, and simulation results provide support for the theoretical properties.

1.2 Extension in complex surveys

In the realm of complex surveys, such as the American Communities Survey conducted by the U.S. Census Bureau, the data collection process involves sampling from strata (e.g., states and major metropolitan areas) using independent samples. Additionally, within each stratum, clusters of observations (e.g., counties within a state and households within communities) are randomly sampled. The inherent dependencies within clusters violate the assumptions of independence required for traditional nonparametric techniques. Therefore, it becomes imperative to develop estimation approaches that explicitly account for the dependence that occurs in complex surveys. In my research on the application of nonparametric methods to complex surveys, I have achieved significant progress.

1.2.1 Smoothing splines estimators for Complex Surveys

In a groundbreaking study, Zhang, Christensen, and Zheng (2015) were the first to investigate the use of smoothing splines for modeling the regression mean structure in the context of complex surveys. We introduced a weighted smoothing spline criterion and derived estimators based on smoothing splines. Through the use of a fully data-driven bandwidth selection method, our simulation studies demonstrated the undesirability of ignoring sampling weights and showcased the superior performance of smoothing splines compared to the local linear estimator. Additionally, the proposed generalized smoothing spline offers a valuable tool for imputing missing data in cases where the assumption of linear regression imputation is not applicable.

1.2.2 Adjusted confidence band for Complex Surveys

Zhang, Mao, and Cheng (2016) extended the development of confidence bands by Zhang and Lu (2008) from the independent and identically distributed (iid) case to complex surveys, and

investigated their asymptotic properties. The proposed confidence bands are constructed in the form of $\hat{m}(x) \pm c \cdot l_\alpha(x)$, where $\hat{m}(x)$ represents the estimated mean, c is an adjusted constant that needs to be estimated to expand the confidence band to account for bias, and $l_\alpha(x)$ is a bound. To derive these estimators, we incorporated both the sampling weights and kernel weights.

To estimate the constant c, simulation studies were conducted. A grid of equally spaced points, c_i for $i = 1, 2, ..., 1000$, ranging between 1 and an upper bound of c, was created. A sequence of binary responses, h_i, was generated corresponding to 1 if the confidence band included the true mean and 0 otherwise. A logistic regression model was then employed to fit the data $\{(h_i, c_i)\}$ for $i = 1, 2, ..., 1000$ in order to obtain an estimate of c.

One potential application of the confidence bands is a lack-of-fit test. For instance, suppose we want to test a parametric linear null hypothesis against a nonparametric alternative. By examining whether the regression function under the null hypothesis is entirely contained within the confidence band, we can make a decision to either reject or accept the null hypothesis based on the test results.

1.2.3 Nonparametric Regression Estimators in Dual Frame Surveys

In the field of survey research, I have also worked extensively on dual frame surveys. Traditionally, large-scale surveys rely on a single sampling frame, which consists of a list of population members used to select the sample. However, despite the U.S. Census Bureau providing an excellent frame for surveys like the American Communities Survey, it is important to note that even the Census frame has limitations and may not accurately capture the entire population. Many organizations conducting surveys do not possess the extensive resources available to the Census Bureau. As the population changes and new data collection methods emerge, relying solely on a single frame may lead to the exclusion of certain segments of the population. For example, the increasing number of individuals who primarily or exclusively use cell phones has rendered surveys conducted solely via landline telephones biased and incomplete.

To achieve better coverage of the population of interest and reduce survey costs, there is a growing recognition of the need to employ dual frame surveys. In such surveys, independent samples are drawn from two or more overlapping sampling frames. The combined frames encompass multiple domains, including domain A, domain B, and the overlapping domain AB, as illustrated in Figure 1. By utilizing two frames, dual frame surveys aim to capture a broader representation of the target population and mitigate coverage biases inherent in single frame surveys.

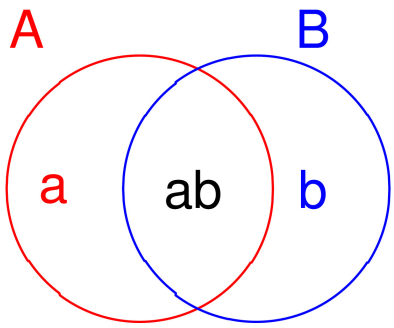

Figure 1: Frames A and B are both incomplete but overlapping

Current research on dual frame surveys primarily focuses on estimating population totals and means. However, in applications such as economic and health surveys, there is often a need to explore relationships among variables, predict new observations, or impute missing values. While linear regression has been discussed in the context of dual frame surveys, it may not adequately capture the complex relationships present in the data.

In a pioneering work on dual frame surveys, Lu, Fu, and Zhang (2021) developed three nonparametric regression estimators and examined their asymptotic properties. Our approach involves transforming the two independent samples into a pseudo single sample by applying adjusted weights. The weights of elements from the overlapping domain (ab) and frame A are adjusted by multiplying θ, while the weights of elements from the overlapping domain (ab) and frame B are adjusted by $(1-\theta)$. This adjustment is necessary due to the overrepresentation of the overlapping domain.

A key challenge is to determine the optimal estimates of θ for the overlapping domain and the bandwidth h for nonparametric regression. To address this, we employed the concepts of Pseudo Maximum Likelihood (Skinner & Rao, 1996) and cross-validation to derive estimates through entirely data-driven methods. Additionally, asymptotic properties were examined under regularity conditions. Simulation results demonstrated the effectiveness of all the proposed methods.

While this research was conducted within the context of survey sampling, the developed techniques can also be applied to other scenarios where data can be combined from two independent sources. Furthermore, these methods have the potential to be extended to situations involving more than two data sources.

1.2.4 Neyman Smooth Type Goodness-of-Fit Tests in Complex Surveys

In the field of categorical data analysis for complex surveys, several goodness-of-fit (GOF) tests have been developed to assess the adequacy of models. These include Wald's test, Fay's jackknifed chi-squared test, Rao and Scott's first and second order corrected tests, and others. However, one limitation of these tests is their lack of sensitivity to slow-varying probabilities. To illustrate this, we can consider an example from Christensen (1997) where a dataset captured information about race, sex, age, and opinions on legalized abortion. The research interest was in testing the hypothesis of no difference in age groups among nonwhite families supporting legalized abortion. Interestingly, despite observing a decreasing trend in the rate of support for legalized abortion among older age groups, both the first and second order corrected tests failed to reject the null hypothesis. This example highlights the need for more sensitive tests that can capture subtle variations in probabilities over different categories or groups.

To address these limitations, Lu, Zhou, Zhang, and Christensen (2021) extended Neyman smooth-type GOF tests to complex surveys. We began by replacing the estimators used in the independent and identically distributed (iid) case with estimators incorporating survey weights. We then introduced basis functions that satisfy orthogonality conditions and employed Fourier transformation to express the test statistic as a sum of a function of the Fourier coefficients. The challenge was to find the optimal estimate of the order q that captures the most information within the first q components. This was achieved by minimizing the mean squared errors through data-driven methods.

Under regularity conditions specific to survey data, we rigorously derived the asymptotic distributions and properties of the Fourier coefficients and test statistics. Simulation studies demonstrated that our proposed methods exhibited improved statistical power while maintaining excellent control over type I error, particularly in cases with slow-varying probabilities. This highlights the advantages of our approach compared to the first and second order corrected tests.

In summary, our work made significant contributions to the field of categorical data analysis for complex surveys by developing Neyman smooth-type GOF tests with improved statistical power and enhanced sensitivity to slow-varying probabilities.

1.3 Extension in technical analysis

Technical analysis, often referred to as "charting," is an investment approach that leverages the analysis of trends in financial markets to make strategic buy and sell decisions with the

aim of maximizing profits. Traditionally, technical analysts have relied on their visual ability to recognize patterns when presented graphically. However, in recent years, there has been a growing interest in applying statistical tools, such as nonparametric kernel regression, to develop systematic and automated approaches for technical pattern recognition. Wang and Zhang (2012) introduced a comprehensive data-driven algorithm for technical analysis that incorporated nonparametric local linear estimators. This algorithm aimed to enhance the objectivity and efficiency of technical analysis by leveraging statistical techniques for pattern identification and decision-making in financial markets.

