# To pool or not to pool: What is a good strategy?*

Richard Paap
*Econometric Institute, Erasmus University Rotterdam*

Wendun Wang
*Econometric Institute, Erasmus University Rotterdam and Tinbergen Institute*

Xinyu Zhang
*Chinese Academy of Sciences and Capital University of Economics and Business*

This version: March 24, 2015

**Abstract:** This paper considers estimating the heterogenous panel data model from a new perspective. We propose a novel optimal pooling averaging estimator that does not require assumptions on the heterogeneity structure and makes an explicit tradeoff between efficiency gains from pooling and bias due to heterogeneity. By theoretically and numerically compare the mean square error of the proposed estimator with existing approaches, we find that there does not exist a uniformly best estimator and our pooling averaging estimator is superior in the non-extreme cases. To decide which estimator to use in practice, practical guidance is provided depending on different features of data and models. We apply our approach to study the cross-country sovereign credit risk. The results shed new light on the determinants of sovereign credit default swamp spreads.

**Keywords:** Credit default swamp spreads; Heterogenous panel; Mean squared error; Model screening; Pooling averaging;

**JEL Classification:** C23, C52, G15

*We are grateful to participants at seminars at Chinese Academy of Sciences, Renmin University, Xiamen University; to Pavel Ciezek, Herman van Dijk, Jan Magnus, and Guohua Zou for useful discussions and constructive comments. All errors are ours.

# 1 Introduction

With the increasing availability of data, panel data models have been widely used in empirical analysis. When estimating a panel data model, researchers have to tackle the challenge that whether the slope coefficients are allowed to be heterogenous across individual units. Assuming homogenous coefficients is amount to estimating the pooled panel, and therefore the question how to model the heterogeneous coefficients is sometimes poetically referred by econometricians to as "to pool or not to pool". This is long-existing question in the panel data analysis, while there is still no consensus on it. On one hand, an increasing number of studies have noted that the homogenous assumption of coefficients is vulnerable in practice, and the violation of this assumption can lead to misleading estimators. For example, Su and Chen (2013), Durlauf et al. (2001), and Phillips and Sul (2007) provided strong cross-country evidence of heterogeneity, and ample microeconomics evidence can be found in Browning and Carro (2007). On the other hand, plenty of empirical studies find that the pooled estimator (assuming homogenous coefficients) often outperforms the estimators obtained from individual time series estimation in terms of mean square (prediction) error (MSE or MSPE), e.g. Baltagi and Griffin (1997), Baltagi et al. (2000), Hoogstrate et al. (2000), et al. The diversifying empirical results suggests that the pooling decision involves the typical bias-variance tradeoff, and the choice of estimation strategies should depend on situations. More specifically, one should balance the efficiency gains from pooling and the bias due to individual heterogeneity. This brings forth two questions. First, how do we make an appropriate tradeoff between the efficiency and bias when estimating a heterogenous panel? Second, is there a fit-for-all estimator that performs well in all cases, and if not how do we make a choice under different situations? This paper addresses these two questions by introducing a novel pooling averaging procedure that makes an appropriate bias-variance tradeoff and by providing practical guidance on how best to handle the coefficient heterogeneity in different situations.

To estimate a heterogenous panel, a simple way is to estimate each individual time series separately. This can lead to consistent coefficient estimates but not efficient. On the other extreme, one could ignore the heterogeneity and estimate the pooled panel. The pooled estimator is biased but can be more efficient. There are various estimators in between these two extremes. For example, the pooled mean group estimator (Pesaran et al., 1999) imposes homogeneity restrictions on the long-run parameters but allows short-run

parameters to vary over individuals; The grouped fixed-effects estimator (Bonhomme and Manresa, 2012) allows a time-varying group pattern in the fixed effect parameters while assuming the slope coefficients are homogenous. These estimation strategies make good sense, but require correct specification of the heterogeneity structure. Another stream of this literature focuses on testing for the homogeneity assumption. The estimator obtained after a preliminary testing step is called the pretest estimator. This type of estimators should be used and explained with caution, because they are conditional on the selected model and ignore the uncertainty in the testing procedure.

Instead of choosing between the pooled and individual estimators or using a selected model, we propose to combine the estimators from different pooling specifications with appropriate weights. This is our first main contribution. The proposed pooling averaging method can be applied to obtain coefficient estimators and to make predictions, and it includes many popular approaches as special cases, such as the shrinkage estimation, pretesting estimation, et al. It has three main advantages: (1) It does not require specifying the heterogeneity structure, and the coefficient parameters can be heterogenous in any pattern. (2) It makes an explicit tradeoff between efficiency gains from pooling and bias due to individual heterogeneity. (3) It avoids the problems caused by pretesting since it is continuous, unconditional (with model uncertainty already taken into account), and has a bounded risk (see, e.g., Danilov and Magnus (2004)). More throughout review and comparison with the literature will be given in Section 2. We theoretically examine the asymptotic and finite sample properties of the Mallows pooling averaging estimator, a particular pooling averaging estimator. This also complements to the model averaging literature that mainly focuses on the asymptotic property of the Mallows averaging estimator.

The second contribution of this paper is to provide empirical researchers with practical guidance on how best to handle the coefficient heterogeneity of the panel data model under different situations. Given the fact that the performance of estimators differs over applications to a large extent, it is of practical use to have such a general guideline. In particular, each empirical data are characterized by unique features, such as the data generation process, amount of information contained, signal-to-noise ratio, sample size, et al. These features are further reflected by the model when researchers try to explore these data, for example, the model specification and fitness. By theoretically comparing the MSE of different estimators, we show that there does not exist a uniformly best method, and

the performance of estimators depends on the features of data and models, i.e. the degree of coefficient heterogeneity, the error variance, the number of regressors, and the choice of weights. This theoretical finding is supported by an extensive simulation experiments where we compare the proposed method with 12 alternatives. Based on the theoretical and numerical results, we offer general guidance how to choose an appropriate estimator, depending on different features of the data and models. For example, the pooling averaging estimator, especially Mallows pooling averaging, is advantageous when the panel is characterized by moderate degree of heterogeneity (not completely heterogenous) and the signal-to-noise ratio of the model is not large, while the mixed and pooled estimators are often recommended when the model is poorly fitted with rather large errors.

To implement the pooling averaging technique in practice, we introduce two model screening strategies to address the issue of a large model space. The distinctive feature of these screening methods is that they do not need to estimate all candidate models, and thus significantly reduce the computation burden. With a smaller model space, the estimation efficiency is also improved.

We apply the proposed pooling averaging estimator to investigate the cross-country sovereign credit risk. Recently, an increasing number of studies have tried to associate the sovereign credit default swamp (CDS) spreads with various macroeconomic fundamentals, e.g. Dieckmann and Plank (2011); Longstaff et al. (2011); Beirne and Fratzscher (2013), and most of these studies analyze individual countries separately. The individual-country analysis suggests that there seems to be a common pattern in the processes of sovereign CDS spreads across countries. To incorporate this potential common pattern, we employ the pooling averaging method to estimate the effect of potential macroeconomic fundamentals and make predictions. In general we find that the pooling averaging provides more intuitive estimates than the individual estimates typically used in this literature. Our empirical results also provide several new insights. For example, the Mallows pooling averaging estimates reveal that the CDS spreads of most small economies are more affected by the global variables, such as Bulgaria, Malaysia, and Slovak, while those of the large economies are largely influenced by both their own local and global stock market. The pseudo out-of-sample prediction shows that the pooling averaging and pooled estimator generally produce more accurate prediction.

The paper is organized as follow. Section 2 briefly reviews the estimation and testing strategies for heterogenous panels in the literature. Section 3 sets up the model and

proposes the pooling averaging estimator in a general form. Section 4 derives the MSE of the pooling averaging estimator and compares it with the pooled and individual estimators. Section 5 discusses the choice of averaging weights and its theoretical properties. Two model screening procedures are introduced in Section 6. To illustrate our theory, simulation studies and an empirical example are provided in Section 7 and 8, respectively. Finally, based on the theoretical and numerical results, we offer some practical suggestions on how best to handle the heterogeneity in Section 9.

# 2 Brief review of literature

The literature on the heterogenous panel mainly focuses on how to estimate a (possibly) heterogenous panel and how to test the homogeneity assumption. Allowing coefficients to vary across individuals can be dated back to Zellner (1969) and Swamy (1970). Swamy (1970) proposed to estimate the mean of the heterogenous coefficient (average effect) using a generalized least squares (GLS) type estimator. Alternatively, one can average the data over time and estimate cross-section regressions to obtain the average effect estimate, as in most cross-country studies. Pesaran and Smith (1995) recommended estimating the average effect using the mean group estimator that equally averages the coefficients obtained from separate regressions for each individual. Another widely used average-effect estimator is to first aggregate the data over individuals and then estimate aggregate time-series regressions. A seminal and comprehensive study of aggregation estimation is given by Pesaran et al. (1989). Hsiao and Pesaran (2008) provided a comprehensive review of random coefficient models and a Bayesian approach for this model. If the average effect is of interest, it can be shown that the FGLS estimator can be written as a weighted average of estimators from different poolings, and it makes a good tradeoff between bias and efficiency.

In contrast to the mean, researchers are sometimes more interested in the individual coefficients (individual effect), especially when heterogenous policy implications or decisions needed to be made. Estimating the individual effect is the focus of this paper. A popular individual effect estimator is the mixed estimator of Lee and Griffiths (1979) that incorporates the average effect in individual coefficient estimation. Maddala et al. (1997) proposed a shrinkage estimator that is a combination of the pooled and individual regressions. If a more structural regression is considered, an interesting estimator proposed by Pesaran et al.

(1999) is to distinguish between the long-run and short-run parameters, and only allows the short-run parameters to be heterogenous (pooled mean group estimator). The validity of this estimator relies on a careful specification of the long-run and short-run parameters. See Baltagi et al. (2008) for an excellent survey. Recent developments in individual-effect estimation involves a latent class/group specification, such as finite mixture models and various types of grouped estimators. A significant approach introduced by Bonhomme and Manresa (2012) is called the grouped fixed-effects estimator that jointly estimates the unknown grouped pattern in the fixed effect parameters and slope coefficients. Similar to the pooled mean group estimator, a heterogeneity structure is imposed here that only the fixed effect parameters are heterogenous. Su et al. (2014) proposed to use penalized estimation to simultaneously estimate the unknown group membership and coefficients. Other grouped estimators include Lin and Ng (2012), Bester and Hansen (forthcoming), among others.[1]

Compared with these existing approaches, our proposed pooling averaging approach includes the shrinkage estimator as a special case, and does not require assumption on the heterogeneity structure. More particularly, we allow the individuals' coefficients to be completely heterogenous or group-wise heterogenous. Besides, although MSE or MSPE is widely used in simulations and applications to evaluate a method, most existing estimators are obtained based on other criteria, and only their asymptotic properties are examined. To our best knowledge, no finite sample properties of existing heterogenous panel estimators are theoretically studied except for the shrinkage estimator (Maddala et al., 2001). Unlike existing estimators, our pooling averaging method explicitly aims at minimizing the MSE, and its finite sample and asymptotic properties are theoretically analyzed.

Recently, a burgeoning literature have provided various tests for parameter homogeneity, differing in the model specifications. For example, Pesaran and Yamagata (2008) proposed dispersion type tests for large panels with cross section dimension $N$ and time dimension $T$ both to infinity. Juhl and Lugovskyy (2010) focused on the typical micro-panel with large $N$ and fixed $T$. Westerlund and Hess (2011) proposed a poolability test for a cointegrated panel. Su and Chen (2013) based on regression residuals and proposed a test applicable in the panel with interactive fixed effects. Jin and Su (2013) considered the case with cross-section dependence and provided a nonparametric test. If the ulti-

[1] If multi-factor error structure or error cross-section dependence are considered, then common correlated effects estimators (Pesaran, 2006; Pesaran and Tosetti, 2011) should be used.

mate interest lies in estimating the parameters of the models, the pretest estimator is not completely satisfactory. It is discontinuous and has unbounded risk. More importantly, it ignores the uncertainty in the testing procedure as it is conditional on the selected model. These are the general problems of pretesting estimators (see Danilov and Magnus (2004) for a detailed discussion). Besides, in the context of heterogenous panel, even if the true model is selected, it does not necessarily produce the best estimator in terms of MSE.

# 3 Pooling averaging estimation

## 3.1 Model setup

Consider the panel data model with heterogeneous slopes

$$y_i^* = \alpha_i + X_i^* \beta_i + u_i^* \qquad i = 1, \ldots, N, \tag{1}$$

where $y_i^* = (y_{i1}^*, \ldots, y_{iT}^*)'$ and $X_i^* = (X_{i1}', \ldots, X_{iT}')'$ is a $T \times k$ matrix of explanatory variables including the intercept i.e. $X_{it1}^* = 1$ for $t = 1, \ldots, T$, and the series $\{y_{it}^*, X_{it}\}$ is assumed to be stationary. The coefficient $\beta_i = (\beta_{i1}, \ldots, \beta_{ik})'$ is *fixed* but allowed to vary over individuals, i.e. some or all of the elements in $\beta_i$ can be different from the elements in $\beta_j$ for $i \neq j$. The error term, $u_i^*$, is assumed to be independently and identically distributed (IID) with mean zero and variance $\sigma_i^2 I_T$, and $u_1^*, \ldots, u_N^*$ are also uncorrelated conditional on $X_i^*$ and $c_i^*$ for all $i$. The unobserved individual effect $\alpha_i$ can be correlated with $X_i^*$. To incorporate this fixed effect, we make a preliminary within-transformation or first-differencing of the data, leading to the transformed dependent and explanatory variables $y_i$ and $X_i$. As it will be shown, the preliminary transformation does not affect the following analysis. Thus model (1) can be written with the transformed variables as

$$y_i = X_i \beta_i + u_i \qquad i = 1, \ldots, N, \tag{2}$$

**Remark 1** *Our setup differs from the random coefficient model proposed by Swamy (1970) mainly in the assumption of slope coefficients. In Swamy's model $\beta_i$ is treated as random, and coefficient heterogeneity are modelled by different realizations of the same distribution or distributions with the same first and second order moments. However, we treat the coefficients as fixed but unknown parameters, a common assumption in the standard regression analysis, and we allow them to vary over individuals in any arbitrary form. Therefore, the way of modelling heterogeneity is more general in (2) than in the random coefficient model.*

**Remark 2** *The error distribution is assumed to be identical for each individual, but allowed to vary across individuals. The IID assumption of error terms is not essential but it simplifies the derivation and notation. If within-individual heteroscedasticity is suspected or serial correlation is allowed in the error (either within or across individuals), the following analysis can be conveniently adjusted by replacing OLS estimation with feasible generalized least square estimation (FGLS). We shall make this point more explicit after introducing our pooling averaging estimator.*

**Remark 3** *To demonstrate our method and stay focused, we first discuss the case of strict exogenous explanatory variables, i.e.* $E(u_i^*|X_i^*, \alpha_i) = 0$ *in (1), and thus* $E(X_i' u_i) = 0$ *in (2). This assumption ensures the unbiasedness and consistency of the fixed effect estimator, but it rules out the dynamic model where the lagged dependent variable is included as explanatory variables. If the lagged dependent variable is included in the model, fixed effect estimation is generally inconsistent, and the degree of inconsistency depends on the degree of state dependence and* $T$*. In that case instrument variable estimation can be applied to model (2), and our pooling averaging estimation can be accordingly adjusted. We shall discuss the dynamic model in more details in Section 5.4.*

Since the fixed effect estimator can be obtained by OLS estimation of the transformed model (2), our following analysis will mainly be based on (2). We further assume that the time series dimension $T$ is large enough, so that we could estimate individual time series separately. We shall refer to the OLS estimators of individual time series in Equation (2) as OLS individual estimators, denoted by $\widehat{\beta}_i = (\widehat{\beta}_{i1}, \ldots, \widehat{\beta}_{ik})'$ (equivalent to system OLS without coefficient restrictions; see, e.g., Wooldridge (2002), Chapter 7). Model (2) can be written in a matrix form as

$$y = X\beta + u, \tag{3}$$

where $y = (y_1', \ldots, y_N')'$, $X = \text{diag}(X_1, \ldots, X_N)$, $\beta = (\beta_1', \ldots, \beta_N')'$, and $u = (u_1', \ldots, u_N')'$.

