Incorporating Prior Knowledge of Latent Group Structure in Panel Data Models

Boyuan Zhang

The assumption of group heterogeneity has become popular in panel data models. Instead of modeling heterogeneity via unit-specific coefficients, the cross-sectional units are assumed to cluster into groups, and within each group units share the same coefficients. This thesis develops an econometric framework to incorporate prior knowledge of groups, which is considered as additional information that does not enter the likelihood function. The prior knowledge aids in clustering units into groups and sharpens the inference of group-specific parameters, particularly when units are not well-separated.

All existing methods for estimating group heterogeneity solve a clustering problem by assuming that units are exchangeable and treating all units equally *a priori*. In a cross-country application of evaluating the impact of climate change on economic growth, countries in different climatic zones are assumed to have equal probabilities of being grouped together. The assumption of exchangeability might not be reasonable since correlations are common between observations at proximal locations and researchers could have knowledge of the underlying group structure based on theories or empirical findings. For instance, Sweden and Finland, which share a border, economic structure, and weather conditions, may have a higher chance of being in the same group than African countries. In such a scenario, it is preferable to use additional information to break the exchangeability between countries to facilitate grouping, as opposed to clustering based solely on observations in the sample.

The whole framework is built on the nonparametric Bayesian method, where I do not impose a restriction on the number of groups, and model selection is not required. The baseline model is a linear panel data model with an unknown group pattern in fixed-effects, slope coefficients, and cross-sectional error variances. I estimate the model using the Dirichlet process (DP) prior, a standard prior in nonparametric Bayesian inference. In this framework, the number of groups is considered as a random variable and is subject to posterior inference. The number of groups and group membership are estimated together with the heterogeneous coefficients. Moreover, since the DP prior implicitly defines a prior distribution on the group partitionings, the posterior analysis takes the uncertainty of the latent group structure into account.

The derivation of the proposed prior starts from summarizing prior knowledge in the form of pairwise constraints, which describe a bilateral relationship between any two units. I consider two types of constraints: *positive-link* and *negative-link* constraints, representing the preference of assigning two units to the same group or distinct groups. Instead of imposing these constraints dogmatically, each constraint is given a level of accuracy that shows how confident the researchers are in their choice. There is a hyperparameter that controls the overall strength of the prior knowledge: a small value partially recovers the exchangeability assumption on units, whereas a large value confines the prior distribution of group partitioning around group structure based on prior knowledge. I choose the optimal value for the hyperparameter by maximizing the marginal data density.

The aforementioned pairwise constraints are used to modify the standard DP prior. In particular, the pairwise constraints are combined with the prior distribution of the group partitioning, shrinking the distribution toward my prior knowledge. I refer to the estimator using the proposed prior as the Bayesian group fixed-effects (BGFE) estimator.

I derive a posterior sampling algorithm for the framework with the modified DP prior. Adopting conjugate priors on group-specific coefficients allows for drawing directly from posteriors using a computationally efficient Gibbs sampler. With the newly proposed prior, it can be shown that, compared to the framework that uses a standard DP prior, all that is needed to implement pairwise constraints is a simple modification to the posterior of the group indices.

The pairwise constraint-based framework is closely related and applicable to other models where group structure plays a role. Although I concentrate primarily on the panel data model, the DP prior with pairwise constraints applies to models without the time dimension, such as the standard clustering problem and the estimation of heterogeneous treatment effects. The framework is also applicable to estimating panel VARs, which involves multiple dependent variables. The group structure is used to overcome overparameterization and overfitting issues by clustering the VAR coefficients into groups, and pairwise constraints add additional information to the highly parameterized model.

I compare the performance of the BGFE estimator to alternative estimators using simulated data. The Monte Carlo simulation demonstrates that the BGFE estimator generates more accurate estimates of the group-specific parameters and the number of groups than the BGFE estimator without including any constraints. The improved performance is mostly attributable to the precise group structure estimation. The BGFE estimator clearly dominates the estimators that omit the group structure by assuming homogeneity or full heterogeneity. I also evaluate the performance of one-step ahead point, set, and density forecasts. Unsurprisingly, the accurate estimates translate into the predictive power of the underlying model; the BGFE estimator outperforms the rest of the estimators.

I apply my method to two empirical applications. An application to forecasting the inflation of the U.S. CPI sub-indices demonstrates that the suggested predictor yields more accurate density predictions. The better forecasting performance is mostly attributable to three key characteristics: the nonparametric Bayesian prior, prior belief on group structure, and grouped cross-sectional heteroskedasticity. In a second application, I revisit the relationship between a country's income and its democratic transition. This question was originally studied by Acemoglu, Johnson, Robinson, and Yared (2008, AER), who demonstrate that the positive income effect on democracy disappears if country fixed effects are introduced into the model. The proposed framework recovers a group structure with a moderate number of groups. Each group has a clear and distinct path to democracy. In addition, I identify heterogeneous income effects on democracy and, contrary to the initial findings, show that a positive income effect persists in some groups of countries, though quantitatively small.