1.3.1 Technical analysis with smoothing splines for bitcoin prices

My interest in technical analysis extends to the Bitcoin market, which is the world's first decentralized digital payment method. Bitcoin has gained recognition and acceptance as a decentralized digital currency and a store of financial value. However, the cryptocurrency market differs from traditional markets due to its global and 24/7 trading nature. Unlike traditional markets, traders are not limited to specific timeframes, and individuals and businesses can engage in global trading at any time. As a result, trading algorithms have become increasingly popular, with a growing share of trading being conducted by trading robots. Miller, Yang, Sun, and Zhang (2019) introduced smoothing splines to analyze the Bitcoin market, specifically aiming to test the effectiveness and profitability of certain technical analysis patterns in this relatively new market.

We collected data from the Global Digital Assets Exchange (GDAX), specifically focusing on Bitcoin prices throughout 2018, with price observations recorded every minute. A Python program was developed to gather the data, which were then stored in a Structured Query Language (SQL) local server using PHPMyAdmin. The data were subsequently imported into R for computation and simulation purposes. Six different technical analysis patterns, such as Moving-Up-Stream and triangle bottoms, were considered in the study. Smoothing splines were employed to reduce noise in the price movements, facilitating the identification of patterns on the smoothed curve.

For example, to identify the triangle bottom pattern, a window of data from every 35 minutes was fitted using smoothing splines. The smoothed data provided local minimum and maximum closing price values denoted by E_1, E_2, E_3, E_4, and E_5, which were necessary for pattern identification. E_5 is the closest extreme point to the end of the time subinterval being fitted. If E_1 represented a minimum and satisfied the conditions $E_1 < E_3 < E_5$ and $E_2 > E_4$, the pattern was recognized as a triangle bottom, as illustrated in Figure 2. This pattern would then generate a buy signal. We also proposed a method to evaluate

the effectiveness of the strategic trading algorithm. We found that using smoothing splines to identify technical analysis patterns, combined with strategies based on these patterns, yielded returns that significantly exceeded the results of unconditional trading strategies.

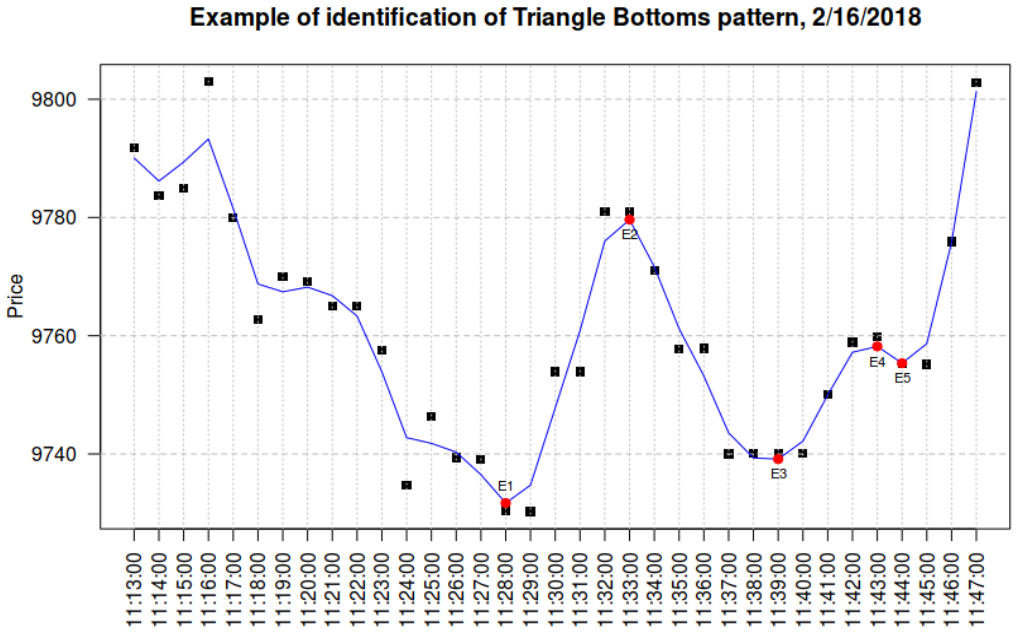

Figure 2: Price pattern: Triangle Bottom

In summary, the study applied smoothing splines to analyze the Bitcoin market, specifically examining the effectiveness of various technical analysis patterns. The findings suggested that using smoothing splines for pattern identification and implementing strategies based on these patterns resulted in returns that outperformed unconditional trading strategies.

1.4 Learning multiple tasks with kernel methods

Spline smoothing is indeed a viable approach for single-task (single-sample) kernel learning. Multitask learning (MTL) is an interesting extension of the kernel approach, aiming to leverage information from multiple related tasks to improve the learning process. In MTL, there are n tasks/samples, each with a different size $m_1, m_2, \ldots, m_n$, and corresponding functions $f_1, f_2, \ldots, f_n$. The objective of MTL is to estimate $f_1, f_2, \ldots, f_n$ simultaneously by minimizing:

$$\frac{1}{\sum_{j=1}^{n} m_j} \sum_{j=1}^{n} \sum_{i=1}^{m_j} L(y_{ji}, f_j(\mathbf{x}_{ji})) + \gamma J(\cdot), \tag{2}$$

Here, y_{ji} represents the ith observation within the jth sample. The term $J(\cdot)$ represents a penalty term that encourages a balance between individual task functions and their average.

The standard single-task kernel methods, such as support vector machines and regularization networks, have been extended to MTL using linear kernels. However, not all data sets can be explained well by linear hyperplanes. To incorporate nonlinearity, Miller and Zhang (2023) developed a model for MTL that incorporates both parametric and nonparametric effects for each task in an additive manner. This approach provides practitioners with flexibility in modeling tasks in a customized manner, leading to increased model performance compared to other modern multi-task methods while maintaining a high degree of model interpretability. The proposed MTL algorithm also incorporates radial basis functions for support vector machine to handle nonlinearity. We leveraged parallel computing techniques to enhance the computational efficiency of our algorithm.

With the rise of big data analysis and deep learning, updating traditional statistical learning approaches becomes crucial. While computational scientists often focus on testing and prediction error, statistical justification remains essential for models and algorithms. In the work of Evgeniou, Micchelli, and Pontil (2005), they applied support vector regression with a multi-task linear kernel to the Inner London Education Authority (ILEA) dataset. However, the explained variance on the test data was only around 34%, and some tasks even yielded negative or zero explained variance when evaluated individually on separate test sets. From a statistical perspective, negative or zero explained variance is unacceptable as it indicates that the suggested model performs worse than a mean model.

To address these issues, Miller and Zhang (2023) proposed new methods for statistical task diagnostics. Our approach allows for the identification and remedy of outlier tasks based on task-specific performance metrics and their empirical distributions. The frameworks we developed were evaluated on the benchmark ILEA dataset and showed significant improvement over other modern multi-task learning methods.

2 Computational and Applied Statistics

I have gained experience working on a diverse range of problems in computational and applied statistics. One of my areas of focus has been the development of parametric bootstrap tests and an R package for conducting heteroscedastic analysis of variance in the case of

unbalanced designs (HeteANOVA). Additionally, I have explored the application of bootstrap and objective Bayes testing methods to address heteroscedastic analysis of variance. These approaches have proven valuable for robust inference in the presence of heteroscedasticity.

I have also delved into general inference problems in computational statistics, including topics such as recursive estimation of time-average variance constants through prewhitening, estimation of correlation coefficients in bivariate log-normal distributions for which the logarithm of the measurements are normally distributed, and estimation of shape parameters in useful family of distributions known as skew normal models.

Another topic I have focused on is developing tests for the median of survival curves, which is particularly useful when dealing with the asymmetry survival data. Comparing survival medians instead of the entire curves is often preferable, especially when there are censored observations. In this regard, we have proposed efficient nonparametric tests specifically tailored for comparing survival medians, with a focus on small sample sizes that are common in pharmaceutical experiments.

Furthermore, I have successfully improved several statistical methodologies. In other work, I have applied the generalized variance function to enhance longitudinal surveys estimators, refined classical Shewhart R and s control charts, and developed random forest regression estimators.