## 3.2 Average pooling strategies

If the slope coefficients are of interest, our goal here is to estimate each individual's coefficient $\beta_i$ (Lee and Griffiths, 1979).[2] This is different from the random coefficient model

[2]If the forecast is of interest, combining pooling strategies is amount to combining different forecasts. Our proposed pooling averaging method includes forecast combination as a special case; See Section 5.3

where the goal is to estimate a common average effect, $\bar{\beta} = \mathrm{E}(\beta_i)$. Individual's coefficient is of particular interest when individualized policies or decisions need to made or individual forecast is desired. A typical example is the brand choice analysis where researchers/companies are especially interested in whether and how a product feature imposes heterogenous effects on the behavior of consumers at different ages, gender, and income. Another example is investigation of the determinants of a country's sovereign CDS spreads (Longstaff et al., 2011). This empirical application raises several interesting and important questions: How do a country's sovereign CDS spreads depend on its own domestic economic performance and global factors? Whether the effect of the same global factor varies across countries? Is there any common feature in the determination process of CDS spreads across countries. See Hoogstrate et al. (2000), Baltagi and Griffin (1997), and Baltagi et al. (2000) for more empirical examples. To estimate this model, one can consider individual estimation, i.e. $\widehat{\beta}_i = (X_i'X_i)^{-1}X_i'y_i$. The individual estimator $\widehat{\beta}_i$ is consistent given that individual $i$'s regression is correctly specified, but it is not efficient as it is a limited information estimator. In the other extreme, one can ignore the heterogeneity and estimate the pooled model, obtaining a common estimator for all individuals, i.e. $b = (\sum_{i=1}^N X_i'X_i)^{-1}\sum_{i=1}^N X_i'X_i\widehat{\beta}_i$. The pooled estimator $\widehat{\beta}_{\text{pool}} = (b', \ldots, b')'$ is more efficient than the individual estimator, but can be severely biased due to incorrect pooling of heterogenous coefficients. The comparison between the individual and pooled estimator suggests a typical bias-variance tradeoff in choosing which estimator to use.

An intermediate estimator between the individual and pooled is to restrict some of the coefficients to be identical. This can be done by imposing equality restrictions to a set of coefficients when estimating (3), i.e.

$$R_m\beta = 0, \tag{4}$$

where $R_m$ is the restriction matrix under the $m$-th pooling strategy with elements $r_{ij}$. Each of its column vector restricts two parameters to be equal, say $r_{i\cdot} = -r_{j\cdot} = 1$ if $\beta_i = \beta_j$. For each $R_m$, we can construct

$$P_m = I_{Nk} - (X'X)^{-1}R_m'(R_m(X'X)^{-1}R_m')^{-1}R_m, \tag{5}$$

then the OLS estimator under the $m$-th pooling strategy is

$$\widehat{\beta}_{(m)} = P_m\widehat{\beta}.$$

for the detailed discussion and Section 8.2 for numerical exercise.

where

$$\widehat{\beta} = (\widehat{\beta}_1', \ldots, \widehat{\beta}_N')',$$

a vector of OLS individual estimator. The individual effect estimator $\widehat{\beta}_{(m)}$ allows estimated coefficients to vary over individuals while restricting some of them to be the same. Different pooling strategies are characterized by different restrictions, $R_m$, and the resulting estimators have different degrees of bias and variance. The individual estimator corresponds to the pooling strategy with no restrictions, while the pooled estimator can be regarded as system OLS restricting identical coefficients over individuals. Then the question is how to determine which pooling strategy to use.

One approach is to test or select the most appealing pooling strategy based on some data-driven criterion. This approach is appealing if one can pick up the true model with correct parameter restrictions. However, in practice, the true model is difficult to obtained for at least two reasons. First, all candidate models can be misspecified in other aspects, such as control variables, linearity, and the error structure, and therefore one cannot distinguish between the bias caused by coefficient heterogeneity and that caused by other misspecification. Without identifying the source of bias, it seems impossible to select the right coefficient restrictions. Second, the coefficient estimates of all candidate models are subjected to estimation errors. Hence, it is difficult to tell whether the efficiency lost is caused by inefficient pooling or estimation noise. Even if one can correctly select the restriction, the true restriction specification does not always produce the best estimator in terms of MSE. This happens, for example, if the heterogeneity of coefficients is small while the signal-to-noise ratio is small. In this case incorrectly pooling heterogenous individuals may lead to lower MSE, because the efficiency gains from pooling dominate the heterogeneity bias. Therefore, if estimation is of central interest, it seems less plausible to focus on testing or selecting the right pooling pattern. Thus motivated, we propose to average estimators from different pooling strategy. By appropriately choose the averaging weights, we are able to make an optimal tradeoff between bias and efficiency. The pooling averaging estimator is given as

$$\widehat{\beta}(w) = \sum_{m=1}^{M} w_m \widehat{\beta}_{(m)} = \sum_{m=1}^{M} w_m P_m \widehat{\beta} = P(w)\widehat{\beta}, \tag{6}$$

where $P(w) = \sum_{m=1}^{M} w_m P_m$ being an $Nk \times Nk$ matrix and $w = (w_1, \ldots, w_M)'$ belongs to the set $\mathcal{W} = \{w \in [0,1]^M : \sum_{m=1}^{M} w_m = 1\}$.

When we estimate (3) under each restriction $R_m$, we use OLS estimation. If the error term $u_i$ is suspected to be serial-correlated or heteroscedastic within individuals, we can use FGLS estimation. If we estimate the variance of $u$ by $\widetilde{\Omega}$, then

$$\widetilde{\beta} = (X'\widetilde{\Omega}^{-1}X)^{-1}X'\widetilde{\Omega}^{-1}y$$

and

$$\widetilde{\beta}_{(m)} = \{I_{Nk} - (X'\widetilde{\Omega}^{-1}X)^{-1}R_m'(R_m(X'\widetilde{\Omega}^{-1}X)^{-1}R_m')^{-1}R_m\}\widetilde{\beta}.$$

To achieve consistency, FGLS require stronger assumptions on the error-regressor correlation, i.e., $\mathrm{E}(X_i \otimes u_i) = 0$, where $\otimes$ denotes the Kronecker product, and its efficiency gains also requires extra conditions (Wooldridge, 2002, pp. 154–161). Hence, the choice of OLS and FGLS estimators depends on the tradeoff between efficiency and robustness. For simplicity, we demonstrate our estimator based on OLS.

We end this section by comparing our pooling averaging (individual effect) estimator with several important estimators for heterogenous panel. First, our pooling averaging estimator includes shrinkage estimator of Maddala et al. (1997) as a special case (see Section 5.3). The shrinkage estimator is defined as

$$\widehat{\beta}_s = \left(1 - \frac{\nu}{F}\right)\widehat{\beta} + \frac{\nu}{F}\widehat{\beta}_{\text{pool}}, \tag{7}$$

where $\nu = [(N-1)k - 2]/[NT - Nk + 2]$ and $F$ is the test statistic for null hypothesis $H_0 : \beta_1 = \ldots = \beta_N$. More specifically, if we denote $\widetilde{R}$ as the restriction matrix associated with $H_0$ and $\widetilde{\sigma}^2 = (Y - X\widehat{\beta})'(Y - X\widehat{\beta})/(NT - Nk)$, then the $F$ statistic is $F = (\widetilde{R}\widehat{\beta})'(\widetilde{R}(X'X)^{-1}\widetilde{R}')^{-1}(\widetilde{R}\widehat{\beta})/[(N-1)k\widetilde{\sigma}^2]$.

By setting zero weights for all $\widehat{\beta}_{(m)}$ except for the individual estimator and pooled estimator, $\widehat{\beta}(w)$ reduces to $\widehat{\beta}_s$. Next, we compare with two average effect estimator: the mean group estimator of Pesaran and Smith (1995) and GLS estimator of Swamy (1970). They can be both written as an average of each individual estimator by

$$b(w) = \sum_{i=1}^{N} w_i^* \widehat{\beta}_i, \tag{8}$$

where $w_i^*$ is some weight. The mean group estimator takes the weight $w_i^* = 1/N$, while the GLS estimator takes the weigth

$$w_i^* = \left\{\sum_{i=1}^{N}[\Delta + \sigma_i^2(X_i'X_i)^{-1}]^{-1}\right\}^{-1}[\Delta + \sigma_i^2(X_i'X_i)^{-1}]^{-1}.$$

Our pooling averaging estimator differs from the average effect estimator $b(w)$ in two main aspects. First, in contrast to common coefficients for all individuals, our pooling averaging estimator allows some of individuals to have common coefficients while the others to have different ones. Besides, the candidate estimator to be averaged in $b(w)$, $\widehat{\beta}_i$, is obtained from each individual time series estimation, using only a subset of the sample. On the contrary, each candidate estimator to be averaged in $\widehat{\beta}(w)$, $\widehat{\beta}_{(m)}$, is obtained using the entire sample. We will further compare the finite sample performance of our estimator with these average effect estimators in the simulation and empirical application.

# 4 Pooling, non-pooling, and averaging: The MSE comparison

Before we discuss the choice of weights for the pooling averaging estimator, we first examine under which situation the pooling averaging exhibits good finite sample performance in general. We theoretically compare the MSE of the pooling averaging estimator with the pooled and individual estimators. The pooling selection estimator, as a special case of the pooling averaging estimator, is also considered. The purpose of this analysis is twofold. Although there have been many empirical studies showing that the performance of estimators differs significantly in applications, there is lack of theoretical explanation, and no consensus is reached on which estimator to use in different practical situations. Hence, the first purpose is to provide theoretical explanations on the diverging performance of the pooled and individual estimators. Second, the theoretical comparison here also sheds some light on how the features of the data and models affect the performance of alternative estimators, such as the degree of coefficient heterogeneity and noise of the model. This further provides guidance on which estimator to choose in practice. For brevity, the comparison here focuses on the mean square error of coefficient estimators. Similar comparison regarding the predicted value of the dependent variable can be made if the interest lies in forecasting.

For notation simplicity, we denote $Q_i = X_i'X_i$, $Q = \sum_{i=1}^N Q_i$, $V_i = \sigma_i^2 Q_i^{-1}$, $V = \text{diag}(V_1, \ldots, V_N)$, and $\|\theta\|^2 = \theta'\theta$ for any vector $\theta$. The pooled estimator can be obtained by the system OLS restricting all coefficients to be the same, i.e. $\widehat{\beta}_{\text{pool}} = (b', \ldots, b')'$, where $b = Q^{-1}\sum_{i=1}^N Q_i\widehat{\beta}_i$. The individual estimator $\widehat{\beta} = (\widehat{\beta}_1', \ldots, \widehat{\beta}_N')'$ is calculated by the

system OLS without coefficient restrictions. Its elements $\widehat{\beta}_i$'s are uncorrelated due to the assumption of uncorrelated $u_i$'s conditional on $X$, and $\widehat{\beta}_i \sim (\beta_i, \sigma_i^2 Q_i^{-1})$.[3] Hence, the MSEs of the pooled and individual estimator can be obtained by

$$\text{MSE}_{\text{ind}} \equiv \text{MSE}(\widehat{\beta}) = \sum_{i=1}^{N} \text{E}\|\widehat{\beta}_i - \beta_i\|^2 = \text{tr}(V) \tag{9}$$

and

$$\begin{aligned} \text{MSE}_{\text{pool}} &\equiv \text{MSE}(\widehat{\beta}_{\text{pool}}) = \sum_{i=1}^{N} \text{E}\|b - \beta_i\|^2 \\ &= \sum_{i=1}^{N} \|Q^{-1} \sum_{i=1}^{N} Q_i \beta_i - \beta_i\|^2 + \sum_{i=1}^{N} \text{tr}(Q^{-1} V_i Q^{-1}). \end{aligned} \tag{10}$$

The first term in Equation (10) captures the bias caused by pooling heterogenous coefficients, and the second term measures the variance. There is no guarantee that $\text{MSE}_{\text{ind}}$ is less than $\text{MSE}_{\text{pool}}$. For example, if we have $Q_i = I$, then $\text{tr}(Q^{-1} V_i Q^{-1}) = k^{-2}\text{tr}(V_i) \leq \text{tr}(V_i)$, and the relation between $\text{MSE}_{\text{ind}}$ and $\text{MSE}_{\text{pool}}$ depends on the magnitude of the bias term in (10) and the difference of variance terms $\text{tr}(V) - k^{-2}\text{tr}(V)$. To derive the MSE of the pooling averaging estimator, we need to distinguish between two cases: fixed weight averaging and random weight averaging.

## 4.1 Fixed weights

Fixed (equal) weighting is a popular method in the forecast combination literature due to its good empirical performance, while its performance is not much investigated in the model averaging literature. Treating $w$ as fixed, we can obtain the MSE of the pooling averaging estimator as

$$\begin{aligned} \text{MSE}_{\text{fw}} &\equiv \text{MSE}(\widehat{\beta}(w)) = \text{E}\|\widehat{\beta}(w) - \beta\|^2 = \|P(w)\widehat{\beta} - \beta\|^2 \\ &= \|P(w)\beta - \beta\|^2 + \text{tr}[P(w) V P'(w)]. \end{aligned} \tag{11}$$

The first term is the deviation between the weighted average and original (non-averaged) true coefficient, and can be thought of as the square of bias of $\widehat{\beta}(w)$. The second term

[3] In the dynamic panel the OLS estimator is biased. Comparing the MSEs of biased estimators is still possible, but it complicates the analysis since the degree of bias differs over models.

represents the variance of $\widehat{\beta}(w)$. We see that the comparison between the MSEs depends on the degree of heterogeneity of the *true* coefficients $\beta$, the error variance of individual regressions $\sigma^2$, and of course the weight choice.

To demonstrate how these parameters affect the performance in an intuitive way, we consider a simple case with $N = 2$. The coefficient vector of interest is then $\beta = (\beta_1', \beta_2')'$. We also assume $Q_i = I$ without loss of generality.[4] The pooling averaging estimator in this case is

$$\widehat{\beta}(w) = w_1\widehat{\beta}_{\text{pool}} + w_2\widehat{\beta}_{\text{ind}} = (w_1\widehat{\beta}' + w_2\widehat{\beta}_1', w_1\widehat{\beta}' + w_2\widehat{\beta}_2')'.$$

The MSEs of the individual estimator, pooled estimator, and pooling averaging estimator are given, respectively, by

$$\text{MSE}_{\text{ind}} = k(\sigma_1^2 + \sigma_2^2), \qquad \text{MSE}_{\text{pool}} = \|\beta_1 - \beta_2\|^2/2 + k(\sigma_1^2 + \sigma_2^2)/2,$$

and

$$\begin{aligned} \text{MSE}_{\text{fw}} &= w_1^2\|\beta_1 - \beta_2\|^2/2 + (w_1^2 + 2w_2^2 + 2w_1w_2)k(\sigma_1^2 + \sigma_2^2) \\ &= w_1^2\|\beta_1 - \beta_2\|^2/2 + (1 + w_2^2)k(\sigma_1^2 + \sigma_2^2)/2. \end{aligned} \tag{12}$$

Under equal weight setting, $w_1 = w_2 = 1/2$, Equation (12) specializes to $\text{MSE}_{\text{fw}} = \|\beta_1 - \beta_2\|^2/8 + 5k/4$. Comparing the pooling averaging and pooled estimation, we see that $\text{MSE}_{\text{fw}} < \text{MSE}_{\text{pool}}$ if and only if

$$\|\beta_1 - \beta_2\|^2 > \frac{k(\sigma_1^2 + \sigma_2^2)w_2^2}{1 - w_1^2}. \tag{13}$$

This suggests that the pooling averaging estimator is superior to the pooled estimator if the difference between individual coefficients is large enough. In an extreme case of completely homogenous panel, $\|\beta_1 - \beta_2\|^2 = 0 \leq k(\sigma_1^2 + \sigma_2^2)w_2^2/(1 - w_1^2)$, and we always have $\text{MSE}_{\text{fw}} \geq \text{MSE}_{\text{pool}}$ as expected. From a different angle, we also note that as the variance of errors and the number of regressors increase, the pooled estimator is more likely to outperform the pooling averaging estimator. To see this, let $\bar{\sigma}^2 = (\sigma_1^2 + \sigma_2^2)/2$ representing the average of error variance, then $\text{MSE}_{\text{fw}} < \text{MSE}_{\text{pool}}$ requires

$$\bar{\sigma}^2 < \frac{(1 - w_1^2)\|\beta_1 - \beta_2\|^2}{2w_2^2 k}, \tag{14}$$

[4]This can be realized by normalizing the regressors by $X_iH_i\Lambda_i^{-1/2}$, where $\Lambda_i$ is a diagonal matrix and $H_i$ is an orthogonal matrix such that $H_i'X_i'X_iH_i = \Lambda_i$.

implying that the pooling averaging estimator is more favorable when the noise is limited.

