UCONN

Zhenhao Gong, PhD
Department of Economics
University of Connecticut
Phone: +1 (315) 870-8326
Email: zhenhao.gong@uconn.edu

Research Statement

My research focuses on large panel models and factor analysis in high-dimensional datasets. My research interests include developing and applying new tools for economists using big data, machine learning, and forecasting.

Graduate Research

My dissertation "Three Essays on Large Panel Econometrics" studies the large panel models with interactive fixed effects and factor analysis with high-dimensional datasets. In my first essay, I propose an improved inference procedure for the interactive fixed effects (IFE) model with cross-sectional dependence and heteroskedasticity. The IFE model includes the standard fixed effects model as a special case but is significantly more flexible to control the unobserved time-varying heterogeneity. My second essay extends the proposed inference procedure in my first essay to the dynamic panel data models with interactive fixed effects. The third essay studies the robustness of the existing methods for choosing the number of strong and weak factors in high-dimensional datasets.

My job market paper proposes an improved inference procedure for the IFE model with cross-sectional dependence and heteroskedasticity. It is well known in the literature that the least square (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 developed a bandwidth selection procedure based on a cluster wild bootstrap method. Monte Carlo simulations show my procedure work well in finite samples.

My procedure can be applied to the broad empirical literature in economics. I illustrate this with two empirical examples. The first one is the well-known problem of the U.S. divorce rates affected by divorce law reforms around the 1970s. Using the standard fixed-effects model with weighted least squares (WLS) estimation, Wolfers (2006) identified the rise of divorce rates in the first eight years after the law reform. However, the robustness of Wolfers (2006)'s results is doubted in two regards. First, the model he uses may not be flexible enough to capture the factors varying over time and across states (e.g., the stigma of divorce; religious belief). This may lead to the observed large discrepancy between the ordinary least squares (OLS) and WLS estimates found by Lee and Solon (2011). Second, the idiosyncratic errors in his model are

assumed to be cross-sectionally independent, which does not seem to be appropriate in practice. Kim and Oka (2013) employed the IFE model for the study. Their results confirm the findings of Wolfers (2006) and are robust to the weighting schemes. However, their bias correction procedure and standard error estimation do not take the cross-sectional dependence into account. I apply the proposed approach and provide inference results for this model. I find the IFE model with the proposed procedure yields smaller estimates with wider confidence intervals than Kim and Oka (2013) 's results.

The second example studies the effects of clean water and effective sewerage systems on U.S. child mortality. An essential question in public health is the cause of the sharp decrease in the U.S. and Massachusetts infant mortality from 1870 to 1930. Alsan and Goldin (2019) exploited the independent and combined effects of clean water and effective sewerage systems on under-5 mortality in Massachusetts, 1880-1920. For empirical strategy, they employ a standard fixed-effects model, which identifies the two interventions together account for approximately one-third of the decline in the log child mortality during the 41 years. Since they use the municipality-level data, the potential unobserved time-varying heterogeneity and cross-sectional correlation in the idiosyncratic errors may affect the results. To check the robustness of their results, I employ the IFE model with the proposed inference procedure for the study. I find that the combined impacts of sewerage and safe water treatments on child mortality are significantly decreased by using the IFE model with the proposed procedure.

My last essay documents the non-robustness issue for estimating the number of factors in high-dimensional data. There are strong and weak factors in the factor model based on the imposed assumptions, and researchers are interested in identifying and 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, 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 be 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 correlated in both dimensions, or have non-gaussian distributions. My simulation results provide helpful recommendations to applied users for choosing the estimation method in dealing with different types of high-dimensional datasets.

Future Plan

My short-term research goals are to continue investigating the theoretical issues in the IFE model and the applications of the IFE model in economics. I plan to develop a test that can be used to detect the cross-sectional dependent errors in the setting of the panel model with interactive fixed effects. It is crucial because if the errors are not cross-sectional correlated, the LS estimator will be more efficient, and we don’t need to apply the proposed inference procedure. The other project I considered is to provide an alternative bias correction procedure based on the 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 estimate and correct it. My long-term research interests are creating and applying new tools for economists to use with big data, machine learning, and forecasting.