2.1 Simultaneous confidence intervals

Analysis of variance (ANOVA) is a versatile statistical technique that finds applications in various fields such as sociology, education, medicine, psychology, and economics, among others. One illustrative example involves studying different weight loss methods, namely dieting, exercising, and a combination of dieting and exercising. ANOVA is commonly employed to assess the equality of group means through an overall test and to perform pairwise multiple comparisons (PMC) to explore differences between individual group means.

In the presence of groups with unequal variances and unbalanced amounts of data, traditional overall and PMC tests are ineffective. Zhang (2015a, 2015b) introduced computational PMC algorithms for one-way and two-way HeteANOVA problems. Building on this work, Alver and Zhang (2021a, 2021b) as well as Zhang (2021) developed Parametric Bootstrap (PB) based solutions for multiway HeteANOVA problems. We now discuss this work in more detail.

2.1.1 Parametric bootstrap tests for HeteANOVA

We present a three-way Heteroscedastic Analysis of Variance (HeteANOVA) model that considers unequal population variances. The model involves observations $Y_{ijk1}, Y_{ijk2}, \ldots, Y_{ijkn_{ijk}}$ from group ijk with a size of n_{ijk}. Here, $i = 1, \ldots, a$, $j = 1, \ldots, b$, and $k = 1, \ldots, c$. The population variances are denoted by σ^2_{ijk}. The full ANOVA model is expressed as:

$$Y_{ijkm} = G + A_i + B_j + C_k + AB_{ij} + AC_{ik} + BC_{jk} + ABC_{ijk} + \epsilon_{ijkm},$$

In this model, G represents the grand mean, A, B, and C represent the main factor effects, AB, AC, and BC represent the two-way interaction terms, and ABC represents the three-way interaction term. The subscript $m = 1, \ldots, n_{ijk}$ identifies a specific observation within the group, and $\epsilon_{ijkm} \overset{iid}{\sim} N(0, \sigma^2_{ijk})$ represents the error term.

In the presence of unequal variances and unbalanced data, traditional overall and pairwise multiple comparison (PMC) tests are ineffective in HeteANOVA. Alver and Zhang (2021a, 2021b) as well as Zhang (2021) developed a PB-based solution for multiway HeteANOVA problems. This solution comprises an overall test and a PMC procedure, incorporating six PB algorithms with one algorithm illustrated Heteroscedastic. A visual representation of these algorithms can be found in Figure 3.

The main idea of our approach is to use the parametric bootstrap method to simulate the distribution of the test statistics, which involves the standardized sum of squares. For instance, when testing the three-way interaction term, the null hypothesis (H_{0ABC}) assumes that $ABC_{ijk} = 0$ for all $i = 1, ..., a$, $j = 1, ..., b$, $k = 1, ...c$, whereas the alternative hypothesis ($H_{\alpha ABC}$) suggests that $ABC_{ijk} \neq 0$ for some i, j, k.

Let Σ be a diagonal matrix with entries σ^2_{111}/n_{111}, σ^2_{112}/n_{112}, ..., σ^2_{abc}/n_{abc}, where n_{ijk} denotes the sample size for each combination. Define J_n as a column vector of n ones, I_n as an $n \times n$ identity matrix, and X_{ABC} as the design matrix formed by combining the following components: J_{abc}, $I_a \otimes J_{bc}$, $J_a \otimes (I_b \otimes J_c)$, $J_a \otimes (J_b \otimes I_c)$, $I_{ab} \otimes J_c$, $I_a \otimes (J_b \otimes I_c)$, and $J_a \otimes I_{bc}$, where $\otimes$ represents the Kronecker product. Let $\bar{Y}$ be a vector of the group means obtained from different factor levels, i.e., $\bar{Y} = (\bar{Y}_{111}, \bar{Y}_{112}, ..., \bar{Y}_{abc})'$. Additionally, let S be the diagonal matrix of the group sample variances, $S = \text{diag}(s^2_{111}/n_{111}, s^2_{112}/n_{112}, ..., s^2_{abc}/n_{abc})$.

Under the null hypothesis, a suitable test statistic for H_{0ABC} is the standardized sum of squares for the three-way interaction. This statistic follows a chi-square distribution with degrees of freedom equal to $(a-1)(b-1)(c-1)$.

$$\bar{Y}^T\Sigma^{-1}\bar{Y} - \bar{Y}^T\Sigma^{-1}X_{ABC}(X^T_{ABC}\Sigma^{-1}X_{ABC})^- X^T_{ABC}\Sigma^{-1}\bar{Y} \sim \chi^2_{(a-1)(b-1)(c-1)} \tag{3}$$

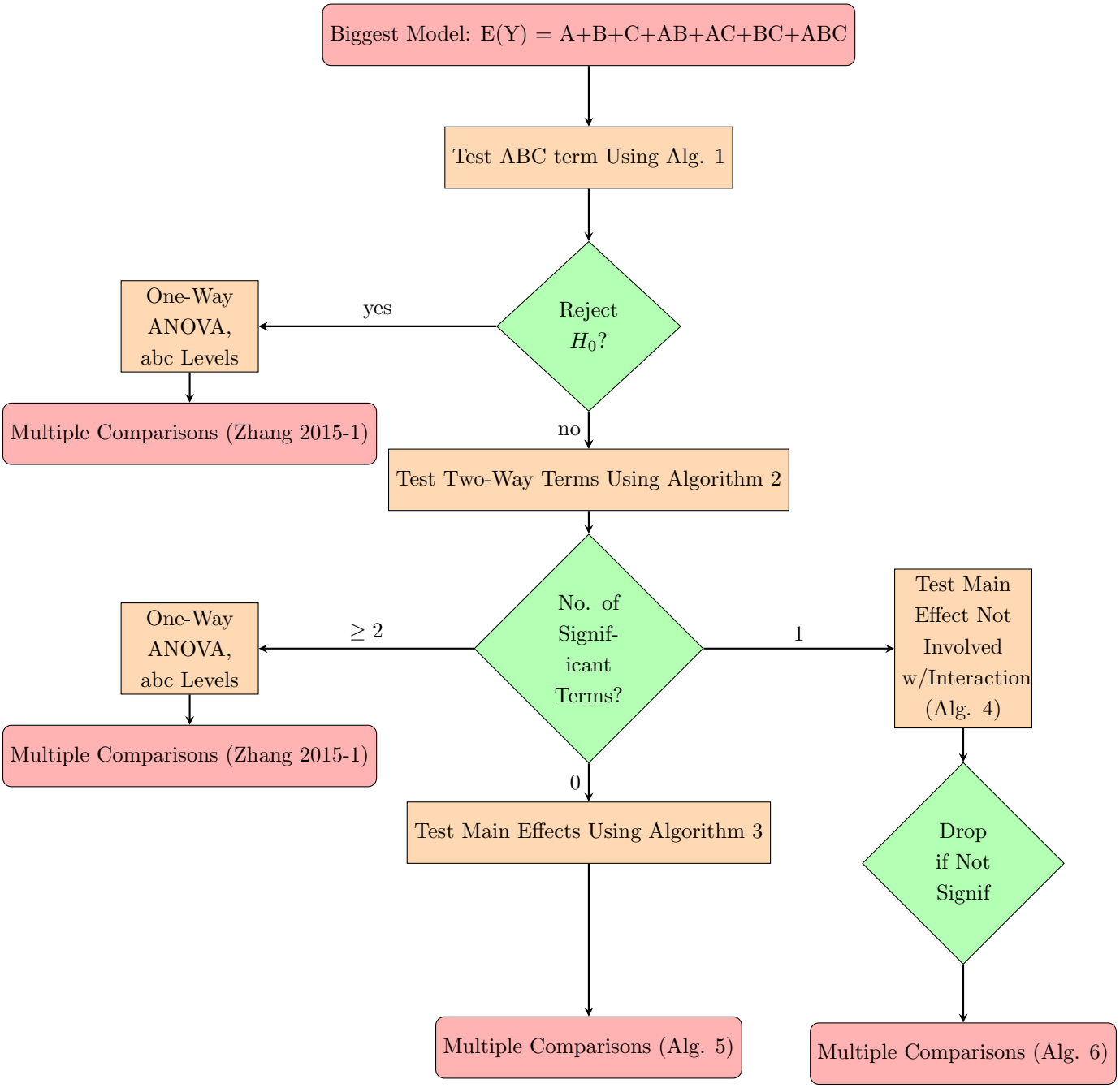

Figure 3: Flowchart: Three-Way ANOVA Testing Using Parametric Bootstrap

The test statistic can be estimated by:

$$\tilde{S}_I = \bar{Y}^T S^{-1} \bar{Y} - \bar{Y}^T S^{-1} X_{ABC} (X_{ABC}^T S^{-1} X_{ABC})^- X_{ABC}^T S^{-1} \bar{Y}. \tag{4}$$