We then compare the pooling averaging with the individual estimation. We have $\text{MSE}_{\text{fw}} < \text{MSE}_{\text{ind}}$ if and only if

$$\|\beta_1 - \beta_2\|^2 < \frac{2\bar{\sigma}^2 k(1 - w_2^2)}{w_1^2}, \tag{15}$$

or from another point of view

$$\bar{\sigma}^2 > \frac{w_1^2 \|\beta_1 - \beta_2\|^2}{2k(1 - w_2^2)}. \tag{16}$$

Inequality (15) shows that pooling averaging is advantageous over individual estimation when coefficient heterogeneity is bounded by the product of $2\bar{\sigma}^2$ and $k$ (since $(1-w_2^2)/w_1^2 > 1$). Even if the panel is completely heterogenous with all coefficients different over individuals, the pooling averaging can still outperform the individual estimation if the variance of errors is large or there are too many explanatory variables in the model. From the viewpoint of $\bar{\sigma}^2$, a larger value of $\bar{\sigma}^2$ favors the pooling averaging as the inequality is more likely to hold in (16).

Finally, we comment on the performance of the pooled and individual estimator. In such a highly simplified case, we see although the pooled estimator introduces a bias term, it can still produce lower MSE than the individual estimator if $\|\beta_1 - \beta_2\|^2 < k(\sigma_1^2 + \sigma_2^2)$. Since the empirical models typically have nontrivial error variances and/or many regressors, this result, to some extent, explains why most empirical researches find individual estimation less desirable.

## 4.2 Random weights

Next, we consider the random weights. Typically, the random weights are correlated with the estimated coefficients as they are estimated from the same data. Also, the weights of different models can be correlated. We define $\rho_m = \text{cov}(w_m, \widehat{\beta})$, $\kappa_{m,l} = \text{cov}(w_m, w_l)$, $\delta_{m,l} = \text{cov}(w_m w_l, \widehat{\beta}' P_m' P_l \widehat{\beta})$ for $m, l \in \{1, \ldots, M\}$, and let $\bar{w} = \text{E}(w)$. Then the MSE of the pooling averaging estimator with random weights can be written as

$$\begin{aligned}
\text{MSE}_{\text{rw}} &= \text{E}\|P(w)\widehat{\beta} - \beta\|^2 \\
&= \text{E}\|P(w)\widehat{\beta}\|^2 - 2\text{E}\beta' P(w)\widehat{\beta} + \|\beta\|^2 \\
&= \|P(\bar{w})\beta - \beta\|^2 + \text{tr}[P(\bar{w}) V P'(\bar{w})] + \iota'\Phi\iota + \Gamma_1 - 2\Gamma_2,
\end{aligned} \tag{17}$$

where $\Phi$ is the matrix with the typical element $\delta_{m,l}$, $\iota$ is a vector of ones,

$$\Gamma_1 = \sum_{m,l} \kappa_{m,l} \beta' P_m' P_l \beta, \quad \text{and} \quad \Gamma_2 = \sum_m \beta' P_m \rho_m.$$

The first two terms of $\text{MSE}_{\text{rw}}$ resembles $\text{MSE}_{\text{fw}}$. In addition, estimating the weights introduces three covariance terms. Although $\text{MSE}_{\text{rw}}$ is not necessary to be always larger than $\text{MSE}_{\text{fw}}$, the forecast combination literature suggests that estimating the weights often leads to poorer finite sample performance than simple average due to the added randomness in $\widehat{\beta}(w)$; See, for example, Stock and Watson (2004) for empirical evidence and Smith and Wallis (2009) and Vasnev et al. (2014) for theoretical explanations. We shall investigate whether this is also true in pooling averaging using simulated and real data.

## 4.3 Model selection weights

Finally, it is worthwhile considering model selection as a special case within the framework of pooling averaging. The model selection estimator assigns the weight one to the estimator from the selected model and zero for the others. Let $\mathbf{b}_M = (\widehat{\beta}_{(1)}', \ldots, \widehat{\beta}_{(M)}')'$, $U = (\widehat{\beta}_{(1)} - \beta, \ldots, \widehat{\beta}_{(M)} - \beta)$, and $\Pi$ is an $M \times M$ matrix with the $m_1 m_2$-th element $\Pi_{m_1 m_2} = \text{E}[(\widehat{\beta}_{(m_1)} - \beta)'(\widehat{\beta}_{(m_2)} - \beta)|w]$. To have an intuitive explanation of model selection MSE, we can rewrite the $\text{MSE}_{\text{rw}}$ in the form of conditional moments as

$$\begin{aligned} \text{MSE}_{\text{rw}} &= \text{E}(w'U'Uw) = \text{E}\big[w'\text{E}(U'U|w)w\big] \\ &= \text{E}\big[w'\text{var}(\mathbf{b}_M|w)w\big] + \text{E}\big[w'\Pi w\big]. \end{aligned}$$

Let $p_m = \text{Prob}(\text{model } m \text{ is selected})$, then the model selection MSE can be written as

$$\begin{aligned} \text{MSE}_{\text{sel}} &= \sum_{m=1}^{M} p_m \text{var}(\widehat{\beta}_{(m)}|w_m = 1) + \sum_{m=1}^{M} p_m \text{Bias}^2(\widehat{\beta}_{(m)}|w_m = 1) \\ &= \sum_{m=1}^{M} p_m \text{MSE}(\widehat{\beta}_{(m)}|w_m = 1). \end{aligned} \tag{18}$$

It is a weighted MSE of the $m$-th estimator conditional on the $m$-th model being selected, and weight is the probability that this $m$-th model is selected. We see that even though only a single model (pooling strategy) is selected, the MSE values of all models are combined. This highlights the important fact that conditional inference only based on the selected model can be misleading.

To summarize, we show that the choice of alternative estimators depends on the degree of coefficient heterogeneity and noise contained in the model. This provides general guidance how to choose among alternative estimators in practice. For example, in the case with large noise but small degree of heterogeneity, the pooled estimator is often preferred. On the contrary, the individual estimator can outperforms others in the case with small noise and large degree of heterogeneity. If we are in the non-extreme cases with moderate degree of heterogeneity and noise, pooling averaging is recommended since it is expected to produce the most accurate estimator. Besides, we see that estimating the averaging weights can introduce extra uncertainty in the mean square error. Thus, it is sometimes more preferred to use fixed (and equal) weights in combining different pooling estimators. We shall further investigate the case-dependent performance of different estimators via simulation, and more detailed guidance would be provided then.

# 5 Choosing pooling averaging weights

We have seen that the pooling averaging estimator can make an appropriate tradeoff between bias and variance, depending on how the weights are chosen. We consider three methods to choose the pooling averaging weights.

## 5.1 Simple average

Simple average assigns equal weights to each candidate models, i.e. $w_m^E = 1/M$. While the simple average is typically favored in the forecast combination literature, its performance in the pooling averaging is not widely studied. Smith and Wallis (2009) argued in the context of forecasting that the simple average systematically outperforms the weighted average (using estimated weights) if they are theoretically equivalent. Vasnev et al. (2014) provided a theoretical framework to show under which situations equal weight averaging is favored. In the pooling averaging context, we expect that the simple average could have good finite sample performance because the estimated weights can be inaccurate due to the errors in the coefficient estimates and the strong correlation between these estimates. Simple average thus gains in efficiency by trading off a small bias against a larger estimation variance. This is in line with the implication of Smith and Wallis (2009).

## 5.2 Smoothed information criteria

Smoothed information criteria (SAIC and SBIC) are introduced by Buckland et al. (1997), which average the estimates using the weights

$$w_m^A = \frac{\exp(-\text{AIC}_m/2)}{\sum_{m=1}^M \exp(-\text{AIC}_m/2)} \quad \text{and} \quad w_m^B = \frac{\exp(-\text{BIC}_m/2)}{\sum_{m=1}^M \exp(-\text{BIC}_m/2)}.$$

Buckland et al. (1997) argued that the weights defined in this way ensure that two models with the same value for information criterion are assigned to the same weight. The weights using $w_m^B$ can also be regarded as an approximation of posterior model probabilities, and thus average using SBIC approximates the Bayesian model average. To avoid the infinity weights, we adjust $\text{AIC}_m$ and $\text{BIC}_m$ by subtracting the minimum AIC from its original values.

## 5.3 Mallows pooling averaging

Hansen (2007) proposed to average the model using the Mallows criterion, which is asymptotically optimal in the sense of achieving the lowest possible squared error. This method is further justified by Wan et al. (2010). In light of the well-established asymptotic property, we consider Mallows pooling averaging (MPA) in estimating a heterogenous panel.

For any vector $\theta$ and symmetric matrix $A$ we set $\|\theta\|_A^2 = \theta' A\theta$. We focus on the squared loss $L_A(w) = \|\widehat{\beta}(w) - \beta\|_A^2$ and the squared risk $R_A(w) = \text{E}\{L_A(w)\}$, then the Mallows criterion can be written as

$$\mathcal{C}_A(w) = \|P(w)\widehat{\beta} - \widehat{\beta}\|_A^2 + 2\text{tr}[P'(w)AV] - \|\widehat{\beta} - \beta\|^2. \tag{19}$$

**Theorem** 5.1 *The Mallows criterion defined in Equation (19) is an unbiased estimator of the squared risk $R_A(w)$.*

**Proof**: See Appendix A1.

We choose weights by minimizing the above criterion, that is

$$\widehat{w} = (\widehat{w}_1, \ldots, \widehat{w}_M)' = \arg\min_{w \in \mathcal{W}} \mathcal{C}_A(w). \tag{20}$$

Note that although $\|\widehat{\beta} - \beta\|^2$ appears in $\mathcal{C}_A(w)$, it is unrelated to the choice of $w$. If we denote $D = (\widehat{\beta}_{(1)} - \widehat{\beta}, \ldots, \widehat{\beta}_{(M)} - \widehat{\beta})$ and $v = \{\text{tr}(P_1 AV), \ldots, \text{tr}(P_M AV)\}'$, then we can rewrite the criterion as

$$\mathcal{C}_A(w) = w'D'ADw + 2w'v - \|\widehat{\beta} - \beta\|^2, \tag{21}$$

which is clearly a quadratic function of $w$. In practice, $\sigma_i^2$ is unknown, and we replace it with $\widehat{\sigma}_i^2 = \widehat{u}_i'\widehat{u}_i/(T-k)$, where $\widehat{u}_i$ is the OLS residual of the $i$-th individual regression.

The criterion $\mathcal{C}_A(w)$ is a generalization of the Mallows pooling averaging criterion defined by Hansen (2007). When $A = X'X$, $\mathcal{C}_A(w)$ simplifies to Hansen's criterion (Equation (11) in Hansen (2007)). In this case one is interested in prediction and focuses on the average (prediction) squared error loss $(X\widehat{\beta}(w) - \mathrm{E}(y))'(X\widehat{\beta}(w) - \mathrm{E}(y))$. Alternatively, one can also concentrate on the accuracy of the estimated coefficients, which is often the research interest in the heterogenous panel. This corresponds to taking $A = I_{Nk}$, and the criterion $\mathcal{C}_A(w)$ aims at minimizing the average squared error of coefficient estimates.

We examine the finite sample property of the MPA estimator by deriving its risk bounds. The risk bounds are widely used as an important theoretical property (or justification) of a estimation procedure. See, for example, Juditsky and Nemirovski (2000), Barron et al. (1999), Yang (2001), Yang (2003), and Yuan and Yang (2005). The risk bound analysis tells us how the Mallows pooling averaging performs in the worst situation.

**Theorem** 5.2 *The upper bound of the risk of MPA estimator is*

$$R_A(\widehat{w}) \quad \leq \quad (1-c)^{-1} \inf_{w \in \mathcal{W}} R_A(w) + (1-c)^{-1}(c^{-1}tr(AV) - 2E[tr\{P'(\widehat{w})AV\}), \quad (22)$$

*where $c$ is a constant belonging to $(0,1)$.*

**Proof**: See Appendix A2.

It shows that up to the constant $(1-c)^{-1}$ and the additive penalty $(1-c)^{-1}[c^{-1}\mathrm{tr}(AV) - 2\mathrm{E}\,\mathrm{tr}\{P'(\widehat{w})AV\}]$, the pooling averaging estimator $\widehat{\beta}(\widehat{w})$ achieves the performance of the optimal non-random weight pooling averaging estimator. The holding of (22) does not depend on sample size. Let $\mathcal{I}_1(\cdot)$ and $\mathcal{I}_2(\cdot)$ be the minimum and maximum eigenvalues of a symmetric matrix and $\sigma^2_{\max} = \max_{i=1,\ldots,M} \sigma_i^2$. Then we can derive the more specific risk bound for coefficient estimation and prediction, respectively.

**Corollary** 5.1 *If $\min_{i=1,\ldots,M} \mathcal{I}_1(T^{-1}X_i'X_i) \geq c_1$, then there exists $c \in (0, 0.5)$ such that when $A = I_{Nk}$,*

$$R_A(\widehat{w}) \leq \frac{1}{1-c} \inf_{w \in \mathcal{W}} R_A(w) + \frac{1-2c}{c(1-c)} \frac{Nk\sigma^2_{\max}}{Tc_1} + \frac{4}{1-c} \frac{Nk\sigma^2_{\max}}{Tc_1}, \quad (23)$$

*and when* $A = X'X$,

$$R_A(\widehat{w}) \leq \frac{1}{1-c} \inf_{w \in \mathcal{W}} R_A(w) + \frac{1-2c}{c(1-c)} Nk\sigma^2_{\max} + \frac{4}{1-c} Nk\sigma^2_{\max}. \tag{24}$$

**Proof**: See Appendix A3.

For the two choices of $A$, the above corollary shows that the risk bound of the Mallow pooling averaging estimator is not dependent on $X$ and $y$, but only determined by a set of constants $\{T, N, k, \sigma^2_{\max}, c_1, c\}$ through additive penalties. As expected, the risk bound in both cases is large if we have a large $N$ panel with many regressors and large variances of residuals. On the contrary, a large time dimension $T$ can reduce the risk bound in the case of $A = I_{Nk}$.

Next, we study the asymptotic property of MPA estimator following the model averaging literature. Here, we assume $N$ to be fixed. The fixed $N$ large $T$ setup is particular relevant in cross-country analysis where the number of countries is typically finite. Let $\xi_T = \inf_{w \in \mathcal{W}} R_A(w)$. We assume the following condition

$$MT^{-1/2} \xi_T^{-1} \mathcal{I}_2(A) \to 0, \tag{25}$$

as $T \to \infty$. For $A = X'X$, under the condition that $\max_{i=1,\ldots,M} \mathcal{I}_1(T^{-1} X_i' X_i) \leq c_2 < \infty$, we know that a necessary condition of (25) is $\xi_T^{-1} = o(M^{-1} T^{-1/2})$, which is similar to the condition (7) of Ando and Li (2014) and requires that candidate models are approximations. For $A = I_{Nk}$, Condition (25) simplifies to $MT^{-1/2} \xi_T^{-1} \to 0$, which constrains the rate of $\xi_T \to 0$.

**Theorem** 5.3 *As* $T \to \infty$, *if the condition (25) is satisfied,* $X'u = O_p(T^{1/2})$, *and* $T^{-1} X'X \to \Psi$ *where* $\Psi$ *is a positive definite matrix, then*

$$\frac{L_A(\widehat{w})}{\inf_{w \in \mathcal{W}} L_A(w)} \to 1, \tag{26}$$

*in probability.*

**Proof**: See Appendix A4.

The result (26) means that MPA estimator $\widehat{\beta}(\widehat{w})$ is asymptotically optimal in the sense that its squared loss is asymptotically identical to that by the infeasible best possible model averaging estimator.

