RESEARCH STATEMENT

My research focuses on developing theories and methodologies for drawing statistical inference based on large and sophisticated data. In addition to my Ph.D. research on the empirical Bayes inference under the supervision of Professor J. T. Gene Hwang, in the last decade, my research interests have expanded to more important and fundamental statistical problems including (i) statistical inference in the presence of unknown yet unequal variances; (ii) model free estimation and hypothesis testing in sufficient dimension reduction; (iii) multiple hypothesis testing for grouped hypothesis and high dimensional regression models; and (iv) bias mitigation on dependent data adaptively collected using multi-armed bandit.

1 Statistical Inference Assuming Unequal and Unknown Variances

While the statistical inference on the mean structures of populations has been extensively studied in literature, inference regarding variances remains largely on inspecting the model assumptions for traditional ANOVA. Statistical inference on general hypotheses on the variance components is challenging and urgently needed. For example, heterogeneity commonly exists in a vast range of applications, such as in the microarray experiments where the variations across a large number of transcripts usually vary widely. As noted by George E. P. Box Box (1953), "*test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port!*" To meet the immediate challenge of drawing simultaneous inference on many variance components, we have proposed the following methods with outstanding performances.

I. Parametric empirical Bayes estimator of variances. In a series of work (Zhao, 2010; Hwang et al., 2009; Zhao and Hwang, 2012; Hwang and Zhao, 2013), we have considered the shrinkage estimator of the variances and its impact on the inference of the mean parameters. In all these works, we assumed that the sample variances s_i^2 follows a chi-squared distribution, i.e., $s_i^2|\sigma_i^2 \sim \sigma_i^2 \frac{\chi_d^2}{d}$. By putting a log-normal prior distribution on the variances, we developed a parametric empirical Bayes estimator of the variances which could substantially improve the performance of inferential procedure for (selected) means.

II. Non-Parametric emprical Bayes estimator of variances. In a recent work with one of my PhD students (Kwon and Zhao, 2022), we imposed an <u>*arbitrary prior*</u> on σ_i^2 as $\sigma_i^2 \sim g(\sigma_i^2)$ and derived a Bayes estimator of the variance as

$$\widehat{\sigma}^2_{i,B} = \frac{k}{2}\left\{\frac{\int_{s_i^2}^{\infty} (s^2)^{-(\frac{k}{2}-2)}\, dF(s^2)}{\int_{s_i^2}^{\infty} (s^2)^{-(\frac{k}{2}-1)}\, dF(s^2)} - s_i^2\right\}. \tag{1}$$

A unique feature of this estimator is that it only depends on $F(s^2)$, the cumulative distribution function of the sample variances. We call this "F-modeling" based estimator in contrast to the "f-

modeling" based estimator that relies on the data via its marginal probability density function. The data-driven version has been easily established by replacing the cumulative distribution function with its empirical counterpart.

III. Robust variance estimator. All the aforementioned approaches have proven to be useful in the presence of heterogeneity. However, a major limitation of these methods is the assumption that the sample variance follows a chi-squared distribution, implying that the parent distribution is Gaussian. It has been known that tests on the variance are more sensitive to the violation of Gaussianity (Box, 1953), in Bar and Zhao (2022) we tackle this by *removing the normality assumption* on the parent distribution and introduce a bias correction term for $\log(s^2)$ that depends on the excess kurtosis of the parent distribution. We have shown that the resulting inference method is robust when the data generation process deviates from the normal distribution profoundly.

2 Sufficient Dimension Reduction

In the general framework of regression analysis, a primary goal is to infer the relation between the response variable Y and a p-predictor $\boldsymbol{X}$. Specifically, one is interested in $Y|\boldsymbol{X}$, namely, the conditional distribution of Y on $\boldsymbol{X}$. Among the literature on sufficient dimension reduction (SDR), one seeks for a minimal subspace $\mathcal{S}_{Y|\boldsymbol{X}}$ of the column space of $\boldsymbol{X}$, called the central subspace, such that

$$Y \perp\!\!\!\perp \boldsymbol{X} | \mathbf{P}_{\mathcal{S}} \boldsymbol{X}. \tag{2}$$

That is, Y is independent from the projection operator $\mathbf{P}_{\mathcal{S}}\boldsymbol{X}$. A prevalent method to estimate the central space $\mathcal{S}$ is the sliced inverse regression (SIR), proposed in Li (1991). From a series of work, we have filled the theoretical gap of SIR under the high dimensional setting.