To develop the parametric bootstrap (PB) variable, we assume without loss of generality that $E(\mathbf{Y}) = 0$. For a given set of $(\bar{y}_{111}, \bar{y}_{112}, ..., \bar{y}_{abc}; s^2_{111}, s^2_{112}, ..., s^2_{abc}))$, we can model $\bar{y}_{Bijk}$ as independent normal random variables with mean 0 and variance s^2_{ijk}/n_{ijk}, while S^2_{Bijk} follows a scaled chi-square distribution with $n_{ijk}-1$ degrees of freedom, i.e., $S^2_{Bijk} \sim \left(\frac{s^2_{ijk}}{n_{ijk}-1}\right)\chi^2_{n_{ijk}-1}$, $i = 1, ...a$, $j = 1, ..., b$, $k = 1, ...c$. Let $\bar{Y}_B$ denote the vector $(\bar{y}_{B111}, \bar{y}_{B112}, ..., \bar{y}_{Babc})^T$, and S_B be the diagonal matrix $\text{diag}(s^2_{B111}/n_{111}, s^2_{B112}/n_{112}, ...s^2_{Babc}/n_{abc})$.

We construct the PB pivot variable by replacing $\bar{Y}$ with $\bar{Y}_B$ and S with S_B in the test statistic (4). This results in the following expression for the PB pivot variable:

$$\tilde{S}_{IB} = \bar{Y}_B^T S_B^{-1} \bar{Y}_B - \bar{Y}_B^T S_B^{-1} X_{ABC} (X_{ABC}^T S_B^{-1} X_{ABC})^- X_{ABC}^T S_B^{-1} \bar{Y}_B. \tag{5}$$

To assess the significance of the test, we compare the observed test statistic $\tilde{s}_I$ with the random distribution of the PB pivot variable $\tilde{S}_{IB}$ which is evaluated by simulations. If $P(\tilde{S}_{IB} > \tilde{s}_I) < \alpha$ for a given significance level α, we reject the null hypothesis H_{0ABC}. The probability $P(\tilde{S}_{IB} > \tilde{s}_I)$ can be estimated using Algorithm 1, as described below:

Algorithm 1:

For a given set of $(n_{111}, n_{112}, ..., n_{abc})$, $(\bar{y}_{111}, \bar{y}_{112}, ..., \bar{y}_{abc})$, and $(s^2_{111}, s^2_{112}, ...s^2_{abc})$, compute the observed test statistic $\tilde{s}_I$ using equation (4) and the sample data.

For $k = 1, ..., m$:

Generate $\bar{y}_{Bijk} \sim N(0, s^2_{ijk}/n_{ijk})$ and $S^2_{Bijk} \sim \left(\frac{s^2_{ijk}}{n_{ijk}-1}\right)\chi^2_{n_{ijk}-1}$, where $i = 1, ...a$, $j = 1, ..., b$, $k = 1, ...c$.

Compute $\tilde{S}_{IB}$ using equation (5).

If $\tilde{S}_{IB} > \tilde{s}_I$, set $Q_k = 1$.

Compute the Monte Carlo estimate of the p-value, $\frac{1}{m}\sum_{k=1}^m Q_k$. Other algorithms can be derived similarly.

2.1.2 R package "pbanova"

To facilitate the application of the parametric bootstrap (PB) methods proposed by Alver and Zhang (2021a, 2021b) and Zhang (2021) for HeteANOVA analysis, we have developed an open-source supplementary R package called "pbanova". This package provides a user-friendly interface, allowing researchers to easily use the PB techniques in their analyses. With the "pbanova" package, researchers can apply the PB methods to address HeteANOVA

problems effectively. Additionally, the methods available in the "pbanova" package can also be used to analyze classical ANOVA problems with equal variance assumptions. By enhancing the accessibility and practicality of PB methods, the "pbanova" package serves as a valuable tool for researchers dealing with HeteANOVA problems and contributes to the advancement of statistical analysis in this field.

2.1.3 Relationship between PB and objective Bayesian (OB) approaches

Zhang, Christensen, and Pesko (2021) investigated the relationship between the parametric bootstrap (PB) and objective Bayesian (OB) approaches for testing the equality of factor level means in one-way HeteANOVA. The OB approach utilizes Bayes' Theorem to obtain posterior distributions of parameters by combining a non-informative or flat prior distribution with the likelihood function. Objective priors are chosen to have minimal influence on the posterior analysis, allowing the data to speak for themselves as much as possible. Interestingly, although the PB and OB tests may appear different, it has been demonstrated that they are asymptotically equivalent.

Simulation studies conducted for one-way heteroscedastic ANOVA indicate that both the PB and OB tests effectively control the type I error of the overall mean tests and exhibit reasonable statistical power when group sizes are not small. This research serves as a catalyst for further investigations, such as examining the robustness of PB and OB tests to outliers, exploring the equivalence of PB and OB methods for multi-way ANOVA problems, and studying their performance under special designs like the randomized complete block design and split plot design. These avenues of inquiry aim to extend our understanding of the applicability and properties of PB and OB approaches in various statistical settings.

2.2 Inferences

I have worked on statistical inference for various problems, including the correlation coefficient of the bivariate lognormal distribution, the medians of survival curves, the shape parameter of a skew normal distribution, and the variance of survey data.

2.2.1 Generalized confidence interval and hypothesis tests for the correlation coefficient

The Pearson product-moment correlation is a widely used measure to assess the linear relationship between two continuous random variables. However, financial data often exhibit skewness, requiring the use of skewed distributions such as the log-normal distribution. When

studying two financial measures like silver and gold returns, the log of a bivariate normal distribution is suitable. A question of interest is determining the correlation between silver and gold. Additionally, in bear markets where both gold and silver prices decline, it becomes crucial to investigate whether the correlation between the two assets differs from the past. However, inference on the correlation coefficient of a bivariate log-normal distribution poses challenges due to skewness. To address this, Zhang and Chen (2015) developed generalized confidence intervals and hypothesis tests for the correlation coefficient, extending the results to compare two correlations based on independent samples. Simulation studies demonstrate the effectiveness of these methods, even for more general sample scenarios. Our future plans involve extending this work to inference on the correlation coefficient of more general multivariate log-normal distributions (Zhang & Chen, 2021).

2.2.2 A nonparametric test to compare survival medians

Survival data often exhibit skewness, making the median of a survival curve a preferred measure of central tendency over the mean. Comparing survival medians, rather than means, is desirable for data analysis. Several nonparametric methods have been proposed to test for equality of survival medians in the presence of skewness. Chen and Zhang (2016) introduced a nonparametric test for comparing several survival medians. Our proposed test statistic measures the deviation from the survival median time to a weighted average of all survival median times. Under the null hypothesis of equal medians for K survival curves ($K \geq 2$), it was proven that the test statistic follows an asymptotic chi-square distribution with $K-1$ degrees of freedom. One limitation of existing methods is the inflation of the type I error rate due to the underestimation of variance using the standard approach based on Greenwood's formula. Consider the variance formula $V(Y) = E[V(Y|X)] + V[E(Y|X)]$. The Greenwood estimate is an approximation of the first term $E[V(Y|X)]$, while ignoring the second term $V[E(Y|X)]$.

In our research, we address the issue of underestimation of variance in existing methods by employing the Greenwood estimate for the first term of the variance, and approximate the second term by bootstrap methods. The bootstrap approximation offers an asymptotic approximation, where the estimated variance has an order of n_i^{-2}, with n_i representing the sample size associated with the ith estimated survival curve. Consequently, for large samples, the second term of the variance $V(Y)$ converges to 0, and the estimator of $V(y)$ converges to that of Greenwood's approach. This observation sheds light on why the tests in the literature perform well for large samples but exhibit inflated type I errors for small samples (Chen & Zhang, 2016). Simulation studies demonstrate that the proposed test effectively controls the

type I error rate, even for small samples—a promising feature given the often limited sample sizes in pharmaceutical and other experiments.