**Remark 4** *Here we shall show how the Mallows pooling averaging estimator is associated with the shrinkage estimator of Maddala et al. (1997) in the context of combining only two estimators,* $\widehat{\beta}$ *and* $\widehat{\beta}_{pool}$*. In this case, the averaged estimator is* $\widehat{\beta}(w) = w_1\widehat{\beta} + w_2\widehat{\beta}_{pool}$*. Following Maddala et al. (1997), we assume* $\sigma_1^2 = \cdots = \sigma_N^2 = \sigma^2$*, and* $\sigma^2$ *can be estimated by* $\widetilde{\sigma}^2$*. Let* $A = X'X$ *and* $P_{pool} = I_{Nk} - (X'X)^{-1}\widetilde{R}'(\widetilde{R}(X'X)^{-1}\widetilde{R}')^{-1}\widetilde{R}$*, where* $\widetilde{R}$ *is defined below (7). By minimizing* $C(w)$*, we have*

$$\widehat{w}_2 = \frac{\widetilde{\sigma}^2(Nk - tr(P_{pool}))}{\|\widehat{\beta} - \widehat{\beta}_{pool}\|^2} = \frac{\widetilde{\sigma}^2 Nk}{\widehat{\beta}'\widetilde{R}'(\widetilde{R}(X'X)^{-1}\widetilde{R}')^{-1}\widetilde{R}\widehat{\beta}}.$$

*Let* $\nu^* = N/(N-1)$*. If* $\nu^*/F \in [0,1]$*, we can write*

$$\widehat{\beta}(\widehat{w}) = \left(1 - \frac{\nu^*}{F}\right)\widehat{\beta} + \frac{\nu^*}{F}\widehat{\beta}_{pool}, \tag{27}$$

*which has the same form as the shrinkage estimator defined by (7).*

## 5.4 Dynamic panel data

So far we have assumed that explanatory variables are strict exogenous, and thus rule out the dynamic models with lagged dependent variables included as explanatory variables. In this section we discuss how our pooling averaging method is affected by this dynamic modeling.

It is well-known that the standard fixed effect estimator is biased and inconsistent if lagged dependent variables are controlled, since the explanatory variables are then correlated with errors. Wooldridge (2002, pp. 301–302) showed that the inconsistency from using fixed effects in this case is of order $T^{-1}$, given that individual time series is weakly dependent. Hence, if we consider the dynamic panel data model

$$y_{it}^* = \alpha_i + \gamma_i y_{it-1}^* + X_{it}^*\beta_i + u_{it}^*, \qquad i = 1, \ldots, N,$$

assuming that $|\gamma_i| < 1$ for all $i$ and $T$ is large, then the bias of fixed effect estimation under each pooling strategy can be ignored, and our pooling averaging procedure can be directly applied. Another justification of pooling averaging in dynamic panel data models is that the properties of the averaging estimator (Theorem 5.1– 5.3) are not affected by bias of candidate estimators. In fact, our pooling averaging estimator allows all candidate models to be misspecified and candidate estimators to be biased.

Although one can consider using instruments to achieve consistency for each model, it is not guaranteed to be advantageous for two reasons. First, our goal here is not obtain a consistent estimator, but an estimator with small MSE. Second, IV estimation typically achieves consistency by sacrificing efficiency or finite sample performance. Given MSE to be our ultimate criterion, IV estimation may not be beneficial compared to OLS.

# 6 Shrinking model space

There are numerous ways of imposing restrictions even when $N$ is moderate. For example, in the case of one regressor and $N$ individuals, we have $B_N$ ways of imposing restrictions, where $B_N$ is a Bell number.[5] This creates a huge model space. Selecting or averaging over the entire model space is thus computational formidable. More importantly, a large proportion of models in the model space can be poorly specified. Averaging over all these models will lead to a rather inaccurate averaging estimator due to poor weight estimation and poor estimators from these models (see also Yuan and Yang, 2005). To avoid these problems, we propose a preliminary screening procedure to rule out the "poor" models that incorrectly impose equality restrictions on far different coefficients. We consider two screening procedures depending on whether the estimation error/uncertainty of coefficients is taken into account. Both procedures start with normalizing the estimated coefficients $\widehat{\beta}_{il} = \widehat{\beta}_{il}/\max\{|\widehat{\beta}_{1l}|, \ldots, |\widehat{\beta}_{Nl}|\}$ for each $l = 1, \ldots, k$, so that the coefficients of different regressors have the same scale between $[-1, 1]$. Normalization avoids the numerical problems caused by extremely large numbers, and also allows us to group coefficients of different regressors using the same criteria.

## 6.1 Screening ignoring estimation uncertainty

We first treat individual coefficient estimates as fixed numbers, and group them simply based on their differences. This is called deterministically screening since no estimation uncertainty are taken into account. We employ agglomerative hierarchical clustering (AHC), one of the most popular data-driven clustering algorithms. It is computationally convenient and avoids the random initials in the K-means algorithms. AHC first computes the (dis)similarity between two estimates, measured by the squared Euclidean distance. (We

[5] $B_N$ is also noted as the $N$-th moment of a Poisson distribution with mean 1.

shall discuss an alternative distance function in the next subsection.) Then it takes the distance information and links pairs of objects that are close together into binary clusters (clusters made up of two estimates) using the linkage function. This function further links these newly formed clusters to each other and to other objects (either estimates or clusters) to create bigger clusters until all the objects in the original data set are linked together in a hierarchical tree. We choose one of the most common linkage functions that takes a simple average over all distance within one cluster. We also try alternative linkage functions in the simulation, and find that the results are hardly affected. In this way each estimate starts in its own cluster, and at each step pairs of clusters are merged until a hierarchical tree is formed. As the last step, one can decide where to cut the hierarchical cluster tree to produce the clustering. We cut the tree by specifying the number of clusters $C$, and the algorithm automatically gives a unique clustering. By varying $C$ from 1 to $N$, we numerate all "reasonable" clusterings.[6] Pooling selection or averaging then takes over different choices of $C$.

We distinguish between two AHC screening methods depending on how the distance is computed. First, we base the clustering on the distance between *individual* regressor's coefficient, namely

$$DE_{ij,l} = (\widehat{\beta}_{il} - \widehat{\beta}_{jl})^2.$$

Thus for each regressor we construct a hierarchical cluster tree, and we cut these trees using a common $C$ for all regressors. By varying $C$ from 1 to $N$, we have $M = N$ candidate models.[7] We call this univariate clustering. Second, the clustering is based on the distance between the *vector* of coefficients $\widehat{\beta}_{i\cdot} = (\widehat{\beta}_{i1}, \ldots, \widehat{\beta}_{ik})$ as

$$DE_{ij} = (\widehat{\beta}_{i\cdot} - \widehat{\beta}_{j\cdot})^2.$$

This leads to a common clustering across all regressors. Again, by letting $C$ vary from 1 to $N$, we have $N$ candidate models over which the selection or averaging is performed. We call this multivariate clustering.

There are several advantages of AHC. For example, it does not require input parameters (only the choice of the distance measure and linkage function) nor a priori information of

[6] One could also cut the tree by specifying the cutoff thresholds, and generate different clusterings using various thresholds. The drawback of this procedure is that it may not numerate all possible clusterings or produce replicated clusterings if the thresholds are not appropriately chosen.

[7] In principle, one can generate $N^k$ candidate models by using different $C$ for each regressors. We do not use different $C$ here to avoid a large number of candidate models.

the number of clusters; It is less arbitrary compared to clustering based on artificially chosen thresholds and easy to implement. It also has disadvantages, such as its sensitivity to noise and outliers and no objective function that is directly minimized.

## 6.2 Screening under estimation uncertainty

We have seen that the distance measure plays an important role in the clustering as it determines the similarity of two estimates and further the structure of the tree. It is possible that the distance between adjacent estimates is not accurate due to estimation error. To incorporate the estimation uncertainty, we employ the Bhattacharyya distance. If we assume that the individual estimators are normally distributed, then the Bhattacharyya distance between two coefficient estimates (univariate clustering) can be obtained by

$$DB_{ij,l} = \frac{1}{4}\frac{(\widehat{\beta}_{i,l} - \widehat{\beta}_{j,l})^2}{\widehat{\sigma}_{i,l}^2 + \widehat{\sigma}_{j,l}^2} + \frac{1}{2}\ln\frac{\widehat{\sigma}_{i,l}^2 + \widehat{\sigma}_{j,l}^2}{2\widehat{\sigma}_{i,l}\widehat{\sigma}_{j,l}},$$

where $\widehat{\sigma}_{i,l}^2$ is the estimated variance of $\widehat{\beta}_{i,l}$, and $\widehat{\sigma}_l^2 = (\widehat{\sigma}_{i,l}^2 + \widehat{\sigma}_{j,l}^2)/2$. This distance measure not only captures the deviation between the means and also the variances between two random variables. It can be seen as a generalization of the famous Mahalanobis distance since it allows the variance of two random variables ($\widehat{\beta}_{i,l}$ and $\widehat{\beta}_{j,l}$ in our case) to be different. It degenerates to the Mahalanobis distance (scaled by a constant) when $\widehat{\sigma}_{i,l}^2 = \widehat{\sigma}_{j,l}^2$, and further to the Euclidean distance (scaled by a constant) when both variances equal 1. When we have multiple regressors, the multivariate version of Bhattacharyya distance (multivariate clustering) is

$$DB_{ij} = \frac{1}{8}(\widehat{\beta}_i - \widehat{\beta}_j)'\widehat{\Sigma}^{-1}(\widehat{\beta}_i - \widehat{\beta}_j) + \frac{1}{2}\ln\left(\frac{\det\widehat{\Sigma}}{\sqrt{\det\widehat{\Sigma}_i\,\det\widehat{\Sigma}_i}}\right),$$

where $\widehat{\beta}_i = (\widehat{\beta}_{i,1}, \ldots, \widehat{\beta}_{i,k})'$, $\det\widehat{\Sigma}_i$ is the determinant of the estimated variance of $\widehat{\beta}_i$, and $\widehat{\Sigma} = (\widehat{\Sigma}_i + \widehat{\Sigma}_j)/2$.

The screening procedure reduces the number of candidate models to be selected or averaged. This not only facilitates computation, and also improves estimation efficiency by reducing the number of (weight) parameters. To avoid the danger of making arguments that are sensitive to the preliminary screening procedure, we consider a variety of clustering methods, for example clustering based on ordering and various predetermined thresholds,

and our simulation suggests that our main conclusion is not affected. Here we mainly focus on the AHC because it consistently produces the best results compared to other clustering algorithms.[8] Finally, we need to point out when shrinking the model space there is a tradeoff between efficiency loss and diversification gains from having many different models. This is, however, beyond the scope of this paper.

# 7 Monte Carlo simulation

## 7.1 Data generation process

Our benchmark setup is the static panel data model with coefficients possibly varying over individuals but constant over time

$$y_{it} = \sum_{l=1}^{3} x_{itl}\beta_{il} + \epsilon_{it}, \quad i = 1, \ldots, N; \quad t = 1, \ldots, T,$$

where $x_{it1} = 1$ and the remaining regressors are independently generated from the standard normal distributions. The idiosyncratic error $\epsilon_{it}$ independently follows a normal distribution with mean zero and variance $\sigma_\epsilon^2$, also uncorrelated with regressors. The slope coefficients are allowed to have different grouping patterns. In particular, we consider four cases with different degrees of heterogeneity in coefficients

(1) Homogenous: $\beta_{il} = 1$ for all $i$ and $l$

(2) Weakly heterogenous:

$$\beta_{i1}, \beta_{i2} = \begin{cases} c_1, & i = 1, \ldots, [N/2] \\ c_2, & i = [N/2]+1, \ldots, N, \end{cases} \qquad \beta_{i3} = \begin{cases} c_1, & i = 1, \ldots, [N/3] \\ c_3, & i = [N/3]+1, \ldots, N, \end{cases}$$

where $[N/2]$ denotes the nearest integer value that is smaller than $N/2$.

(3) Strongly heterogenous: For $\beta_{i1}$ and $\beta_{i2}$, they take six different values from $q = (q_1, \ldots, q_6)$ for every 1/6 of the sample, respectively; For $\beta_{i3}$, it also takes six different values from the same vector, but the sample size of each value is different, i.e. $q_1$ for $i = 1, \ldots, [N/8]$, $q_2$ for $i = [N/8]+1, \ldots, [N/4]$, $q_3$ for $i = [N/4]+1, \ldots, [3N/8]$, $q_4$ for $i = [3N/8]+1, \ldots, [N/2]$, $q_5$ for $i = [N/2]+1, \ldots, [5N/8]$, and $q_6$ for $i = [5N/8]+1, \ldots, N$.

[8] Details of other clustering methods and simulation results are available upon request.

(4) Completely heterogenous: $\beta_{il} = 0.1il$ for all $i$ and $l$

We consider three sets of $q$, a benchmark $q_B = [1.0, 1.5, 3.3, 3.0, 2.5, 2.7]$, a larger scale of coefficients $q_L = [1.0, 3.5, 6.3, 4.0, 5.0, 5.4]$ with greater difference in-between, and a smaller scale $q_S = [1.0, 1.2, 2.3, 2.0, 1.5, 1.7]$. We also consider two error structures. First, the variance of errors $\sigma_\epsilon^2$ is identical for all individuals (homoskedastic errors). Second, the variance of errors varies across individuals, and in this case we could use theoretical $R^2$ to determine $\sigma_{\epsilon i}^2$ (heteroskedastic errors across individuals). The sample size varies from $N \in \{5, 10\}$ and $T \in \{10, 40\}$, leading to four combinations of $N$ and $T$. Our simulation results are based on 3000 replications.

## 7.2 Methods

We compare 12 methods: pooled estimation, individual (unrestricted) estimation, two FGLS estimation (FGLS of $\bar{\beta}$ and $\beta$), mixed estimation, and shrinkage estimation, two pretesting methods that select the "best" model based on AIC and BIC, respectively, and four pooling averaging methods discussed in Section 6. We describe these methods in turn as below.[9]

The pooled estimator, denoted by $\widehat{\beta}_{\text{pool}}$, assumes that coefficients are homogenous across individuals and estimate the pooled panel using OLS. The individual estimator $\widehat{\beta}_{\text{ind}}$ can be obtained by OLS estimation of model (2) without coefficient restrictions. The FGLS estimator of $\bar{\beta}$ can be obtained by

$$\widehat{\bar{\beta}}_{\text{FGLS}} = \left( \sum_{i=1}^{N} X_i' \widehat{\Psi}_i^{-1} X_i \right)^{-1} \left( \sum_{i=1}^{N} X_i' \widehat{\Psi}_i^{-1} y_i \right),$$

where $\widehat{\Psi}_i = X_i \widehat{\Delta} X_i' + \widehat{\sigma}_i^2 I_T$ and

$$\widehat{\Delta} = \frac{1}{N-1} \sum_{i=1}^{N} \left( \widehat{\beta}_i - \frac{1}{N} \sum_{i=1}^{N} \widehat{\beta}_i \right) \left( \widehat{\beta}_i - \frac{1}{N} \sum_{i=1}^{N} \widehat{\beta}_i \right)'.$$

It produces a common estimate for all individuals (as the pooled estimator) but incorporates the variation of coefficients; see Hsiao (2003, Sec. 6.2.2.b) for detailed discussions. The FGLS estimator of $\beta$ is computed by

$$\widehat{\beta}_{\text{FGLS}} = \left( \sum_{i=1}^{N} X_i' \widehat{\Sigma}_{uu}^{-1} X_i \right)^{-1} \left( \sum_{i=1}^{N} X_i' \widehat{\Sigma}_{uu}^{-1} y_i \right),$$

[9]The mean group estimator is dominated by the FGLS estimator, and therefore not reported.

where $\widehat{\Sigma}_{uu} = N^{-1}\sum_{i=1}^{N}(y_i - X_i\widehat{\beta}_{\text{ind}})'(y_i - X_i\widehat{\beta}_{\text{ind}})$, and it can be more efficient than the individual OLS estimator but requires a stronger assumption on the error-regressor correlation to achieve consistency. In our case with $\epsilon$ and $X$ independent, GLS of $\beta$ is equivalent to individual OLS. Lee and Griffiths (1979) suggested a mixed estimation of the individual effect as

$$\widehat{\beta}_{i,\text{mix}} = \widehat{\bar{\beta}}_{\text{FGLS}} + \widehat{\Delta}X_i'(X_i\widehat{\Delta}X_i' + \widehat{\sigma}_i^2 I_T)^{-1}(y_i - X_i\widehat{\bar{\beta}}_{\text{FGLS}}),$$

which incorporates the average effect in estimating the individual effect. Finally, the shrinkage estimator is defined in Equation (refeq:shrinkage), which is a special case of the pooling averaging estimator. All pretesting and pooling averaging estimators are based on the preliminary model screening described in Section 6.

