Research Statement

Zhenhao Gong

I have been interested in solving practical problems using economic models and theories since my undergraduate study. During my Ph.D. study at UConn, I focus my research on factor analysis in high-dimensional datasets and large panel models with interactive fixed effects. My research interests include developing and applying new tools for economists with the use of big data, machine learning, and forecasting.

Current Work

My dissertation "Three Essays on Large Panel Econometrics" focuses on the large panel data models and factor analysis with high-dimensional datasets. The first essay proposes an improved inference procedure for the panel data model with interactive fixed effects in the presence of cross-sectional dependence and heteroskedasticity. The second essay extends the proposed inference procedure to the dynamic panel data model with interactive fixed effects. The third essay studies the robustness of the current methods for choosing the number of strong and weak factors in high-dimensional datasets.

My job market paper, "Improved Inference for Interactive Fixed Effects Model with Cross-sectional Dependence", proposes an improved inference procedure for the panel data model with interactive fixed effects in the presence of cross-sectional dependence and heteroskedasticity. It is well known in the literature that the least squares (LS) estimator in this model by Bai (2009) is asymptotically biased when the error term is cross-sectionally dependent, and I address this problem. My procedure involves two parts: correcting the asymptotic bias of the LS estimator and employing the cross-sectional dependence robust covariance matrix estimator. I prove the validity of the proposed procedure in the asymptotic sense. Since my approach is based on the spatial HAC estimation, e.g., Conley (1999), Kelejian and Prucha (2007) and Kim and Sun (2011), I need a distance measure that characterizes the dependence structure. Such a distance may not be available in practice and I address this by considering a data-driven distance that does not rely on prior information. I also develop a bandwidth selection procedure based on a cluster wild bootstrap method. Monte Carlo simulations show my procedure work well in finite samples. As empirical illustrations, I apply the proposed method to make inference on the effects of divorce law reforms on the U.S. divorce rate, and the effects of clean water and sewerage interventions on the U.S. child mortality.

The second chapter of my dissertation extents the inference procedure I proposed in my job market paper to the dynamic panel data model with interactive fixed effects. As Moon and Weidner (2015) shown, there are two sources of asymptotic biases of the LS estimator in the dynamic panel data model with interactive fixed effects. The first type of bias is the same bias as Bai (2009), which is caused by the correlations and heteroskedasticities in the idiosyncratic errors. The other type of bias arises from the predetermined regressors, which is analogous to the incidental parameter bias of Nickell (1981) in finite T. The bias correction procedure proposed

by Moon and Weidner (2015) only consider heteroskedasticities, assuming no correlations in the idiosyncratic errors. Thus, their estimators are not valid when the idiosyncratic errors are correlated in both dimensions. The bias caused by the time-series correlated errors and the predetermined regressors can be estimated by the truncated kernel method of Newey and West (1987). In the presence of cross-sectional correlation and heteroskedasticity, we can apply the proposed procedure to improve the inference of the LS estimator.

The last chapter of my dissertation documents the non-robustness issue for estimating the number of factors in high dimensional data. In my first two chapters, I assume the number of factors is known but we need to estimate it in practice. There are strong and weak factors in the factor model based on the imposed assumptions and researchers are interested in estimating the total number of them. Most methods in the literature for choosing the number of strong and weak factors are based on the results from random matrix theory (RMT), which studies the distribution of sample eigenvalues and requires i.i.d. and gaussian assumption on the error terms in the factor model. These restrictions may not appropriate when we want to apply those methods in practice. My paper shows that those methods are not robust by simulation when the error terms in the factor model are serially or/and cross-sectionally correlated or have non-gaussian distributions. My simulation results provide useful recommendations to applied users for how to choose the estimation method in dealing with different types of high-dimensional datasets.

Future Plans

In the future, I'm planning to develop a test that can be used to detect the cross-sectional correlation in the error terms. This is important because if the errors are not cross-sectional correlated, then the LS estimator will be more efficient and we don't need to apply the proposed inference procedure. The other future project I considered is to provide an alternative bias correction procedure based on bootstrap method. The advantage of this method is that we don't need to know the specific structure of the cross-sectional correlation bias to correct it.