I. Phase transition. In Lin et al. (2018), we filled the theoretical vacancy on the sufficient dimension reduction literature by showing that the SIR is consistent if and only if the ratio $\rho = \frac{p}{n} \to 0$. This implies that the phase transition of SIR is the same as that of the PCA. When ρ does not converge to zero, one must impose sparsity condition to achieve the estimation consistency. This result provides a theoretical justification of many regularized methods in the sufficient dimension reduction.

When p is of the same or a higher order of n, we introduced a Lasso regression method based on the SIR to obtain an estimate of the SDR space (Lin et al., 2021). The resulting algorithm, Lasso-SIR, is shown to be consistent and achieves the optimal convergence rate under certain sparsity conditions when p is of order $o(n^2\lambda^2)$, where λ is the generalized signal-to-noise ratio.

II. Detection boundary of single index model. In Lin et al. (2021), we developed the detection boundary of the single index model $y = f(\boldsymbol{\beta X}, \epsilon)$. Note that the norm of the parameter vector, a commonly used quantity for measure the strength of the signal, is not identifiable and no longer applicable. We then defined the *generalized signal-to-noise ratio (gSNR)* λ as the unique non-zero eigenvalue of $\mathrm{var}[(\mathbb{E}(\boldsymbol{x}|y)]$ and developed the detection boundary of the single index model in terms of λ. This is the first result to study the detection boundary for the single index model.

III. Limiting distribution of SIR. Regardless of the large amount of investigations on the consistency of SIR, results on the limiting distribution of SIR are largely missing when p diverges in n. All the existing distributional results either assume a fixed p or a fixed number of signals (Wu and Li, 2011). In an on-going project (Zhao and Xing, 2022), based on the newly developed theory of Gaussian approximation for high dimension, we have derived the limiting distribution of the SIR by allowing the dimension p diverges to ∞. This paves the way for providing further statistical inference for the SIR under high dimensionalities.

3 Multiple Hypothesis Testing

Multiple hypotheses testing is the key in modern sciences when the number of hypotheses of interest is very large. A tremendous upsurge of research has taken place in this area in the last three decades. Since I joined Temple University, I have been working on this area and developed a series of important methods to address various fundamental issues.

I. Group hypothesis. In an sole-authored paper (Zhao, 2022), I have theoretically investigated the pros and cons of the p-value based testing method and the local false discovery rate (FDR) based testing method. In He, Sarkar, and Zhao (2015), we have studied optimal multiple testing procedure incorporating a severity function reflecting unequal penalty of type II errors. When hypotheses admit grouping structures, we proposed an optimal method in Liu et al. (2016) controlling both the overall and the within-group FDR. In Sarkar and Zhao (2017), we continues the line of research initiated in Liu et al. (2016) on developing a novel framework for multiple testing of hypotheses grouped in a one-way classified form using the hypothesis-specific local FDRs, which effectively captures the dependence structure due to the grouping.

II. Multiple testing for linear regression models. Many widely-accepted FDR controlling methods, such as the Benjamini-Hochberg (BH) method, start with independent and valid p-values. There are two challenges when applying these techniques to the high-dimensional linear regression model: (i) the p-values are difficult to obtain, and (ii) the p-values are usually dependent. In Ji and Zhao (2014), we studied the rate optimal multiple testing procedure from the perspective of variable selection for high-dimensional regression models with the signals so weak and rare that the *"selection consistency"* is not possible. This is the first result on the rate optimality of testing procedures for high-dimensional regression models. In Xing et al. (2021), we proposed the Gaussian Mirror method, which creates for each predictor variable a pair of mirror variables by adding and subtracting a randomly generated Gaussian perturbation. The mirror variables naturally lead to test statistics which are symmetric with respect to zero under the null. This symmetry property allows us to estimate the false discovery proportion and easily control the FDR.

III. Model-free multiple testing. The Gaussian mirror we introduced in Xing et al. (2021) is powerful in detecting the signals for the high-dimensional linear regresion models. In an on-going project (Zhao and Xing, 2022), we further studied the model-free multiple testing problem. We consider the general multiple index model

$$Y = f(\boldsymbol{\beta}_1 \boldsymbol{X}, \cdots, \boldsymbol{\beta}_D \boldsymbol{X}, \epsilon),$$

where the link function $f(\cdot)$ is unknown. Let $\mathcal{S} = span(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_D)$ be the central space. For each covariate, we want to test whether this covariate plays any role in the central subspace $\mathcal{S}$. In this paper, we estimated β_i's using SIR without knowing the link function and constructed the mirror statistic. Based on the limiting distribution of SIR, these statistics are symmetric with respect to zero under the null. We then proposed a method, named Model-free Multiple Testing using Mirror Statistics (MMM). It is shown that MMM controls the FDR at the desired level well and is more powerful than its competitors. We are in the final stage of writing this paper.