We evaluate all methods based on the square loss of coefficients

$$L(w) = \|\widehat{\beta} - \beta\|^2.$$

For the average effect estimators (the pooled estimator and the FGLS estimator of $\bar{\beta}$) we expand the single estimate of a regressor to an $N \times 1$ vector, so that the comparison can be made with other methods.

We emphasize that the purpose of the simulation studies is not to show the dominant superiority of a method. Instead, we try to demonstrate that the performance depends on various factors, thus providing evidence for the theory in Section 4. Also, we provide the applied researchers with some implications, based on the simulation results, that which methods are more likely to give reliable results in a particular situation.

## 7.3 Results

Since there are many varying parameters in the experiment designs, we shall discuss the results in three layers, fixing a part of parameters on each layer. We first present the results based on the benchmark parameterization with heteroskedastic errors ($q_B$, $R^2 = 0.9$). We examine how the clustering methods affect the performance and compare various estimation methods at different degrees of heterogeneity and different sample sizes. Next, we vary the scale of coefficient parameters ($q_L$ and $q_S$), which also shed some light on the effect of parameter heterogeneity, but from a different perspective. Finally, we add more noise by changing the $R^2$ *ceteris paribus*. In addition to the three-layer discussion, we also summarize the results of homoskedastic errors and allowing autoregressive regressors and omitted variables. All following results are based on $A = I_{Nk}$.

***Benchmark parameterization***

Table 1 presents the results using deterministic clustering (clustering ignoring estimation errors) and uncertain clustering (clustering taking into account the estimation errors). The univariate and multivariate AHC produce the same results under deterministic clustering[10], and therefore we only report the univariate case. For uncertain clustering, univariate and multivariate clustering give the different results, and we use the subscript "*uc*" to denote the univariate clustering, and "*mc*" the multivariate version. All the MSEs are normalized by dividing the MSE of the individual estimates. Since the individual estimates (as well as GLS, mixed, pooled, and shrinkage estimates) are not affected by the clustering method, normalization does not affect the comparability. We highlight the minimum numbers of each row in bold. Results of AIC, SAIC, and two GLS are not reported in this table (available upon request) because they are always dominated by other methods.

TABLE 1

First, we note that the MPA and equal weight estimators based on uncertain clustering generally provide lower MSE than those based on deterministic clustering except when $N = 5$ and $T = 40$. This is, however, not the case for the SBIC estimator, whose MSE based on deterministic clustering is often in between the MSEs based on two versions of uncertain clustering, and SBIC under deterministic clustering even performs best in case (2) with $N = 5$ and $T = 40$. Comparing the univariate and multivariate clustering with uncertainty, there is no a strictly dominant method. Univariate clustering produces lower MSE in most cases except in the strongly and completely heterogenous cases with $N = 10$.

Next, we focus on comparing the methods under different degrees of heterogeneity and sample sizes. In case (1) of a homogeneous panel the pooled estimator always performs best as expected, regardless of the clustering procedure. Also, BIC and SBIC estimators are almost as good as the pooled one, especially when $T$ is large. This suggests that the BIC criterion can consistently pick up the true model in this case.

When the panel is characterized by some degree of heterogeneity (case (2) and (3)), the pooling averaging estimators dominate others in all cases. Particularly, MPA produces the lowest MSE in 4 out of 8 cases (50%), followed by Equal (25%) and SBIC (25.5%).

[10]This is not obtained by construction, but due to the choice of DGP coefficients.

Particularly, Equal performs best in both case (2) and (3) under small $N$ and small $T$. As $T$ increases, BIC is more likely to pick up the right model when the degree of heterogeneity is not so strong. Thus we observe that SBIC produces the most accurate estimate in case (2) when $T = 40$. When the panel is characterized by strong heterogeneity, the true model is difficult to identify, and in this case MPA is often the most preferable method. Besides, MPA also outperforms SBIC in the weakly heterogenous panel when $T$ is relatively small, since the consistency of BIC requires a large sample.

Coming to the completely heterogenous case (4), we find that Mixed works uniformly best. When $N = 5$, SHK, as a special case of pooling averaging, is the second best choice, while MPA is the second best when $N = 10$. This observation may seem counterintuitive at the first glance, since one may expect that the individual estimator should perform well in this case. However, the simulation results in fact support our theoretical argument in Section 4 that individual estimation can be inferior to pooling averaging even when all coefficients are completely heterogenous. This is because although the individual estimators are unbiased, they are inefficient, especially under small $T$ or large noise.

In general, we observe that pooling averaging (with different weight choices) performs best in the partially heterogenous panel, and the pooled and mixed are the best choices in the homogenous and completely heterogenous panel, respectively. Among the pooling averaging estimators, MPA exhibits its advantages when the number of individuals is large, which is a typical situation for the macroeconomic data. These results also partly explain the inconsistency between theoretical justification of individual estimation and empirical findings in favor of pooled estimation.

***Alternative slope coefficients***

We then deviate from the benchmark by considering alternative choices of slope coefficients, i.e. $q_L$ and $q_S$. In $q_L$, the difference between difference is larger (a larger degree of coefficient heterogeneity), and the results are given in Table 2.

TABLE 2

We see that Pool and Mixed again produce the most accurate estimators in the homogenous and completely heterogenous, respectively. For the partially heterogenous panel, equal weight averaging is the best method when $N$ is small and Mallows averaging is the

best for a larger $N$. The good performance of Equal under small $N$ may be explained by the fact that the candidate models are rather competing under small $N$, because the difference between the clusterings (and thus the models) based on $C = 1$ and $C = N$ is small. This result, on one hand, supports the pooling averaging procedure, and on the other hand is line with a huge literature on forecasting combination that the equal-weight combination often outperforms the combination using estimated optimal weights when the models are competing (Smith and Wallis, 2009). We also note that when the difference between coefficients is large, treating the coefficient vector as a whole (multivariate clustering) leads to less accurate results. Thus univariate clustering that groups individuals over each regressor separately beats multivariate clustering in all cases.

TABLE 3

Table 3 presents the results under $q_S$ where the difference between coefficients is smaller than the benchmark. We see that the cases where MPA performs best decrease, while SBIC often produces the most accurate estimates in the weakly heterogenous panel. The mixed estimator performs particularly well in the strongly heterogenous panel when $N = 10$ and the completely heterogenous panel. One possible reason that pooling averaging estimators become less favorable in this case is that the coefficient difference is smaller, leading to closer candidate models. This further results in inaccurate weight estimation. Therefore, we observe that equal weights and BIC-based weights are more favored, and the mixed estimator also performs considerably well.

***The effect of noise***

So far, the results are all based on a fixed $R^2 = 0.9$. The theory in Section 4 suggests that more noise will weaken the advantages of pooling averaging estimates, but support the use of the pooled estimator. To verify this argument, we examine the effect of adding more noise to the model by decreasing $R^2$. We consider two choices of $R^2$, the moderate fitness with $R^2 = 0.75$ and relatively poor fitness with $R^2 = 0.5$. The results are given in Table 4. As above, the estimates that never perform best (AIC, SAIC, BIC, SHK, and FGLS of $\beta$) are omitted.

TABLE 4

We see that increasing errors does not affect the performance of the pooled estimator in the homogeneous panel. In fact, the pooled estimator is more favorable when there are more errors, because it employs the whole sample and estimates the least parameters. When it comes to the partially heterogenous panel (cases (2) and (3)), MPA still performs best most often with $R^2 = 0.75$, followed by the FGLS estimator of $\bar{\beta}$, and then the mixed estimator. If we further decrease the $R^2$ to 0.5, we see that FGLS of $\bar{\beta}$ and Pool are especially favorable in the heterogenous panel. These two estimators both give an average effect estimate, and they possess the minimum MSE because they are the most efficient among others and the efficiency gains dominate the bias in this case.[11] We also consider homoskedastic errors by setting $\sigma_\epsilon^2 = 0.3$ (less noise) and $\sigma_\epsilon^2 = 0.7$ (more noise). The comparison is similar to heteroskedastic errors.

Finally, we briefly consider two other extensions, both analyzed in the benchmark setting: autoregressive regressors and models with omitted variables. Autocorrelation is introduced through an AR(1) process, and in the omitted variable case regressors are generated with nonzero correlations. The simulation results are largely similar therefore not reported. In particular, the pooling averaging estimators are generally preferred in the partially heterogenous panel, among which MPA and Equal often produce the most accurate estimators, and SBIC performs best in large $T$. The mixed estimator also performs considerably well and sometimes the best in the omitted variable cases. The performance of the competing methods depends on various factors in the same way as discussed above.

# 8 Explain and predict sovereign credit risk

We apply the proposed pooling averaging estimator to examine the determinants of sovereign credit default swap (CDS) spreads of a panel of countries. A CDS contract is an insurance contract that protects the buyer from the credit event, e.g. a load default. Its spread, expressed in basis points, is the insurance premium that buyers have to pay, and thus reflects the credit risk. Recently, an increasing number of studies have tried to associate the CDS spreads with various macroeconomic fundamentals, e.g. Dieckmann and Plank (2011); Longstaff et al. (2011); Beirne and Fratzscher (2013); Aizenman et al. (2013). Despite the

[11] More values of $R^2$ are experimented, and they show that the performance of averaging estimators is poor when $R^2$ is less than 0.6.

availability of a cross-country panel, most of these studies analyze individual countries separately. The individual-country analysis shows that there is indeed some common pattern in the processes of sovereign CDS spread across countries. It thus raises a question whether a determinant has similar impacts on CDS spreads in different sovereigns, and whether it is useful to pool some of the countries when estimating the sovereign CDS spread panel.

We use the data set from Longstaff et al. (2011) and examine the determinants of sovereign CDS spreads of 18 countries.[12] It contains the monthly data of five-year sovereign CDS and selective financial indicators of macroeconomic fundamentals (local variables and global variables). The local variables include local stock market returns, changes in local exchange rates (*fxrate*), and changes in foreign currency reserves (*fxres*). The global variables include the U.S. stock market returns (*gmkt*), treasury yields (*trsy*), investment-grade corporate bond spreads (*ig*), high-yield corporate bond spreads (*hy*), equity premium (*eqp*), volatility risk premium (*volp*), term premium (*termp*), equity flows (*ef*), and bond flows (*bf*). To have a balance panel, we use the sample starting from January 2003 to January 2010. Preliminary unit root tests shows that the CDS spreads of all countries are clearly autocorrelated, and the first-order difference eliminates the autocorrelation to a large extent. Therefore, we take the first-order difference of each series, and the forecast is then based on the differenced CDS spreads.

## 8.1 Estimation results

We compute the MPA estimates with and without taking account the estimation uncertainty. Both MPA estimates are based on univariate clustering in the screening step. Since two screening methods lead to similar estimated coefficients, we only report the estimates based on the uncertain clustering in Table 5, but we shall compare their prediction accuracy in the next subsection. Like other model averaging estimators, MPA does not provide a variance estimator, and the statistical inference of significance is not feasible here. However, since all variables are normalized, we can compare the effects of different determinants within a country, and also the effect of a variable between countries.

TABLE 5

[12] We exclude 8 countries from Longstaff et al. (2011)'s data set, i.e. Pakistan, Panama, Isreal, Peru, Qatar, S.Africa, Ukrain, and Venezuela, due to a large proportion of missing data in these countries.

We first examine the local variables. MPA reports a negative effect of the local stock return in all countries, while the individual estimation shows a positive effect in Bulgaria.[13] The negative effect given by MPA is in line with the theory that the good local economic performance is associated with less risk of default. This effect is relatively small in Bulgaria, Chile, and Romania (absolute value around 0.12), but strong in Brazil, Hungary, Japan, Mexico, Phillipines, Russia, and Turkey (absolute value at least 0.20), in line with the significance shown by Longstaff et al. (2011). In contrast to Longstaff et al. (2011), MPA reports a positive effect of foreign exchange rate in all countries except China, while Longstaff et al. (2011) reported a negative effect in 8 countries. The MPA estimates are more intuitive because the depreciation of sovereign's currency is expected to be associated with an increase in the sovereign credit spreads. The coefficient for the change in foreign currency reserves is negative in 11 countries and positive in 7 countries, and the negative effect is generally strong. This is also in line with the literature.

As for the global variables, we find that the U.S. stock return is the most salient determinant, explaining the most proportion of the credit spread variation. It has a negative impact on all sovereigns' credit spreads. Among 18 countries, the credit spreads of Turkey and Poland are least affected by the U.S. stock return, while Bulgaria, Korea, and Malaysia are most influenced. It appears that the countries with a weak local-stock-return effect are generally associated with a strong effect of U.S. stock returns, and they are typically small economies. On the contrary, some large economies, such as Brazil, Mexico, Japan, and Russia, are largely affected by both their own stock markets and the global stock market. Turkey is the only country whose own stock market effect is larger than the global market effect. Such heterogeneity effects cannot be captured by the pooled estimator, and the OLS individual estimator reports a counterintuitive positive effect of the U.S. stock return on Turkey's credit spreads. Another important global variable is the U.S. high-yield spread. It imposes a positive impact on all sovereign CDS spreads, and the impact is particularly strong in eastern European countries, such as Croatia, Poland, Romania, and Russia. These two global variables have the most explanatory power of all of the variables in the regression. Also robust (in the sense of the same sign for all countries) but slightly less powerful global variables include the volatility risk premium, which has a negative

[13]Our individual estimation results are slightly different from those in Longstaff et al. (2011) since the starting point of the sample is different.

coefficient for all countries.

## 8.2 Out-of-sample prediction

To further compare alternative methods, we consider the pseudo out-of-sample prediction. We divide the entire time period into two sub-samples: the first 90% of the sample period are used to estimate the coefficients, and the remaining are used for prediction and evaluation. The prediction is made based on the estimated coefficients and out-of-sample values of explanatory variables. We evaluate the predictors using the average absolute deviation (AAD) from the true value and the root mean square prediction error (RMSPE), both of which are averaged across 18 countries.

TABLE 6

Table 6 presents the out-of-sample prediction performance of alternative methods. The subscript "*det*" indicates the deterministically screening (without considering estimation uncertainty), and "*unc*" represents screening that takes account the uncertainty. All numbers are normalized with respect to the individual estimation. We first compare all methods using the panel of all 18 countries. The first two columns show that SBIC (and BIC) and Pool perform equally well. The good performance of Pool can be partly explained by possible misspecification of the model. This is supported by a large variation of individual regression $R^2$, from a minimum 0.45 to a maximum 0.82. Also, the CDS spreads of several countries are characterized by heteroskedasticity (the Engel's ARCH test rejects the null hypothesis of no ARCH effects for 5 countries). Under a potentially large degree of model misspecification, the efficiency lost of having more parameters (pooling less countries) can be prominent. Therefore, the pooled estimators demonstrates its superiority because the efficiency gains from pooling dominates the bias. Since the pooled estimator has a dominant BIC, BIC and SBIC produce the same predictor as Pool.