2.2.3 Estimators of the shape parameter for a scalar skew normal model

Skew normal distributions have gained significant attention due to their appealing mathematical and statistical properties, as well as their flexibility in fitting data. These distributions find wide applications in various fields, including selective sampling, stochastic frontier models, compositional data, and financial markets, where the assumption of normality is inadequate. The skew normal distribution, denoted as $Y \sim SN(\mu, \sigma, \lambda)$, has a density function

$$f(y; \lambda, \mu, \sigma) = \frac{2}{\sigma}\phi(\frac{y-\mu}{\sigma})\Phi(\lambda\frac{y-\mu}{\sigma}), \tag{6}$$

where ϕ and Φ represent respectively the density and cumulative distribution functions of a standard normal distribution with mean 0 and standard deviation 1. The parameters μ, σ, and λ control the location, scale, and shape of the distribution, respectively. It is worth noting that the skew normal distribution reduces to the normal distribution when the shape parameter λ is set to 0. The shape parameter has a somewhat complicated relationship to the skewness of the distribution.

One challenge associated with the skew normal model is the difficulty in estimating the shape parameter. In cases where the sample size is moderate or small, the maximum likelihood estimator of the shape parameter can be infinite. Moreover, existing estimators for the shape parameter in the literature suffer from significant bias even for moderate sample sizes. To address this issue, Zhang and Liu (2017) proposed three estimators for the shape parameter in a scalar skew normal model. The first two estimators employed a bias correction approach, while the third estimator was derived by solving a modified score equation. Simulation studies demonstrated that the proposed estimators exhibit smaller bias compared to the existing estimators in the literature, particularly for small and moderate sample sizes. This research provides more reliable and accurate estimation methods for the shape parameter in skew normal distributions.

2.2.4 Longitudinal generalized variance functions for survey data

In the field of survey research, my interest has expanded to include an extension of generalized variance functions (GVFs) to longitudinal data analysis. Generalized variance functions are widely used to generate convenient estimates of variances for large-scale surveys, such as the Census Bureau's Current Population Survey and the Canadian Labour Force Survey.

Traditionally, GVFs in survey research are constructed by pooling data from a specific time period, typically one year, under the assumption of a constant population throughout that period. However, this approach overlooks the fact that the size of the population may change over time, leading to substantial variation in the standard errors of the estimators. Moreover, with access to longitudinal data, it is possible to use multiple years of data to estimate the variance of the estimators.

To overcome these limitations, Zhang, Cheng, and Lu (2019) introduced the concept of longitudinal generalized variance functions (LGVFs), which incorporate a time effect into the modeling framework. The LGVFs account for the fluctuations in population size over time and use a larger amount of data for estimation, resulting in more accurate variance estimates for longitudinal data. The study investigates the asymptotic properties of the estimators, which involve linear combinations of cluster means obtained from stratified two-stage cluster samples.

The implementation of these LGVF methods to the Current Population Survey demonstrates their effectiveness in producing proper standard errors for longitudinal data. By considering the changing population size over time, and using a larger amount of data, the LGVFs significantly improve the accuracy of variance estimation, thereby enhancing the reliability and validity of survey research findings in longitudinal settings.

2.3 Methodology improvement

2.3.1 Recursive estimation of time-average variance constants through prewhitening

The time-average variance constant (TAVC) plays a significant role in various time series inference problems, such as unit root testing and statistical inference of the mean. Consider a stationary process $(X_i), i \in Z$, with a mean $\mu = E(X_i)$ and finite variance, and let $\gamma(k) = cov(X_0, X_k), k \in Z$, be the covariance function. An estimate of μ can be obtained by taking the sample mean $\bar{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Under suitable conditions, as n increases, $\bar{X_n}$ converges in distribution to a normal distribution:

$$n^{1/2}(\bar{X_n} - \mu) = n^{-1/2}\sum_{i=1}^{n}(X_i - \mu) \longrightarrow N(0, \sigma^2),$$

where σ^2 represents the TAVC, long-run variance, or asymptotic variance parameter. In a stationary process, the TAVC is the sum of all covariances (and is a multiple of the spectral density at 0). In many applications, it is a good idea to update the estimate of σ^2 sequentially

as new observations are accumulated. Wu (2009) proposed an efficient recursive algorithm to compute the TAVC, resulting in a memory complexity of $O(1)$ and computational complexity scaling linearly with n.

Prewhitening is a technique used to identify a filter that transforms X_i into residuals or white noise e_i that are serially independent or free of autocorrelation. The TAVC of the prewhitened series $\{e_i\}, i \in Z$, is denoted as σ_e^2. The estimation of σ_e^2 can be more accurate than that of σ^2. Zheng, Jin, and Zhang (2016) applied an AR(1) prewhitening filter to construct a recursive estimate of the TAVC. By using the AR(1) prewhitening filter, we achieved substantial improvements in efficiency under certain circumstances. The paper provides theoretical conditions for deciding whether prewhitening is necessary or not. The memory complexity of $O(1)$ is maintained, and the accuracy of the estimate is enhanced, as supported by both theoretical results and simulation studies.

2.3.2 Exploring Market Dynamics in Cryptocurrency and Quality Control Techniques in Industrial Settings

In our research titled "Reversion and Location Trends in the Bitcoin Market" (Lewinski, Yang, Chen, & Zhang, 2019), we explore the dynamics of the cryptocurrency market, particularly Bitcoin, and investigate whether certain phenomena observed in traditional stock markets also exist in cryptocurrency markets. Specifically, we examine the application of the 75 percent reversion rule in cryptocurrency markets.

Using local linear regression, we find that active market trading, as indicated by peaks in the fitted regression curve, tend to occur around the opening or closing times of traditional markets. Additionally, we observe that the stance of governments on cryptocurrencies and news events can impact specific regions and potentially have a ripple effect on other regions. These insights can assist investors and institutions in understanding the cryptocurrency market and making informed decisions based on the timing of market activity in different geographic locations.

In another aspect of my research, I focus on variability charts commonly used in industrial quality control, such as the R control chart (R chart) and the s control chart (s chart), initially introduced by Shewhart in 1931 (Shewhart, 1931). Zhang (2014b) proposed improved versions of these charts that provide more accurate approximation of control limits. This improvement is achieved by incorporating cumulative distribution functions of the sample range and standard deviation. These enhanced control charts offer better performance in monitoring and maintaining the quality of industrial processes.

2.3.3 Mixed model and extensions

Mixed models, involving models with some effects being random rather than fixed, have been at the forefront of my research interests, driving my exploration into their various aspects. In my initial investigation (Lu & Zhang, 2010), I focused on the problem of testing null hypotheses concerning the variance component in a balanced one-way random effects model. Building upon this foundation, my ongoing research aims to expand these findings to encompass unbalanced designs, addressing a broader range of practical scenarios. I'm pleased to share that my work has garnered attention from fellow researchers, indicating its significance and potential impact. Notably, one of the Ph.D. students in our department has further extended this research to Generalized Split-plot design models, broadening its applicability. Additionally, researchers from the field of biostatistics have expressed keen interest. I am excited to continue this line of inquiry and collaborate with fellow scholars to advance my understanding in this area.

3 Applications

Besides developing statistical theory, methodology and algorithms, I am also interested in applying statistical theory and methods to solve health care and financial engineering problems. This work can be seen in Zhang et al. (2011) , X. Liu, Li, Zhang, Sheng, and Xu (2007) and Zhang (2012b).

3.1 Application of Statistical Theory and Methods in Health Care

In the field of health care, my research focuses on applying statistical theory and methods to understand and solve problems related to human diseases. Specifically, I have been involved in studies investigating genetic mutations in human endometrial cancer and the characterization of serum low-molecular weight proteins/peptides in liver injury patients.