V. Non-parametric testing. Along with the century-long advancement of data analysis across a tremendous variety of scientific fields, the regression model is the most powerful and the most widely used tool to reveal scientific laws. For the regression model, three fundamental statistical assumptions serve as the cornerstones yet have been often adopted incautiously: (A1) the population of random noises admits normality or similar distributional assumptions; (A2) the random noises are identically and independently distributed; and (A3) the exogeneity such that the covariates are not dependent on the random noises. Though statistical tests developed specifically for one of the three assumptions scattered in the literature, a unified powerful validation procedure on all three assumptions remains a long-lasting missing component in statistics and related data-driven research.

In Zhang et al. (2021), we studied nonparametric dependence detection with the proposed **B**inary **E**xpansion **A**pproximation of **U**niformi**TY** (BEAUTY) approach, which generalizes the celebrated Euler's formula, and approximates the characteristic function of any copula with a linear combination of expectations of binary interactions from marginal binary expansions. This novel theory enables a unification of many important tests through approximations from some quadratic forms of symmetry statistics, where the deterministic weight matrix characterizes the power properties of each test. To achieve a robust power, we study test statistics with data-adaptive weights, referred to as the **B**inary **E**xpansion **A**daptive **S**ymmetry **T**est (BEAST). It is shown that the BEAST is powerful in testing of the uniformity and testing of dependence which can be used to test the assumptions (A1) and (A3). In an on-going project, we are using the BEAUTY framework to construct a non-parametric testing method to detect whether a set of observations are independent (A2). The preliminary results are very promising.

4 Statistical Inference for Adaptively Collected and Dependent Data

In both statistical research and broad data-driven applications, the random sampling is an important technique ensures that results from the sample approximate what would have been obtained if the entire population had been measured. Many inferential methods are legitimate when we have a random sample. However, when the data are collected adaptively and are *dependent* by nature, even the sample mean becomes a biased estimator. We investigated statistical issues for the multi-armed bandit problem (MAB).

The MAB can be seen as a set K-arms and each arm is associated with a distribution P_k. The customer iteratively plays one lever per round (t) by selecting one arm based on the historical data, denoted as I_t, and generating an observation Y_t from the distribution associated with this arm. Let $(Y_t, I_t), t = 1, 2, \cdots, T$ be the data generated from the MAB algorithm.

I. Pre-data collection bias mitigation. Let μ_k be the mean of P_k and the sample mean

$$\widehat{\mu}_k = \frac{\sum_{t=1}^{n} Y_t \mathbf{1}(I_t = k)}{N_k(n)}$$

be the estimator of μ_k where $N_k(n) = \sum_{t=1}^{n} \mathbf{1}(I_t = k)$ is the number of times that the k-th arm is pulled. Because of the genuine dependence among the observations due to the sequential allocation, $\widehat{\mu}_k$ is known to be biased (Nie et al., 2018; Neel and Roth, 2018). In Wang et al. (2022), we provided a new MAB algorithm, randomized multi-arm bandits, which combines a randomization step with any chosen MAB algorithm. This randomziation step weaken the serial dependence and can mitigate the bias substantially without affecting its regret asymptotically.

II. Post-data collection bias mitigation. With data collected by a certain MAB algorithm, we considered the statistical inference on parameters associated with the distributions P_k's (Wang and Zhao, 2022). We provided an explicit formula to understand the source of bias. For example, the bias of the sample mean $\widehat{\mu}_k$ admits

$$\mathbb{E}\left[\widehat{\mu}_k\right] - \mu_k = -\frac{\mathrm{Cov}(\widehat{\mu}_k, N_k(n))}{\mathbb{E}[N_k(n)]}.$$

For a given data, we use the resampling method to approximate the bias term. It is shown that the proposed method could substantially mitigate the bias and lead to methods with good inferential properties. We have further developed the bias formulas and the resampling methods for more general cases including the context bandits and other estimators such as the sample variance and empirical distribution functions. This bias correction could help with the inference, which in turn could improve the adative design of the experiments. It has the potential to be further applied in the reinforcement learning.