Next, we compare the predictors in two subsamples, Latin American countries (columns 3 and 4) and Asian countries (columns 5 and 6). Individual estimation suggests that four countries in the Latin America sample have largely different coefficients. For example, Brazil has the second largest effect of *lstock* (absolute value 0.3233) in all 18 countries, while this effect of Chile is the second least (absolute value 0.0170); The effect of *gstock* is

the fourth strongest in Columbia (absolute value 0.5710), but especially weak in Mexico (absolute value 0.2727). In this subsample, we find that $\text{MPA}_{unc}$ produces the best predictor in terms of both AAD and RMSPE. It beats Pool because coefficient heterogeneity is more significant, and completely ignoring it leads to an inaccurate predictor. This result is in line with our simulation finding that $\text{MPA}_{unc}$ performs considerably well in the strongly heterogenous panel. Different from the Latin American sample, six Asian countries share more similar patterns. For example, individual estimation suggests that most of coefficients of Japan and Philippines are similar, and Korea and Malaysia also share great similarity. In this subsample, AIC, SBIC, and Pooled produce the best predictor.

Finally, to examine how each method perform when the model is well fitted, we focus on the countries with the individual estimation adjusted $R^2$ larger than 0.75.[14] The last two columns show that MPA predictor has the least RMSPE in this subsample, while SBIC and pooled estimator has the least AAD, but the results of three estimation methods are very close. We also consider the cases where we use a shorter period for estimation (and correspondingly a longer period for prediction). We find that the predictor based on the pooled model outperforms others, because there seems a shock at round 80% of the period in most of sovereign CDS spread series.

In general, the out-of-sample prediction analysis suggests that the proposed pooling averaging produces good predictors of sovereign CDS spreads. The performance of the MPA predictor depends on the degree of heterogeneity and the model fitness. Besides, we see that the predictor based on the pooled model performs considerably well in most of cases. This is partly because it is more flexible to the future changes of the dynamic process, and also because the pooled predictor suffers less from efficiency lost due to model misspecification. This also explains why most empirical studies advocate the pooling estimation in the real data analysis.

# 9 Implications and discussions

Based on our theoretical and numerical analysis, we offer practitioners with general guidance on how to determine which estimator to use for heterogenous panel in different situations. Considering the fact that the performance of different estimators have been shown

[14]We choose 0.75 because our simulation studies suggest that MPA has good performance when $R^2$ close or larger than 0.75.

to vary significantly over applications, this guidance with checkable conditions provides useful tools for choosing an appropriate estimator in practice.

As the first step, it seems plausible to estimate individual time series separately, and compute the coefficient estimates and $R^2$ for each regression. Estimation of individual regression can be used as a starting point because the coefficient estimates are consistent, although may be inefficient. If most individual regressions produce low $R^2$, the pooled estimator could be a safe choice. On the contrary, if large $R^2$ values are observed in most regressions and coefficient estimates are completely different over individuals, the mixed estimator is often recommended. When we are in a situation with large $R^2$ but individual coefficient estimates are "partially" different in the sense that some individuals share similar coefficients, the pooling averaging estimator can be a good choice. In case with the moderate or large number of individual $N$ and regressors $k$, averaging over the entire model space is not feasible, and then a preliminary screening is needed. To determine which screening method to use, one can check the estimated variance of individual coefficients. If the variance is negligible, one can consider implementing screening without taking into account the estimator error. However, if the coefficient variance is not negligible, then uncertainty screening for MPA or equal-weight averaging are likely to be more reliable. Besides, the sample size should also be considered when choosing methods. For example, if the time dimension is large, the mixed or individual estimation approaches are likely to produce accurate results. Regarding the individual dimension $N$, when we have a small $N$ panel, a fixed weight averaging is often more preferable than the pooling averaging using estimated weights.

This practical guidance mainly focuses on choosing among the 12 candidate methods considered in the simulation and application. However, we do not rule out other estimators, especially if there is other interest than the mean square error. For example, if the long-run and short-run distinction is of interest, the pooled mean group estimation (Pesaran et al., 1999) that allows different treatment of two types of parameters is especially useful. If the individuals' group pattern is of interest, it is a good idea to consider grouped estimator, such as Bonhomme and Manresa (2012) and Lin and Ng (2012).

To summarize, this paper contributes to the literature by proposing a novel optimal pooling averaging method and by providing some practical guidance for empirical researchers which estimator to use in different situations. To address the issue of a possibly large model space, we also propose a preliminary model screening procedure, which not

only reduces computation burden and also improves the estimation efficiency. We compare the finite sample performance of 13 methods under extensive experiment designs, and also apply our method to examine which determinants are salient in affecting the sovereign CDS spreads using an 18 country panel. We show that the estimates produced by pooling averaging method capture the heterogeneity that is ignored by the pooled estimator, and they are generally more intuitive than the individual estimates.

# References


J. Aizenman, M. Hutchison, and Y. Jinjarak. What is the risk of European sovereign debt defaults? Fiscal space, CDS spreads and market pricing of risk. *Journal of International Money and Finance*, 34:37–59, 2013.

B. Baltagi, G. Bresson, and A. Pirotte. To pool or not to pool? In L. Mátyás and P. Sevestre, editors, *The Econometrics of Panel Data*, Advanced Studies in Theoretical and Applied Econometrics, pages 517–546. Springer Berlin Heidelberg, 2008.

B. H. Baltagi and J. M. Griffin. Pooled estimators vs. their heterogeneous counterparts in the context of dynamic demand for gasoline. *Journal of Econometrics*, 77:303–327, 1997.

B. H. Baltagi, J. M. Griffin, and W. Xiong. To pool or not to pool: Homogeneous versus hetergeneous estimations applied to cigarette demand. *The Review of Economics and Statistics*, 82:117–126, 2000.

A. Barron, B. Lucien, and P. Massart. *Probability Theory and Related Fields*, 113:301–413, 1999.

J. Beirne and M. Fratzscher. The pricing of sovereign risk and contagion during the European sovereign debt crisis. *Journal of International Money and Finance*, 34:60–82, 2013.

C. A. Bester and C. B. Hansen. Grouped effects estimators in fixed effects models. *Journal of Econometrics*, forthcoming.

S. Bonhomme and E. Manresa. Grouped patterns of heterogeneity in panel data. Working Papers wp2012-1208, CEMFI, 2012.

M. Browning and J. Carro. Heterogeneity and microeconometrics modeling. In R. Blundell, W. Newey, and T. Persson, editors, *Advances in Economics and Econometrics*, volume 3, pages 47–74. Cambridge University Press, 2007.

S. T. Buckland, K. P. Burnham, and N. H. Augustin. Model selection: An integral part of inference. *Biometrics*, 53:603–618, 1997.

D. Danilov and J. R. Magnus. On the harm that ignoring pretesting can cause. *Journal of Econometrics*, 122:27–46, 2004.

S. Dieckmann and T. Plank. Default risk of advanced economies: An empirical analysis of credit default swaps during the financial crisis. *Review of Finance*, 0:1–32, 2011.

S. N. Durlauf, A. Kourtellos, and A. Minkin. The local Solow growth model. *European Economic Review*, 45:928–940, 2001.

B. E. Hansen. Least squares model averaging. *Econometrica*, 75:1175–1189, 2007.

A. J. Hoogstrate, F. C. Palm, and G. A. Pfann. Pooling in dynamic panel-data models: An application to forecasting gdp growth rates. *Journal of Business and Economic Statistics*, 18:274–283, 2000.

C. Hsiao. *Analysis of Panel Data*. Cambridge University Press, New York, 2003.

C. Hsiao and H. Pesaran. Random coefficient models. In L. Mátyás and P. Sevestre, editors, *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*, chapter 6, pages 185–213. Springer Publishers, Netherlands, 2008.

S. Jin and L. Su. Nonparametric tests for poolability in panel data models with cross section dependence. *Econometric Reviews*, 32:469–512, 2013.

A. Juditsky and A. Nemirovski. Functional aggregation for nonparametric regression. *The Annals of Statistics*, 28:681–712, 2000.

T. Juhl and O. Lugovskyy. A test for slope heterogeneity in fixed effects models. *working paper*, 2010.

L.-F. Lee and W. E. Griffiths. The prior likelihood and best linear unbiased prediction in stochastic coefficient linear models. Working Papers in Econometrics and Applied Statistics No. 1., University of New England, 1979.

C.-C. Lin and S. Ng. Estimation of panel data models with parameter heterogeneity when group membership is unknown. *Journal of Econometric Methods*, 1:42–55, 2012.

F. A. Longstaff, J. Pan, L. H. Pedersen, and K. J. Singleton. How sovereign is sovereign credit risk? *American Economic Journal: Macroeconomics*, 3:75–103, 2011.

G. S. Maddala, R. P. Trost, H. Li, and F. Joutz. Estimation of short-run and long-run elasticities of energy demand from panel data using shrinkage estimator. *Journal of Business and Economic Statistics*, 15:90–100, 1997.

G. S. Maddala, H. Li, and V. K. Srivastava. A comparative study of different shrinkage estimators for panel data models. *Annals of Economics and Finance*, 2:1–30, 2001.

M. H. Pesaran. Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 74:967–1012, 2006.

M. H. Pesaran and R. Smith. Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics*, 68:79–113, 1995.

M. H. Pesaran and E. Tosetti. Large panels with common factors and spatial correlation. *Journal of Econometrics*, 161:182–202, 2011.

M. H. Pesaran and T. Yamagata. Testing slope homogeneity in large panels. *Journal of Econometrics*, 142:50–93, 2008.

M. H. Pesaran, R. G. Pierse, and M. S. Kumar. Econometric analysis of aggregation in the context of linear prediction models. *Econometrica*, 57:861–888, 1989.

M. H. Pesaran, Y. Shin, and R. P. Smith. Pooled mean group estimation of dynamic heterogeneous panels. *Journal of the American Statistical Association*, 94:621–634, 1999.

P. C. Phillips and D. Sul. Transition modeling and econometric convergence test. *Econometrica*, 75:1771–1855, 2007.

J. Smith and K. F. Wallis. A simple explanation of the forecast combination puzzle. *Oxford Bulleting of Economics and Statistics*, 71:331–355, 2009.

J. H. Stock and M. W. Watson. Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, 23:405–430, 2004.

L. Su and Q. Chen. Testing homogeneity in panel data models with interactive fixed effects. *Econometric Theory*, forthcoming, 2013.

L. Su, Z. Shi, and P. C. Phillips. Identifying latent structures in panel data. Cowles Foundation Discussion Papers 1965, Cowles Foundation for Research in Economics, Yale University, 2014.

P. A. V. B. Swamy. Efficient inference in a random coefficient regression model. *Econometrica*, 38:311–323, 1970.

A. L. Vasnev, J. R. M. Gerda Claeskens, and W. Wang. A simple theoretical explanation of the forecast combination puzzle. *Working paper*, 2014.

A. T. K. Wan, X. Zhang, and G. Zou. Least squares model averaging by mallows criterion. *Journal of Econometrics*, 156:277–283, 2010.

J. Westerlund and W. Hess. A new poolability test for cointegrated panels. *Journal of Applied Econometrics*, 26:56–88, 2011.

J. M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, 2002.

Y. Yang. Adaptive regression by mixing. *Journal of the American Statistical Association*, 96:574–588, 2001.

Y. Yang. Regression with multiple candidate models: selecting or mixing? *Statistica Sinica*, 13:783–809, 2003.

Z. Yuan and Y. Yang. Combining linear regression models: When and how? *Journal of the American Statistical Association*, 100:1202–1214, 2005.

A. Zellner. On the aggregation problem: A new approach to a troublesome problem. In K. Fox, J. Sengupta, and G. Narasimham, editors, *Economic Models, Estimation and Risk Programming: Essays in Honor of Gerhard Tintner*, volume 15 of *Lecture Notes in Operations Research and Mathematical Economics*, pages 365–374. Springer Berlin Heidelberg, 1969.

# Appendix

## A1: Proof of Theorem 5.1

From (6), it is straightforward to show that

$$\begin{aligned} \mathrm{E}\{L_A(w)\} &= \mathrm{E}\{\|\widehat{\beta}(w)-\beta\|_A^2\} = \mathrm{E}\{\|P(w)\widehat{\beta}-\beta\|_A^2\} \\ &= \mathrm{E}\{\|P(w)\widehat{\beta}\|_A^2\} + \|\beta\|_A^2 - 2\beta' P'(w)A\beta \end{aligned}$$

and

$$\begin{aligned} \mathrm{E}\{\mathcal{C}_A(w)\} &= \mathrm{E}\{\|P(w)\widehat{\beta}-\widehat{\beta}\|^2\} + 2\mathrm{tr}\{P'(w)AV\} - \mathrm{tr}(AV) \\ &= \mathrm{E}\{\|P(w)\widehat{\beta}\|_A^2\} + \mathrm{E}\{\|\widehat{\beta}\|_A^2\} - \mathrm{tr}(AV) \\ &\quad -2\mathrm{E}\{\widehat{\beta}' P'(w)A\widehat{\beta}\} + 2\mathrm{tr}\{P'(w)AV\} \\ &= \mathrm{E}\{\|P(w)\widehat{\beta}\|_A^2\} + \|\beta\|_A^2 - 2\beta' P'(w)A\beta. \end{aligned} \tag{28}$$

So $L_A(w)$ and $\mathcal{C}_A(w)$ have the same expectations.

## A2: Proof of Theorem 5.2

Let $\widetilde{w} = \arg\min_{w\in\mathcal{W}} R_A(w)$, and the corresponding risk is

$$R_A(\widetilde{w}) = \inf_{w\in\mathcal{W}} R_A(w), \tag{29}$$

which means that $\widetilde{w}$ is theoretically (non-random) optimal weight vector.

We rewrite $\mathcal{C}_A(w)$ as

$$\begin{aligned} \mathcal{C}_A(w) &= \|P(w)\widehat{\beta}-\widehat{\beta}\|_A^2 + 2\mathrm{tr}\{P'(w)AV\} - \|\widehat{\beta}-\beta\|_A^2 \\ &= \|P(w)\widehat{\beta}-\beta-(\widehat{\beta}-\beta)\|_A^2 + 2\mathrm{tr}\{P'(w)AV\} - \|\widehat{\beta}-\beta\|_A^2 \\ &= \|P(w)\widehat{\beta}-\beta\|_A^2 + 2\mathrm{tr}\{P'(w)AV\} - 2\{P(w)\widehat{\beta}-\beta\}'A(\widehat{\beta}-\beta). \end{aligned}$$

Letting

$$a(w) = 2\mathrm{tr}\{P'(w)AV\} - 2\{P(w)\widehat{\beta}-\beta\}'A(\widehat{\beta}-\beta), 3 \tag{30}$$

we have

$$\mathcal{C}_A(w) = L_A(w) + a(w). \tag{31}$$

It is straightforward to show that for any non-random weight vector $w$,

$$\mathrm{E}\{a(w)\} = 2\mathrm{tr}\{P'(w)AV\} - 2\mathrm{E}[\{P(w)\widehat{\beta} - \beta\}'A(\widehat{\beta} - \beta)] = 0. \tag{32}$$

It follows from (20) and (31) that

$$L_A(\widehat{w}) = \mathcal{C}_A(\widehat{w}) - a(\widehat{w}) \leq \mathcal{C}_A(\widetilde{w}) - a(\widehat{w}) = L_A(\widetilde{w}) + a(\widetilde{w}) - a(\widehat{w}). \tag{33}$$

Taking expectations of both sides of above formula, by (29) and (32) we have

$$\begin{aligned}
R_A(\widehat{w}) &\leq \mathrm{E}\{L_A(\widetilde{w})\} + \mathrm{E}\{a(\widetilde{w})\} - \mathrm{E}\{a(\widehat{w})\} \\
&= \inf_{w\in\mathcal{W}} R_A(w) - \mathrm{E}\{a(\widehat{w})\} \\
&= \inf_{w\in\mathcal{W}} R_A(w) - 2\mathrm{Etr}\{P'(\widehat{w})AV\} + \mathrm{E}2\{P(\widehat{w})\widehat{\beta} - \beta\}'A(\widehat{\beta} - \beta) \\
&\leq \inf_{w\in\mathcal{W}} R_A(w) - 2\mathrm{Etr}\{P'(\widehat{w})AV\} + \mathrm{E}\{c\|P(\widehat{w})\widehat{\beta} - \beta\|_A^2 + c^{-1}\|\widehat{\beta} - \beta\|_A^2\} \\
&= \inf_{w\in\mathcal{W}} R_A(w) - 2\mathrm{Etr}\{P'(\widehat{w})AV\} + cR_A(\widehat{w}) + c^{-1}\mathrm{E}\|\widehat{\beta} - \beta\|_A^2 \\
&= \inf_{w\in\mathcal{W}} R_A(w) - 2\mathrm{Etr}\{P'(\widehat{w})AV\} + cR_A(\widehat{w}) + c^{-1}\mathrm{tr}(AV),
\end{aligned} \tag{34}$$

where $c$ is a constant belonging to $(0, 1)$. Thus,

$$R_A(\widehat{w}) \leq \frac{1}{1-c}\inf_{w\in\mathcal{W}} R_A(w) + \frac{1}{1-c}(\frac{1}{c}\mathrm{tr}(AV) - 2\mathrm{E}[\mathrm{tr}\{P'(\widehat{w})AV\}]). \tag{35}$$