3.1.1 Genetic Mutations in Human Endometrial Cancer

With the advancements in genome sequencing technology, my research aimed to unravel the genetic basis of human endometrial cancer. Through transcriptome sequencing on an endometrial tumor paired with normal cervical tissue, Zhang et al. (2011) developed a phylogenetic approach to characterize individual genetic mutations in cancer cell proliferation within a single resected patient tumor. By assuming that all cells have the same mutations

from their parents and using the highest frequency mutation to split the cluster, the method facilitated the identification of mutations belonging to each group. The resulting tree model provided a clear display of the order of genetic mutations, making it more interpretable and meaningful compared to classical tree models. Using this approach, we successfully identified five ubiquitous mutations presumed to occur in the cancer founder cell of the tumor, and may collectively play critical roles in endometrial oncogenesis. Further testing in 10 additional endometrial tumors failed to show overlapping mutations in the cancer founder cells, indicating the lack of a single common oncogenic pathway for these endometrial tumors. The effects of individual mutations in cancer cell proliferation were calculated based on descendant cell number and time span since acquiring each mutation.

3.1.2 Characterization of Serum Low-Molecular Weight Proteins/Peptides in Liver Injury Patients

In collaboration with other researchers, we conducted a study that aimed to characterize serum low-molecular weight proteins/peptides associated with liver injury using factor analysis with surface-enhanced laser desorption ionization-time of flight-mass spectrometry (SELDI-TOF MS) data (X. Liu et al., 2007). The SELDI-TOF MS data posed challenges due to its high dimensionality and the need to address correlations among peaks. By comparing mass spectra of the hepatitis group with those from the control group, we identified 43 peaks associated with liver function impairment. Through factor analysis, we extracted four common factors (cholestasis, coagulation, attenuation, and 9292) that provided insights into the underlying processes related to liver injury. Notably, the factors related to coagulation disorders (coagulation and 9292) in liver injury shed light on the role of serum low-molecular weight proteins/peptides in coagulation processes. Additionally, we proposed plausible interpretations for some undetermined peaks, contributing to the understanding of liver injury mechanisms.

3.2 Application of Statistical Theory and Methods in Financial Engineering

In addition to health care applications, my research extends to the field of financial engineering. Specifically, I have conducted investigations into the optimal geometric mean returns of stocks and options and their relationship to the Kelly criterion (Zhang, 2012b).

In this study, I proved that the optimal geometric mean returns of a stock and its corresponding option are the same according to the Kelly criterion. This result was demonstrated

using both the binomial option pricing model and continuous stochastic models with a self-financing assumption. Through simulation studies, the approximate validity of this relationship was established for the continuous option pricing model. Furthermore, for the discrete case, I showed that the ratio of the optimal fractions of a stock and its option is unrelated to the probability distribution of the return. This implies that, we can use a small amount of options to replace the underlying asset without changing the optimal geometric mean return and knowing the probability distribution of the return. Hence, in practice, either there are sure win chances or the price of options are more expensive than their theoretical values. Otherwise, one should always hold more uncorrelated options instead of stocks.

By conducting research in both health care and financial engineering domains, I aim to contribute to the advancement of statistical theory and methods while addressing real-world challenges in these fields. These applications demonstrate the practical relevance and potential impact of statistical approaches in understanding diseases, identifying biomarkers, and making informed financial decisions.

4 Ongoing research

In the following sections, I will provide an overview of my current work, which includes collaborative research with the National Center for Health Statistics (NCHS), as well as my future research directions in survey sampling and missing data imputation.

4.1 A general procedure for evaluating models and ensemble support vector regression

In addition to the classical goodness-of-fit (GOF) and lack of fit tests for model assessment, a commonly used measurement for evaluating model performance is the mean squared test set error or prediction error (PE). This approach involves randomly dividing the data into two parts: a training set and a test set. The training set is used to build models, while the test set is used to evaluate the model's performance based on the PE. However, from a statistical perspective, relying on the PE from a single test set may introduce bias, and using PE as the sole measure of model fit without considering other diagnostics may be insufficient.

To overcome this limitation, Zhang and He (2022(a)) introduces a comprehensive framework and a general procedure for assessing the performance of various regression methods in the context of big data. The key concept behind this procedure involves randomly partitioning the data into multiple training and test sets. By applying the regression method to each

training set and calculating the PE on corresponding test sets, an average PE is obtained by averaging the PEs from different test sets. This approach provides a robust and reliable measure of the performance of different regression methods, allowing for a comprehensive comparison and evaluation of their effectiveness in handling big data.

4.1.1 Algorithm for the general procedure

Suppose we are interested in comparing r different methods or models. The proposed Monte Carlo simulation algorithm for conducting these comparisons is as follows:

Algorithm 1: General Monte Carlo simulation procedure

```
Randomly create K splits and save the splits to K different files
for each split do
    read the split from the data file;
    for each method (1,2,..., r) do
        fit the model with the training data;
        use the fitted model to do prediction on the testing data;
        compute the measures of prediction performance;
    record the results to the output file;
Use statistical tools to analyze the results in the output file;
```

Algorithm 1 can be carried out by parallel computing.

4.1.2 Extended Paired t-test

We used the binomial test and paired t-test in the general procedure. In the realm of machine learning, comparing two methods based on prediction errors (PEs) by averaging them across multiple test sets is a common practice. However, we hold a different perspective on this matter. We propose extending the paired t-test to compare methods based on the same test set, followed by estimating the average difference and testing whether the value 0 falls within the confidence interval. To demonstrate this extended paired t-test, we outline the procedure as follows: let $PE_i^{(1)}$ and $PE_i^{(2)}$ denote the prediction errors obtained from model 1 and model 2, respectively, for the ith test set. We define the PE improvement as

$$\delta_i = PE_i^{(1)} - PE_i^{(2)}, \tag{7}$$

which serves as a measure for the ith split. We reject the null hypothesis H_0 if the confidence interval for the true PE improvement $D = E(\delta_i)$ is positive. This approach allows us to more precisely assess the performance difference between the two methods and draw conclusions based on the inclusion or exclusion of 0 within the confidence interval.

4.1.3 Ensemble Support Vector Regression (ESVR)

Support vector regression (SVR) is a regression method based on the principles of support vector machines (SVM) (Vapnik, 1995). It addresses regression problems by using an ϵ-insensitive loss function, which penalizes data points that deviate from a specified margin ϵ. SVR incorporates concepts from the theory of reproducing kernel Hilbert space and shares similarities with other regularization approaches such as ridge regression, cubic smoothing splines, and thin plate splines. In the field of machine learning, ensemble methods such as bagging (Breiman, 1996), stacking, and boosting have gained prominence. These methods enhance predictive performance by combining predictions from multiple models.

Motivated by the principles of SVR and ensemble methods, we have proposed two modified support vector regression models for regression analysis. Here, we will focus on illustrating one of these models: the ensemble support vector regression (ESVR). The ESVR model combines multiple reduced models in E_m (a collection of subsets) to achieve improved predictive accuracy. Additionally, it helps reduce the spread or dispersion of predictions. The procedure for h-step ESVR is outlined below:

Algorithm 2: h-step ensemble SVR procedure (ESVR)

```
for each m do
    for each subset in E_m do
        fit SVR with the training data;
        use 10 fold cross-validation to find optimal parameters;
        use the model with optimal parameters to do prediction on the testing data;
        record the result;
average all the results.
```

By averaging the results obtained from different subsets and models, the h-step ensemble ESVR provides a more robust and accurate prediction. The procedure, as described in Algorithm 2, allows for the selection of optimal SVR models within each subset and accounts for the variability across different subsets.

4.2 The general information criterion (GIC)

My research has made notable progress in variable selection for regression analysis by introducing the general information criterion (Zhang & Pleis, 2022(b)). Various criteria, such as mean squared error, adjusted R^2, Akaike information criterion (AIC), Bayesian information criterion (BIC), and others, are commonly used for model selection. However, none of these criteria directly measure the model's predictive power. To address this gap, our research introduces a general information criterion (GIC) that directly quantifies the predictive power

of models.

Both AIC and BIC add penalty for the number of parameters in the criterion. The larger the penalty, the less the number of predictors will be selected. Unfortunately, the penalty from AIC is too small and the penalty from BIC is too big. Therefore, we propose a GIC to adjust the penalty term so that AIC and BIC provide the lower and upper bounds, i.e., AIC and BIC are retrieved when $\lambda = 2$ and $\lambda = \log(n)$ respectively,

$$GIC = n * \log\frac{SSE}{n} + n + n * \log(2 * \pi) + \lambda * h, \tag{8}$$

where h is the number of predictors, and $\lambda \in [2, \log(n)]$.