5 Summary

My research focused on developing methods and theories to address some fundamental issues such as heterogeneity and dependence. My research after tenure promotion has resulted in publications in the prestigious journals, including, two papers on Journal of American Statistician Association, one paper on Annals of Statistics, one paper on Biometrika, one single-author paper on TEST, and four others. I have four papers under different stages of the reviewing process and four papers to be submitted.

I have received the grant (IIS-1633283) from the National Science Foundation. Additionally, I have served as an associate editor of the ASA journal "Statistical Analysis and Data Mining" from 2013. I have served as the Editorial board of reviewers for the "Journal of Machine learning research" from 2020. I have served in the NSF panel. I have been invited to peer-review manuscripts for top journals such as JASA, JRSSB, Biometrika and Annals of Statistics. I have been invited to present my research for multiple conferences and department colloquiums. In the future, I will endeavor to maintain my productivity in high quality research and collaborative works.

References

Bar, H. and Z. Zhao (2022). Robust varince estimation. Technical report.

Box, G. E. P. (1953). Non-normality and tests on variances. Biometrika 40, 318–335.

He, L., S. K. Sarkar, and Z. Zhao (2015). Capturing the severity of type II errors in high-dimensional multiple testing. Journal of Multivariate Analysis 142, 106–116.

Hwang, J. T., J. Qiu, and Z. Zhao (2009). Empirical Bayes confidence intervals shrinking both means and variances. Journal of the Royal Statistical Society. Series B 71(1), 265–285.

Hwang, J. T. and Z. Zhao (2013). Empirical Bayes confidence intervals for selected parameters in high dimension with application to microarray data analysis. Journal of the American Statistical Association 108(502), 607–618.

Ji, P. and Z. Zhao (2014). Rate optimal multiple testing procedure in high-dimensional regression. Submitted.

Kwon, Y. and Z. Zhao (2022). On F-modeling based empirical bayes estimation of variances. Biometrika.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. Journal of the American Statistical Association 86(414), 316–327.

Lin, Q., Z. Zhao, and J. S. Liu (2018). On consistency and sparsity for sliced inverse regression in high dimensions. The Annals of Statistics 46(2), 580 – 610.

Lin, Q., Z. Zhao, and J. S. Liu (2021). Testing model utility for single index models under high dimension. Festschrift in Honor of R. Dennis Cook: Fifty Years of Contribution to Statistical Science, 65.

Liu, Y., S. K. Sarkar, and Z. Zhao (2016). A new approach to multiple testing of grouped hypotheses. Journal of Statistical Planning and Inference 179, 1–14.

Neel, S. and A. Roth (2018). Mitigating Bias in Adaptive Data Gathering via Differential Privacy. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright, pp. 1–10.

Nie, X., J. Taylor, and J. Zou (2018). Why Adaptively Collected Data Have Negative Bias and How to Correct for It. In Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain, Volume 84.

Sarkar, S. K. and Z. Zhao (2017). Local false discovery rate based methods for multiple testing of one-way classified hypotheses. arXiv preprint arXiv:1712.05014.

Wang, T., B. Ji, and Z. Zhao (2022). Bias, regret and statistical inference in adaptive data collection. Submitted.

Wang, T. and Z. Zhao (2022). Resampling-based bias adjustment for adaptively collected data. Technical report.

Wu, Y. and L. Li (2011). Asymptotic properties of sufficient dimension reduction with a diverging number of predictors. Statistica Sinica 2011(21), 707.

Xing, X., Z. Zhao, and J. S. Liu (2021). Controlling false discovery rate using gaussian mirrors. Journal of the American Statistical Association, 1–20.

Zhang, K., Z. Zhao, and W. Zhou (2021). Beauty powered beast. arXiv preprint arXiv:2103.00674.

Zhao, Z. (2010). Double shrinkage empirical Bayesian estimation for unknown and unequal variances. Statistics and Its Interface 3, 533–541.

Zhao, Z. (2022). Where to find needles in a haystack? TEST 31(1), 148–174.

Zhao, Z. and J. T. Hwang (2012). Empirical Bayes false coverate rate controlling confidence interval. Journal of the Royal Statistical Society. Series B 74(5), 871–891.

Zhao, Z. and X. Xing (2022). Model-free multiple testing using mirror statistics (MMM). Manuscript in preparation.