## A3: Proof of Corollary 5.1

To derive the risk bounds for two specific choices of $A$, we define $B_m$ be a projection matrix associated with model $m$ as

$$B_m = (X'X)^{-1/2}R_m'(R_m(X'X)^{-1}R_m')^{-1}R_m(X'X)^{-1/2},$$

and $\mathbf{B(w) = \sum_{m=1}^{M} w_m B_m}$. It can be verified that $B_m$ is a symmetric and idempotent matrix and thus

$$\mathcal{I}_2(B_m) \leq 1, \quad \mathcal{I}_2\{B(w)\} \leq 1, \quad \mathrm{rank}(B_m) = Nk. \tag{36}$$

Then, we can rewrite the risk bound of (22) in terms of $B(\widehat{w})$ as

$$\begin{aligned}
&R_A(\widehat{w}) \\
&\leq \frac{1}{1-c}\inf_{w\in\mathcal{W}} R_A(w) + \frac{1}{1-c}\{\frac{1}{c}\mathrm{tr}(AV) - 2\mathrm{E}(\mathrm{tr}[\{I_{Nk} - (X'X)^{1/2}B(\widehat{w})(X'X)^{-1/2}\}AV])\}
\end{aligned}$$

$$\leq \frac{1}{1-c}\inf_{w\in\mathcal{W}} R_A(w) + \frac{1-2c}{c(1-c)}\mathrm{tr}(AV) + \frac{2}{1-c}\mathrm{E}[\mathrm{tr}\{B(\widehat{w})(X'X)^{-1/2}AV(X'X)^{1/2}\}]. \quad (37)$$

We first consider the case of $A = I_{Nk}$. Using (36), the second term of (37) satisfies

$$\begin{aligned}
\mathrm{tr}(AV) &= \mathrm{tr}(V) = \sum_{i=1}^{N}\mathrm{tr}(\sigma_i^2 Q_i^{-1}) \\
&= T^{-1}\sum_{i=1}^{N}\mathrm{tr}\{\sigma_i^2(T^{-1}X_i'X_i)^{-1}\} \\
&\leq T^{-1}\sum_{i=1}^{N}\sigma_i^2\mathrm{rank}(X_i)\mathcal{I}_2\{(T^{-1}X_i'X_i)^{-1}\} \\
&\leq T^{-1}Nk\sigma_{\max}^2 c_1^{-1} \qquad (38)
\end{aligned}$$

and the third term of (37) satisfies

$$\begin{aligned}
&\mathrm{E}[\mathrm{tr}\{B(\widehat{w})(X'X)^{-1/2}AV(X'X)^{1/2}\}] = \mathrm{E}[\mathrm{tr}\{B(\widehat{w})(X'X)^{-1/2}V(X'X)^{1/2}\}] \\
&= 2^{-1}\mathrm{E}[\mathrm{tr}\{B(\widehat{w})(X'X)^{-1/2}V(X'X)^{1/2} + (X'X)^{1/2}V(X'X)^{-1/2}B(\widehat{w})\}] \\
&\leq 2^{-1}\mathrm{E}[\mathrm{rank}\{B(\widehat{w})(X'X)^{-1/2}V(X'X)^{1/2} + (X'X)^{1/2}V(X'X)^{-1/2}B(\widehat{w})\} \\
&\quad \times\mathcal{I}_2\{B(\widehat{w})(X'X)^{-1/2}V(X'X)^{1/2} + (X'X)^{1/2}V(X'X)^{-1/2}B(\widehat{w})\}] \\
&\leq 2\mathrm{E}[\mathrm{rank}\{B(\widehat{w})(X'X)^{-1/2}V(X'X)^{1/2}\}\mathcal{I}_2\{B(\widehat{w})(X'X)^{-1/2}V(X'X)^{1/2}\}] \\
&\leq 2Nk\mathrm{E}[\mathcal{I}_2\{B(\widehat{w})(-X'X)^{1/2}V(X'X)^{1/2}\}] \\
&\leq 2Nk\mathcal{I}_2\{(X'X)^{-1/2}V(X'X)^{1/2}\} \\
&= 2T^{-1}Nk\max_{i\in\{1,\ldots,N\}}\sigma_i^2\mathcal{I}_2\{(T^{-1}X_i'X_i)^{-1}\} \\
&= 2T^{-1}Nk\sigma_{\max}^2 c_1^{-1}. \qquad (39)
\end{aligned}$$

Plugging the inequalities (38) and (39) into (22), we can obtain the risk bound for $A = I_{Nk}$ as in (23). For the case of $A = X'X$, we have

$$\mathrm{tr}(AV) = \mathrm{tr}(X'XV) = k\sum_{i=1}^{N}\sigma_i^2 = Nk\sigma_{\max}^2, \qquad (40)$$

and

$$\begin{aligned}
&\mathrm{E}[\mathrm{tr}\{B(\widehat{w})(X'X)^{-1/2}AV(X'X)^{1/2}\}] = \mathrm{E}[\mathrm{tr}\{B(\widehat{w})(X'X)^{1/2}V(X'X)^{1/2}\}] \\
&\leq 2Nk\mathcal{I}_2\{(X'X)^{1/2}V(X'X)^{-1/2}\} = 2Nk\max_{i\in\{1,\ldots,N\}}\sigma_i^2 = 2Nk\sigma_{\max}^2. \qquad (41)
\end{aligned}$$

Plugging (40) and (41) into (37) leads to the risk bound given in (24).

## A4: Proof of Theorem 5.3

Observe that

$$\begin{aligned} R_A(w) &= \mathrm{E}\{L_A(w)\} = \mathrm{E}\{\|\widehat{\beta}(w) - \beta\|_A^2\} = \mathrm{E}\{\|P(w)\widehat{\beta} - \beta\|_A^2\} \\ &= \|P(w)\beta - \beta\|_A^2 + \mathrm{tr}\{P(w)AP'(w)V\} \\ &= \|P(w)\widehat{\beta} - \beta - P(w)(\widehat{\beta} - \beta)\|_A^2 + \mathrm{tr}\{P(w)AP'(w)V\} \\ &= L_A(w) + \|P(w)(\widehat{\beta} - \beta)\|_A^2 + 2\beta' AP(w)(\widehat{\beta} - \beta) \\ &\quad -2\widehat{\beta}' P(w)AP(w)(\widehat{\beta} - \beta) + \mathrm{tr}\{P(w)AP'(w)V\} \\ &\equiv L_A(w) + \Xi(w). \end{aligned} \tag{42}$$

From (31), (42), and the proof of Theorem 1' of Wan et al. (2010), to prove (26), we need only verify that

$$\sup_{w\in\mathcal{W}} \frac{|a(w)|}{R_A(w)} = o_p(1) \tag{43}$$

where $a(w)$ is defined in (30), and

$$\sup_{w\in\mathcal{W}} \frac{|\Xi(w)|}{R_A(w)} = o_p(1). \tag{44}$$

Using the proving steps of (38) and (39), from (36) and $T^{-1}X'X \to \Psi$, we have

$$\begin{aligned} &\sup_{w\in\mathcal{W}} |\mathrm{tr}\{P'(w)AV\}| \le \sum\nolimits_{m=1}^{M} |\mathrm{tr}(P_m' AV)| \\ &= \sum\nolimits_{m=1}^{M} [|\mathrm{tr}(AV)| + |\mathrm{tr}\{(X'X)^{1/2} B_m (X'X)^{-1/2} AV\}|] \\ &\le \sum\nolimits_{m=1}^{M} \{\lambda(A)\mathrm{tr}(V) + 2\lambda(A)\lambda((X'X)^{1/2})\lambda((X'X)^{-1/2})\mathrm{rank}(B_m)\lambda(V)\} \\ &= \lambda(A)O(MT^{-1}). \end{aligned} \tag{45}$$

From $X'u = O_p(n^{1/2})$ and $T^{-1}X'X \to \Psi$, we have $\widehat{\beta} - \beta = O_p(T^{-1/2})$, which, along with $T^{-1}X'X \to \Psi$, implies

$$\begin{aligned} &\sup_{w\in\mathcal{W}} |\{P(w)\widehat{\beta} - \beta\}' A(\widehat{\beta} - \beta)| \le \|\widehat{\beta} - \beta\|\lambda(A) \sup_{w\in\mathcal{W}} \|P(w)\widehat{\beta} - \beta\| \\ &\le \|\widehat{\beta} - \beta\|\lambda(A) \sum\nolimits_{m=1}^{M} \|P_m\widehat{\beta} - \beta\| \\ &= O_p(MT^{-1/2})\lambda(A). \end{aligned} \tag{46}$$

Now, from (45), (46) and the condition (25), we can obtain (43). Similarly, we can also obtain (44). This completes the proof.

Table 1: MSE comparison: Benchmark

| | | Determinstic clustering | | | Uncertain clustering | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | DGP | MPA | Equal | SBIC | MPA$_{uc}$ | MPA$_{mc}$ | Equal$_{uc}$ | Equal$_{mc}$ | SBIC$_{uc}$ | SBIC$_{mc}$ | Mixed | Pool | SHK |
| | 1 | 0.2428 | 0.3715 | 0.1828 | 0.3715 | 0.3064 | 0.6528 | 0.5989 | 0.1845 | 0.1744 | 0.6393 | **0.1611** | 0.7562 |
| $N = 5$ | 2 | 0.9604 | 0.8738 | 1.0570 | 0.7581 | 0.8179 | **0.6718** | 0.7747 | 0.8614 | 1.0839 | 0.8077 | 2.8287 | 0.9557 |
| $T = 15$ | 3 | 0.8907 | 0.7349 | 1.0282 | 0.6188 | 0.6594 | **0.5905** | 0.6360 | 0.6693 | 0.8914 | 0.7705 | 1.5603 | 0.9220 |
| | 4 | 1.1551 | 7.4286 | 1.1708 | 1.1392 | 1.0246 | 3.9303 | 3.9361 | 1.3245 | 1.1630 | **0.7580** | 24.0241 | 0.9941 |
| | 1 | 0.3742 | 0.6710 | 0.1927 | 0.3736 | 0.3140 | 0.6704 | 0.6203 | 0.1921 | 0.1884 | 0.6535 | **0.1843** | 0.9168 |
| $N = 5$ | 2 | 0.7222 | 0.9286 | **0.6759** | 0.6817 | 0.7924 | 0.9008 | 1.3127 | 0.6942 | 0.8740 | 0.8506 | 8.5772 | 0.9938 |
| $T = 40$ | 3 | 0.7888 | 0.7927 | 0.8824 | **0.7613** | 0.7746 | 0.7617 | 0.9189 | 0.8681 | 1.0168 | 0.8572 | 4.4980 | 0.9881 |
| | 4 | 1.1246 | 8.2212 | 1.0626 | 1.0363 | 1.0016 | 11.4912 | 11.5407 | 1.0428 | 1.0169 | **0.7646** | 73.4702 | 0.9991 |
| | 1 | 0.3011 | 0.7696 | 0.0838 | 0.2979 | 0.2088 | 0.7647 | 0.6812 | 0.0835 | 0.0812 | 0.4358 | **0.0795** | 0.6635 |
| $N = 10$ | 2 | 0.7198 | 0.7799 | 0.7868 | **0.5488** | 0.6237 | 0.6103 | 0.6262 | 0.5883 | 0.8849 | 0.5821 | 2.1644 | 0.9141 |
| $T = 15$ | 3 | 0.7315 | 0.7892 | 0.8994 | 0.5350 | **0.4925** | 0.6053 | 0.5611 | 0.5996 | 0.6106 | 0.5340 | 1.3416 | 0.8780 |
| | 4 | 1.2057 | 1.0555 | 1.4831 | 0.9392 | 0.7866 | 1.0583 | 0.9896 | 1.1446 | 0.9707 | **0.5736** | 7.5001 | 0.9715 |
| | 1 | 0.2976 | 0.7793 | 0.0929 | 0.2972 | 0.2094 | 0.7783 | 0.7025 | 0.0930 | 0.0934 | 0.4457 | **0.0922** | 0.8821 |
| $N = 10$ | 2 | 0.5805 | 0.7579 | 0.4140 | 0.4998 | 0.5962 | 0.6545 | 0.7219 | **0.3941** | 0.7482 | 0.6788 | 6.5729 | 0.9874 |
| $T = 40$ | 3 | 0.7797 | 0.7712 | 0.8573 | 0.6243 | **0.5566** | 0.6346 | 0.5574 | 0.6790 | 0.6739 | 0.6022 | 4.0247 | 0.9802 |
| | 4 | 1.3906 | 1.6819 | 1.9814 | 1.1018 | 0.9189 | 2.6735 | 2.6751 | 1.4965 | 1.2799 | **0.5927** | 23.3102 | 0.9962 |

Table 2: MSE comparison: Large coefficient difference (uncertain clustering, $R^2 = 0.9$)

| | DGP | $\text{MPA}_{uc}$ | $\text{MPA}_{mc}$ | $\text{Equal}_{uc}$ | $\text{Equal}_{mc}$ | $\text{SBIC}_{uc}$ | $\text{SBIC}_{mc}$ | Mixed | Pool | SHK |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 0.3726 | 0.3084 | 0.6519 | 0.6006 | 0.1923 | 0.1761 | 0.6404 | **0.1595** | 0.7580 |
| $N = 5$ | 2 | 0.7216 | 0.7492 | **0.5867** | 0.6355 | 0.9086 | 0.9800 | 0.7510 | 1.2556 | 0.9137 |
| $T = 15$ | 3 | 0.5467 | 0.5774 | **0.5416** | 0.5847 | 0.6168 | 0.6771 | 0.7095 | 0.7578 | 0.8739 |
| | 4 | 1.1393 | 1.0255 | 3.9513 | 3.9617 | 1.3188 | 1.1571 | **0.7592** | 24.1246 | 0.9944 |
| | 1 | 0.3721 | 0.3116 | 0.6703 | 0.6205 | 0.1898 | 0.1886 | 0.6523 | **0.1841** | 0.9157 |
| $N = 5$ | 2 | 0.7992 | 0.8547 | **0.7450** | 0.9070 | 1.0476 | 1.3168 | 0.8225 | 3.5609 | 0.9868 |
| $T = 40$ | 3 | 0.6822 | 0.6840 | **0.6388** | 0.6963 | 0.7597 | 1.0072 | 0.7934 | 2.0261 | 0.9770 |
| | 4 | 1.0407 | 1.0077 | 11.5694 | 11.6314 | 1.0468 | 1.0233 | **0.7662** | 73.6364 | 0.9993 |
| | 1 | 0.3039 | 0.2122 | 0.7634 | 0.6786 | 0.0816 | 0.0810 | 0.4381 | **0.0782** | 0.6670 |
| $N = 10$ | 2 | **0.5164** | 0.5638 | 0.5866 | 0.6324 | 0.6512 | 0.7338 | 0.5215 | 0.9322 | 0.8486 |
| $T = 15$ | 3 | **0.4336** | 0.4560 | 0.5973 | 0.6325 | 0.4775 | 0.5014 | 0.4899 | 0.5324 | 0.7942 |
| | 4 | 0.9402 | 0.7914 | 1.0498 | 0.9832 | 1.1480 | 0.9718 | **0.5779** | 7.4181 | 0.9710 |
| | 1 | 0.2991 | 0.2118 | 0.7783 | 0.7007 | 0.0939 | 0.0942 | 0.4466 | **0.0930** | 0.8829 |
| $N = 10$ | 2 | **0.5565** | 0.6743 | 0.6174 | 0.6786 | 0.6689 | 1.1366 | 0.6104 | 2.7270 | 0.9729 |
| $T = 40$ | 3 | **0.5560** | 0.5915 | 0.6098 | 0.6220 | 0.6670 | 0.8404 | 0.5828 | 1.4956 | 0.9574 |
| | 4 | 1.1003 | 0.9179 | 2.6708 | 2.6712 | 1.4882 | 1.2750 | **0.5929** | 23.1922 | 0.9962 |