To determine the optimal value of the parameter λ in a data-driven manner, a grid of λ values can be set up, such as $\lambda_1 = 2, \lambda_2, \lambda_3, \ldots, \lambda_r = \log(n)$. The selection of λ is based on minimizing the prediction error, which can be achieved through the proposed Algorithm 1, i.e., a Monte Carlo simulation algorithm using multiple splits of the data into training and test sets.

In addition, from Algorithm 1 GIC employs inclusion frequencies as a measure of variable importance. These frequencies indicate the proportion of times a predictor is selected among the K models generated from the multiple splits. By analyzing the inclusion frequencies, variables with high frequencies can be identified as important contributors.

To facilitate the implementation of GIC variable selection, we have developed algorithms tailored for this purpose. These algorithms enable efficient and flexible variable selection using the GIC procedure. To demonstrate the effectiveness of GIC, we have applied it to two real-world examples, illustrating its ability to accurately and reliably identify relevant predictors when compared to traditional methods such as AIC and BIC.

4.3 Semi-parametric models for small area estimation (SAE) using machine learning methods

4.3.1 Area level SAE using support vector regression

In the domain of small area estimation, which involves sub-populations or domains with limited sample sizes for obtaining reliable estimates, my collaborative research with NCHS has yielded significant advancements.

The Fay-Herriot model for SAE assumes a linear linking function, which may not always be suitable when dealing with nonlinear relationships. To address this limitation, Zhang and Pleis (2022(c)) proposed a semi-parametric model for small area estimation (SP-SAE), which allows for a more general approach by estimating the nonparametric component using

support vector regression (SVR). Consider m areas $i = 1, \cdots, m$. The SP-SAE model can be expressed as follows:

Sampling model: $\bar{y}_i = \bar{Y}_i + e_i$, where $e_i \sim N(0, \psi_i)$ and ψ_i is known.

Linking model: $\bar{Y}_i = \mathbf{z}'_{ai}\boldsymbol{\beta} + f(\mathbf{z}_{bi}) + v_i$, where $\mathbf{z}_{ai}$ and $\mathbf{z}_{bi}$ are disjoint subsets of $\mathbf{z}_i$, $f(\mathbf{z}_{bi})$ is a smooth function, and $v_i \sim N(0, \sigma_v^2)$.

Combined model: $\bar{y}_i = \mathbf{z}'_{ai}\boldsymbol{\beta} + f(\mathbf{z}_{bi}) + v_i + e_i$.

The estimation of parameters in the SP-SAE model is accomplished using the back-fitting (BF) algorithm. This iterative algorithm involves fitting the linear and nonlinear components of the model repeatedly until convergence is achieved. To assessment the performance of linear and nonlinear models and emphasize the necessity of a nonparametric model like SP-SAE, we employed a procedure that incorporates multiple random splits, as well as the binomial test and paired t-test methods proposed by Zhang and He (2022(a)). This procedure allows for a comprehensive comparison between the two types of models.

We also conducted an empirical analysis using ACS 2015 health coverage data, comparing the performance of the direct estimator, linear Fay-Herriot model, and the nonparametric SP-SAE model. The results showed that the prediction errors of the linear model were significantly larger than that of the SP-SAE. Moreover, the estimate of σ_v^2 in the Fay-Herriot model was larger than that in the SP-SAE model, indicating that adding a nonparametric component reduces the between-area variation, allowing small areas to borrow strength from other areas more effectively. Comparisons of coefficients of variation demonstrated that SP-SAE outperformed Fay-Herriot, while Fay-Herriot outperformed direct estimators. It is worth noting that SVR was used in this research to estimate the regression function, but other machine learning methods could also be employed within this framework.

4.3.2 Area and Unit level SAE using random forests

Random forests (RF), introduced by Breiman (2001), have emerged as a powerful nonparametric machine learning tool and have shown competitiveness across various data mining techniques. RF is particularly effective in handling regression problems when the number of observations is smaller than the number of predictors, which is often encountered in small area estimation. Additionally, RF offers the advantage of addressing prediction tasks even in the presence of missing values in the predictor variables.

In addition to area-level SAE, my ongoing collaborative research with the NCHS aims to examine unit-level SAE. For this purpose, we propose the use of random forests (RF) and its bias-corrected variant, known as bias-corrected random forests (BCRF) (Zhang & Lu, 2012), for both area and unit-level estimation.

The unit-level model considers each unit j within area i individually, instead of using summarized totals or means. Let y_{ij} represent unit j in area i, where $i = 1, 2, \cdots, m$ and $j = 1, \cdots, n_i$. The total number of units is given by $n = n_1 + n_2 + \cdots + n_m$. The response vector is denoted as $\mathbf{y} = (y_{11}, y_{12}, \cdots, y_{1n_1}, y_{21}, y_{22}, \cdots, y_{2n_2}, \cdots, y_{m1}, y_{m2}, \cdots, y_{mn_m})'$, and the corresponding error term is $\boldsymbol{\epsilon} = (\epsilon_{11}, \cdots, \epsilon_{1n_1}, \cdots, \epsilon_{m1}, \cdots, \epsilon_{mn_m})'$. Let I_m be the $m \times m$ identity matrix, $\mathbf{1}_{n_i}$ be a vector of length n_i consisting of ones, $\mathbf{f} = (f_1(\mathbf{z}), f_2(\mathbf{z}), \cdots, f_m(\mathbf{z}))'$, and $\mathbf{Z} = I_m \otimes (\mathbf{1}'_{n_1}, \mathbf{1}'_{n_2}, \cdots, \mathbf{1}'_{n_m})'$, where $\otimes$ represents the Kronecker product.

We fit the following model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{f} + \boldsymbol{\epsilon}, \tag{9}$$

where $\mathbf{X}$ is a design matrix related to the linear term structure, and the error term $\boldsymbol{\epsilon}$ has a mean of zero and a block diagonal variance matrix for each area. We plan to tackle the problem by employing the backfitting algorithm, which enables the estimation of parameters in the proposed small area model.

One advantage of unit-level SAE is that the vector $\mathbf{f} = (f_1(\mathbf{z}), f_2(\mathbf{z}), \cdots, f_m(\mathbf{z}))'$ is estimated using all the unit-level information simultaneously. Each area has its own area-specific estimator $\hat{f}_i(\mathbf{z})$. However, it is possible that model (9) may not adequately explain some areas when evaluated individually on separate test sets. Therefore, we propose developing appropriate lack-of-fit tests to identify outlying areas and exploring possible remedial measures.

4.4 Partially linear model for dual frame surveys

I am working on a semiparametric model for dual frame surveys, which employ a parametric component to account for the domain effect and a nonparametric component to capture the underlying regression function. The model is defined as follows:

$$y_{ij} = \beta_i + m(\mathbf{x}_{ij}) + \epsilon_{ij}, \quad i = a, ab, b, \tag{10}$$

Here, a, ab, and b represent the three non-overlapping domains within the dual frame survey. The vector $\mathbf{x}_{ij} = (x_{ij1}, \ldots x_{ijp})^T$ contains p covariates corresponding to y_{ij}. The unknown parameter β_i is associated with domain i, while the unknown function $m(\cdot) \in W_2^m[0, 1]$ represents the underlying nonparametric regression function. The error terms ϵ_{ij} are zero-mean random variables with a variance-covariance matrix of Σ.

In this model, we assume that the underlying nonparametric regression function $m(x_{ij})$ is the same across the three domains, while the domain effects β_i can differ. This model can also be employed to test for the presence of a domain effect, i.e., to assess the hypothesis H_0 :

$\beta_a = \beta_{ab} = \beta_b$. However, there are a few challenges we anticipate when adopting a partially linear model for complex surveys with dual frame designs. These challenges include the bias introduced by using adjusted weights in the overlap domain and the boundary problem associated with each domain. To examine the asymptotic properties of the estimators, we intend to employ a combined inference framework.

4.5 Exploring Estimation Methods for Missing Data Handling in Survey Nonresponse

In addition to small area estimation, my research plan includes investigating techniques for handling missing data, specifically focusing on missing data imputation and weighting adjustment. Missing data is a common issue in survey practice, and it can lead to biased and flawed inferences if not addressed appropriately. One common approach to handle missing data is to impute the missing values using auxiliary variables, while weighting adjustment is suggested when unit missing occurs.

4.5.1 Imputation by machine learning methods

Several issues need to be considered in missing data imputation research, including survey design issues, missing response from auxiliary variables, and the evaluation of imputation methods' effectiveness. Based on initial studies, I propose three estimators for survey nonresponse imputation when auxiliary variables are available. These estimators include the Support Vector Regression (SVR) estimator ($\hat{f}_S$), the bias-corrected Random Forest (RF) estimator ($\hat{f}_{RC} = \hat{f}_R - \hat{B}(X, Y)$), and a combined estimator that incorporates both RF and SVR ($\hat{f}_\theta(\mathbf{x}) = \theta\hat{f}_R + (1-\theta)\hat{f}_S$).

To determine the optimal value of θ, we set up a grid of values between 0 and 1. For each value of θ, we calculate the mean squared PE based on the combined estimator $\hat{f}_\theta(\mathbf{x})$. The goal is to find the value of θ that minimizes the mean squared PE, indicating the best trade-off between the RF and SVR estimators.

The research topics related to these estimators involve investigating the effects of sampling design and weights, comparing them to existing methods in the literature, and developing a diagnostic procedure to evaluate the imputation methods, such as assessing prediction error.

4.5.2 Weighting adjustment methods

One challenge in practice is that auxiliary variables are usually available only for sampled respondents. In such cases, weighting adjustment is suggested. Deville and Sarndal (1992)

derived a weighting system using a distance measure and calibration equations, and Slud and Thibaudeau (2009, 2010) adapted the generalized-raking calibration methodology for nonresponse adjustment. They also proposed linearization-based large-sample variance formulas. However, their weighting adjustment method relies on exact constraints, which I plan to modify. My plan is to use machine learning methods to optimize the weights, either through high-dimensional parametric models or nonparametric nonlinear models, thus relaxing the exact constraints and enhancing the weighting adjustment process. This aspect of the research is currently under investigation.

References

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*, 123-140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Chen, Z., & Zhang, G. (2016). Comparing survival curves based on medians. *BMC Medical Research Methodology*, 16–33.

Deville, J.-C., & Sarndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, *87*(418), 376–382.

Evgeniou, T., Micchelli, C., & Pontil, M. (2005). Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, *6*, 615-637.

Lewinski, D., Yang, Y., Chen, Z., & Zhang, G. (2019). Reversion and location trends in the bitcoin market. *International Journal of Data Science*, *4*, 275 - 287.

Liu, Härdle, W. K., & Zhang, G. (2017). Statistical inference for generalized additive partially linear models. *Journal of Multivariate Analysis*, *162*, 1–15.

Liu, X., Li, L., Zhang, G., Sheng, G., & Xu, W. (2007). An approach to characterize serum low molecular weight proteins/peptides in liver injury with SELDI-TOF MS and factor analysis. *Clinical Biochemistry*, *40*, 1266-1271.

Lu, Y., Fu, Y., & Zhang, G. (2021). Nonparametric regression estimators in dual frame surveys. *Communications in Statistics - Simulation and Computation*, *50*, 854-864.

Lu, Y., & Zhang, G. (2010). The equivalence between likelihood ratio test and F test for testing variance component in a balanced one way random effects model. *Journal of Statistical Computation and Simulation*, *80*, 443-450.

Lu, Y., Zhou, L., Zhang, G., & Christensen, R. (2021). Neyman smooth type goodness-of-fit tests in complex surveys. *Technical Report*.

Miller, N., Yang, Y., Sun, B., & Zhang, G. (2019). Identification of technical analysis patterns with smoothing splines for bitcoin prices. *Journal of Applied Statistics*, *46*,

2289-2297.

Miller, N., & Zhang, G. (2023). Additive multi-task learning models and task diagnostics. *Communications in Statistics - Simulation and Computation*.

Shewhart, W. A. (1931). *Economic control of quality of manufacturing processes*. Milwaukee: John Wiley & Sons.

Skinner, C., & Rao, J. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, *91*, 349-356.

Slud, E., & Thibaudeau, Y. (2009). Simultaneous calibration and nonresponse adjustment. In *Proceedings of jsm2009, section on survey research methods, washington, dc* (p. 2263-2272).

Slud, E., & Thibaudeau, Y. (2010). Simultaneous calibration and nonresponse adjustment [Computer software manual]. (Research report, U.S. Census Bureau)

Vapnik, V. (1995). *The nature of statistical learning theory*. Berlin, Heidelberg: Springer-Verlag.

Wang, S., & Zhang, G. (2012). Price pattern recognition utilizing local polynomial regression. *Journal of Trading*, *7*, 37-43.

Wu, W. (2009). Recursive estimation of time-average variance constants. , *19*, 1529-1552.

Zhang, G. (2011). *Smoothing splines using compactly supported positive definite radial basis functions*. New York: ProQuest, UMI Dissertation Publishing.

Zhang, G. (2012a). Smoothing splines using compactly supported positive definite radial basis functions. *Computational Statistics*, *27*, 573-584.

Zhang, G. (2012b). Optimal geometric mean returns of stocks and their options. *International Journal of Stochastic Analysis*, *2012*.

Zhang, G. (2014b). Improved R and s control charts for monitoring the process variance. *Journal of Applied Statistics*, *41*, 1260-1273.

Zhang, G. (2015a). A parametric bootstrap approach for one-way ANOVA under unequal variances with unbalanced data. *Communications in Statistics-Simulation and Computation*, *44*, 827–832.

Zhang, G. (2015b). Simultaneous confidence intervals for pairwise multiple comparisons in a two-way unbalanced design with unequal variances. *Journal of Statistical Computation and Simulation*, *85*, 2727–2735.

Zhang, G., Beck, B., Luo, W., Wu, F., Kingsmore, S. F., & Dai, D. (2011). Development of a phylogenetic tree model to investigate the role of genetic mutations in endometrial tumors. *Oncology Report*, *25*, 1447-1454.

Zhang, G., & Chen, Z. (2015). Inferences on correlation coefficients of bivariate log-normal

distributions. *Journal of Applied Statistics*, *42*, 603–613.

Zhang, G., & Chen, Z. (2021). Inferences on correlation coefficients of multivariate log-normal distributions. *Manuscript under preparation*.

Zhang, G., Christensen, F., & Zheng, W. (2015). Nonparametric regression estimators in complex surveys. *Journal of Statistical Computation and Simulation*, *85*, 1026–1034.

Zhang, G., Christensen, R., & Pesko, J. (2021). Parametric bootstrap and objective bayesian testing for heteroscedastic one-way anova. *Statistics and Probability Letters*, *174*.

Zhang, G., & He, Y. (2022(a)). *A general procedure for evaluating models and ensemble support vector regression* (Technical Paper). Hyattsville, Maryland: National Center for Heath Statistics.

Zhang, G., & Liu, R. (2017). Bias-corrected estimators for scalar skew normal. *Communications in Statistics - Simulation and Computation*, *46*, 831–839.

Zhang, G., & Lu, Y. (2008). Adjusted confidence bands in nonparametric regression. *Communications in Statistics-Simulation and Computation*, *37*, 106-113.

Zhang, G., & Lu, Y. (2012). Bias-corrected random forests in regression. *Journal of Applied Statistics*, *39*, 151-160.

Zhang, G., Mao, G., & Cheng, Y. (2016). Adjusted confidence band for complex survey data. *Communications in Statistics-Simulation and Computation*, *45*, 1896–1904.

Zhang, G., & Pleis, J. (2022(b)). *Variable selection procedure in regression for large data* (Technical Paper). Hyattsville, Maryland: National Center for Heath Statistics.

Zhang, G., & Pleis, J. (2022(c)). *A semi-parametric model for small area estimation using support vector machine* (Technical Paper). Hyattsville, Maryland: National Center for Heath Statistics.

Zheng, W., Jin, Y., & Zhang, G. (2016). Recursive estimation of time-average variance constants through prewhitening. *Statistics and Probability Letters*, *114*, 30–37.