Table 3: MSE comparison: Small coefficient difference (uncertain clustering, $R^2 = 0.9$)

| | DGP | $\text{MPA}_{uc}$ | $\text{MPA}_{mc}$ | $\text{Equal}_{uc}$ | $\text{Equal}_{mc}$ | $\text{SBIC}_{uc}$ | $\text{SBIC}_{mc}$ | Mixed | Pool | SHK |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 0.3681 | 0.3052 | 0.6513 | 0.5980 | 0.1846 | 0.1770 | 0.6384 | **0.1593** | 0.7574 |
| $N = 5$ | 2 | 0.6912 | 0.7857 | **0.6877** | 0.8067 | 0.7012 | 0.9663 | 0.8274 | 3.6256 | 0.9637 |
| $T = 15$ | 3 | 0.7922 | 0.7973 | **0.6942** | 0.7578 | 0.9706 | 1.1488 | 0.8315 | 2.1664 | 0.9412 |
| | 4 | 1.1468 | 1.0287 | 3.9902 | 4.0002 | 1.3306 | 1.1685 | **0.7558** | 24.3258 | 0.9945 |
| | 1 | 0.3692 | 0.3099 | 0.6711 | 0.6204 | 0.1911 | 0.1898 | 0.6473 | **0.1868** | 0.9146 |
| $N = 5$ | 2 | 0.5922 | 0.7362 | 0.9293 | 1.3787 | **0.5089** | 0.6813 | 0.8706 | 10.7730 | 0.9953 |
| $T = 40$ | 3 | 0.9450 | **0.9081** | 1.0272 | 1.2215 | 1.1521 | 1.3136 | 0.9140 | 6.3827 | 0.9918 |
| | 4 | 1.0438 | 1.0092 | 11.6052 | 11.6682 | 1.0479 | 1.0214 | **0.7687** | 73.8860 | 0.9994 |
| | 1 | 0.3031 | 0.2109 | 0.7633 | 0.6803 | 0.0836 | 0.0812 | 0.4391 | **0.0787** | 0.6671 |
| $N = 10$ | 2 | 0.4882 | 0.5775 | 0.6331 | 0.6067 | **0.4201** | 0.7600 | 0.6158 | 2.8414 | 0.9306 |
| $T = 15$ | 3 | 0.7245 | 0.6474 | 0.7150 | 0.6288 | 0.8991 | 0.8680 | **0.6083** | 1.9015 | 0.9068 |
| | 4 | 0.9441 | 0.7922 | 1.0695 | 1.0037 | 1.1626 | 0.9773 | **0.5793** | 7.5645 | 0.9715 |
| | 1 | 0.2958 | 0.2109 | 0.7778 | 0.7004 | 0.0916 | 0.0919 | 0.4463 | **0.0912** | 0.8827 |
| $N = 10$ | 2 | 0.4137 | 0.4566 | 0.6664 | 0.6744 | **0.2229** | 0.3470 | 0.7043 | 8.6009 | 0.9902 |
| $T = 40$ | 3 | 0.8816 | 0.8271 | 0.8063 | 0.7789 | 1.1940 | 1.3438 | **0.6996** | 5.8831 | 0.9858 |
| | 4 | 1.0985 | 0.9162 | 2.6774 | 2.6806 | 1.4806 | 1.2792 | **0.5942** | 23.2004 | 0.9963 |

Table 4: More noise in the model ($N = 10$, $T = 40$)

| $R^2$ | $q$ | DGP | MPA$_{uc}$ | MPA$_{mc}$ | Equal$_{uc}$ | Equal$_{mc}$ | SBIC$_{uc}$ | SBIC$_{mc}$ | Mixed | FGLS$_{\bar{\beta}}$ | Pool |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 0.2965 | 0.2066 | 0.7788 | 0.7014 | 0.0932 | 0.0943 | 0.4472 | 0.0959 | **0.0924** |
| | $q_B$ | 2 | 0.5423 | 0.5335 | 0.6440 | 0.6266 | 0.7033 | 0.7354 | **0.5275** | 0.8271 | 0.8218 |
| | | 3 | 0.4478 | **0.4295** | 0.6267 | 0.6213 | 0.5040 | 0.5051 | 0.4898 | 0.5324 | 0.5298 |
| | | 4 | 0.7735 | 0.6847 | 0.6461 | **0.5633** | 1.1466 | 1.1756 | 0.5689 | 2.6111 | 2.6260 |
| | | 1 | 0.2968 | 0.2068 | 0.7780 | 0.6995 | 0.0916 | 0.0920 | 0.4457 | 0.0949 | **0.0913** |
| 0.75 | $q_L$ | 2 | 0.4308 | 0.4270 | 0.5989 | 0.6445 | 0.3986 | 0.3915 | 0.4461 | **0.3777** | 0.3843 |
| | | 3 | 0.3323 | 0.3307 | 0.5947 | 0.6548 | 0.2563 | 0.2541 | 0.4381 | **0.2406** | 0.2492 |
| | | 4 | 0.7752 | 0.6809 | 0.6450 | 0.5612 | 1.1720 | 1.1869 | **0.5609** | 2.6549 | 2.6723 |
| | | 1 | 0.2957 | 0.2068 | 0.7776 | 0.7017 | 0.0928 | 0.0931 | 0.4466 | 0.0955 | **0.0922** |
| | $q_S$ | 2 | **0.5407** | 0.5525 | 0.6597 | 0.6196 | 0.6707 | 0.8177 | 0.5577 | 1.0392 | 1.0350 |
| | | 3 | 0.5672 | **0.5128** | 0.7070 | 0.6434 | 0.6945 | 0.6823 | 0.5430 | 0.7296 | 0.7289 |
| | | 4 | 0.7779 | 0.6908 | 0.6511 | 0.5709 | 1.1621 | 1.1803 | **0.5705** | 2.6446 | 2.6582 |
| | | 1 | 0.2975 | 0.1970 | 0.7772 | 0.6955 | 0.0925 | 0.0928 | 0.4460 | 0.0960 | **0.0922** |
| | $q_B$ | 2 | 0.3462 | 0.3240 | 0.6223 | 0.6564 | 0.1924 | 0.1854 | 0.3793 | **0.1457** | 0.1743 |
| | | 3 | 0.2942 | 0.2612 | 0.6194 | 0.6523 | 0.1491 | 0.1443 | 0.3909 | **0.1197** | 0.1427 |
| | | 4 | 0.4599 | 0.4578 | 0.5448 | 0.5319 | 0.4307 | 0.4217 | 0.4294 | **0.3615** | 0.3743 |
| | | 1 | 0.2975 | 0.1970 | 0.7772 | 0.6955 | 0.0925 | 0.0928 | 0.4460 | 0.0960 | **0.0922** |
| 0.5 | $q_L$ | 2 | 0.2975 | 0.3001 | 0.5915 | 0.6628 | 0.1377 | 0.1377 | 0.3454 | **0.0853** | 0.1256 |
| | | 3 | 0.2433 | 0.2351 | 0.5898 | 0.6649 | 0.1155 | 0.1130 | 0.3800 | **0.0832** | 0.1107 |
| | | 4 | 0.4599 | 0.4578 | 0.5448 | 0.5319 | 0.4307 | 0.4217 | 0.4294 | **0.3615** | 0.3743 |
| | | 1 | 0.2975 | 0.1970 | 0.7772 | 0.6955 | 0.0925 | 0.0928 | 0.4460 | 0.0960 | **0.0922** |
| | $q_S$ | 2 | 0.3710 | 0.3309 | 0.6540 | 0.6540 | 0.2190 | 0.2083 | 0.4059 | **0.1783** | 0.1996 |
| | | 3 | 0.3456 | 0.2833 | 0.6849 | 0.6605 | 0.1759 | 0.1694 | 0.4087 | **0.1457** | 0.1647 |
| | | 4 | 0.4599 | 0.4578 | 0.5448 | 0.5319 | 0.4307 | 0.4217 | 0.4294 | **0.3615** | 0.3743 |

Table 5: Effects of CDS spreads determinants: Mallows pooling averaging estimates

| | Brazil | Bulgaria | Chile | China | Columbia | Croatia | Hungary | Japan | Korea |
|---|---|---|---|---|---|---|---|---|---|
| *lstock* | −0.2283 | −0.1209 | −0.1209 | −0.1667 | −0.1667 | −0.1667 | −0.2159 | −0.2159 | −0.1366 |
| *fxrates* | 0.1312 | 0.0640 | 0.0640 | −0.0043 | 0.0509 | 0.0509 | 0.0509 | 0.0509 | 0.0982 |
| *fxres* | 0.0426 | 0.0426 | 0.0296 | −0.0212 | −0.0074 | −0.0074 | 0.0760 | 0.0029 | −0.0475 |
| *gstock* | −0.3862 | −0.4943 | −0.4050 | −0.4050 | −0.4591 | −0.4591 | −0.3733 | −0.3733 | −0.4670 |
| *trsy* | 0.1018 | 0.0233 | 0.0233 | 0.0233 | 0.0084 | 0.0828 | 0.0115 | −0.0804 | 0.0282 |
| *ig* | 0.0868 | 0.0232 | −0.0538 | 0.0561 | −0.0172 | 0.0143 | −0.0245 | −0.1313 | −0.0749 |
| *hy* | 0.2165 | 0.3344 | 0.2096 | 0.3805 | 0.2006 | 0.4026 | 0.2908 | 0.2179 | 0.1798 |
| *eqp* | 0.0236 | 0.0168 | 0.1271 | 0.0951 | 0.0331 | −0.0369 | 0.0313 | 0.0129 | 0.0099 |
| *volp* | −0.0888 | −0.2177 | −0.2177 | −0.2177 | −0.0925 | −0.1940 | −0.1940 | −0.2108 | −0.1959 |
| *termp* | 0.0205 | −0.0052 | −0.0159 | 0.0053 | 0.0496 | −0.0442 | −0.0257 | 0.0193 | −0.0204 |
| *ef* | 0.0930 | 0.0197 | −0.0565 | 0.0523 | 0.0329 | −0.0026 | −0.0026 | 0.0085 | 0.0198 |
| *bf* | 0.0847 | −0.0157 | −0.0499 | 0.0115 | 0.0036 | −0.0235 | −0.0559 | 0.0488 | −0.0086 |
| $R^2$ | 0.45 | 0.71 | 0.78 | 0.70 | 0.51 | 0.77 | 0.73 | 0.54 | 0.75 |

*Notes:* $R^2$ is obtained from individual estimation of each country.

Table 5: Effects of CDS spreads determinants: Mallows pooling averaging estimates (Continued)

| | Malaysia | Mexico | Philippines | Poland | Romania | Russia | Slovak | Thailand | Turkey |
|---|---|---|---|---|---|---|---|---|---|
| *lstock* | −0.1468 | −0.2194 | −0.2194 | −0.1610 | −0.1151 | −0.1998 | −0.1350 | −0.1556 | −0.4115 |
| *fxrates* | 0.0982 | 0.0982 | 0.0402 | 0.0402 | 0.0402 | 0.0402 | 0.0402 | 0.0402 | 0.0931 |
| *fxres* | 0.0256 | −0.0360 | −0.0446 | −0.0412 | −0.0246 | −0.0246 | −0.0246 | −0.0246 | 0.0088 |
| *gstock* | −0.4670 | −0.3808 | −0.3808 | −0.3177 | −0.4329 | −0.4284 | −0.4284 | −0.4306 | −0.1854 |
| *trsy* | 0.0282 | −0.0065 | 0.0341 | 0.0372 | 0.0450 | 0.1298 | 0.0922 | 0.0145 | 0.0876 |
| *ig* | −0.0749 | −0.0783 | 0.0228 | 0.0115 | 0.0298 | 0.0943 | −0.0196 | −0.0196 | 0.0796 |
| *hy* | 0.2656 | 0.1621 | 0.1963 | 0.4175 | 0.4175 | 0.4175 | 0.3328 | 0.3328 | 0.2665 |
| *eqp* | 0.0099 | 0.0870 | 0.0972 | 0.1055 | 0.0241 | 0.0241 | 0.0241 | 0.0241 | 0.0404 |
| *volp* | −0.1959 | −0.1839 | −0.2162 | −0.2493 | −0.2261 | −0.2261 | −0.2261 | −0.0922 | −0.0922 |
| *termp* | −0.0204 | 0.0244 | −0.0188 | −0.0233 | 0.0088 | −0.0800 | −0.0387 | −0.0387 | −0.0676 |
| *ef* | 0.0074 | 0.0177 | 0.0251 | −0.0056 | 0.0207 | 0.0534 | 0.0107 | 0.0107 | 0.0404 |
| *bf* | −0.0086 | −0.0095 | 0.0377 | −0.0175 | 0.0026 | −0.0294 | −0.0151 | −0.0151 | 0.0785 |
| $R^2$ | 0.74 | 0.82 | 0.52 | 0.82 | 0.73 | 0.79 | 0.69 | 0.66 | 0.50 |

*Notes:* $R^2$ is obtained from individual estimation of each country.

Table 6: Out-of-sample prediction comparison

| | *Full sample* | | *Latin America* | | *Asia* | | *Large* $R^2$ | |
|---|---|---|---|---|---|---|---|---|
| | AAD | RMSPE | AAD | RMSPE | AAD | RMSPE | AAD | RMSPE |
| $\text{MPA}_{det}$ | 0.8938 | 0.9023 | 0.8916 | 0.8788 | 0.9372 | 0.9434 | 0.9791 | 0.9820 |
| $\text{MPA}_{unc}$ | 0.8498 | 0.8805 | **0.8451** | **0.8320** | 0.9321 | 0.9257 | 0.9611 | **0.9638** |
| $\text{Equal}_{det}$ | 0.9702 | 0.9702 | 0.9270 | 0.9091 | 0.9483 | 0.9542 | 0.9743 | 0.9755 |
| $\text{Equal}_{unc}$ | 0.9466 | 0.9539 | 0.9340 | 0.9193 | 0.9629 | 0.9617 | 0.9705 | 0.9702 |
| $\text{SAIC}_{det}$ | 0.9572 | 0.9505 | 0.9625 | 0.9484 | 0.9386 | 0.9416 | 0.9982 | 1.0019 |
| $\text{SAIC}_{unc}$ | 0.9063 | 0.9220 | 0.8476 | 0.8321 | 0.9096 | 0.9031 | 0.9869 | 0.9859 |
| $\text{AIC}_{det}$ | 0.9572 | 0.9505 | 0.9625 | 0.9484 | 0.9419 | 0.9420 | 0.9982 | 1.0019 |
| $\text{AIC}_{unc}$ | 0.9063 | 0.9220 | 0.8541 | 0.8381 | **0.8948** | **0.8891** | 0.9866 | 0.9858 |
| SBIC* | **0.8368** | **0.8579** | 0.8541 | 0.8381 | **0.8948** | **0.8891** | **0.9607** | 0.9653 |
| $\text{FGLS}_{\beta}$ | 0.9948 | 0.9830 | 0.9882 | 0.9390 | 0.9795 | 0.9795 | 0.9761 | 0.9896 |
| $\text{FGLS}_{\bar{\beta}}$ | 0.8392 | 0.8675 | 1.7005 | 1.6284 | 1.0562 | 1.0095 | 0.9848 | 0.9907 |
| Mixed | 0.9700 | 0.9689 | 2.7459 | 2.9217 | 1.4370 | 1.5806 | 1.0754 | 1.0816 |
| Pool | **0.8368** | **0.8579** | 0.8541 | 0.8381 | **0.8948** | **0.8891** | **0.9607** | 0.9653 |
| SHK | 0.9693 | 0.9732 | 0.9774 | 0.9699 | 0.9777 | 0.9771 | 0.9954 | 0.9946 |

*Notes:* * The results of BIC are generally the same as SBIC, and therefore omitted. Both results are unaffected by the screening methods.