To pool or not to pool:
What is a good strategy for parameter estimation and forecasting in panel regressions?*

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

Xinyu Zhang  
*Academy of Mathematics and Systems Science, Chinese Academy of Sciences*

Richard Paap  
*Econometric Institute, Erasmus University Rotterdam*

**Abstract:** This paper considers estimating the slope parameters in potentially heterogeneous panel data regressions for explaining and forecasting cross-country sovereign credit risk. We propose a novel optimal pooling averaging estimator that makes an explicit trade-off between efficiency gains from pooling and bias due to heterogeneity. By theoretically and numerically comparing various estimators, we find that a uniformly best estimator does not exist and that our new estimator is superior in non-extreme cases. The results provide practical guidance for the best estimator depending on features of data and models.

**Keywords:** Credit default swap spreads; Heterogeneous panel; Mean squared error; Model screening; Pooling averaging;  

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

---

*Address for correspondence: R. Paap, Econometric Institute (H11-17), Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands, e-mail: paap@ese.eur.nl. We thank participants at seminars at Chinese Academy of Sciences, Renmin University, Xiamen University, the 10th NESG at Maastricht, and the 21st international panel data conference at Budapest; Econometric Society 2016 China Meeting at Chengdu and Australasia Meeting at Sydney, to Pavel Ciezek, Herman van Dijk, Jan Magnus, and Guohua Zou for useful discussions and constructive comments. All errors are ours.*
1 Introduction

Since the breakout of US and European debt crisis, most countries have experienced a large and rapid increase in their sovereign government debt. Considerable attention has thus been paid to the sovereigns’ credit risk, especially those under great economic pressure. To fully recover from the crisis and hopefully avoid future debt crises, a crucial task is to understand the nature of sovereign credit risk. Is the sovereign credit risk only driven by local economic performance or by forces from the global market? Besides, it is of equally significance and interest to forecast the future sovereign credit risk. A good forecast not only serves as a crucial fundamental of an effective policy but also directly influences the ability of financial market participants to diversify the risk. As a popular indicator of sovereign credit risk, the determinants of sovereign credit default swap (CDS) spreads have been widely studied for both developed and emerging countries, see, among others, Longstaff et al. (2011); Dieckmann and Plank (2012). The literature suggests that the effect of potential determinants exhibits significant degree of heterogeneity across countries but also possesses some common pattern within certain sub-samples. Hence, this raises an important issue how to model heterogeneity the best way for examining the determinants of sovereign CDS spreads and for forecasting the future spreads.

Existing studies on sovereign CDS spreads are either based on individual time series regression or on a pooled regression model. The former considers separate models for each country as in Longstaff et al. (2011) and Dieckmann and Plank (2012) and estimates the parameters of the individual model independently. The pooled model ignores the heterogeneity and assumes homogeneous coefficients for all countries. A natural question is now which assumption leads to more reliable results, and whether there exists any alternative and better way to handle the heterogeneity.

The issue on how to model the potentially heterogeneous parameters across individual units is sometimes poetically referred by econometricians to as “to pool or not to pool”. This is a long-existing issue in the panel data analysis, but there is still no consensus on it. On the one hand, an increasing number of studies have noted that the homogeneity assumption of parameters is vulnerable in practice, and that the violation of this assumption can lead to misleading estimates. For example, Su and Chen (2013) and Durlauf et al. (2001) 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 empir-
ical studies find that the pooled estimator and forecast often outperform those obtained from individual time series analysis in terms of mean square (forecasting) error (MSE or MSFE), see, for example, Baltagi and Griffin (1997), Baltagi et al. (2000), and Hoogstrate et al. (2000).

The above mentioned empirical results suggest that the pooling decision involves the typical bias-variance trade-off, and that the amount of pooling should depend on the situation at hand. 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 trade-off between the efficiency and bias when estimating or forecasting in a heterogeneous panel data model? Second, is there a fit-for-all estimator that performs well in all situations, 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 trade-off. Furthermore, we provide practical guidance on how best to handle parameter heterogeneity in empirical research.

The first contribution of this paper concerns a pooling averaging method that makes an optimal bias-variance trade-off in estimating and forecasting in heterogeneous panel data models. The optimal trade-off is achieved by combining the estimators or forecasts from different pooling specifications with appropriate weights. There are of course several ways to address heterogeneity in the literature, for example, the random coefficient model (Swamy, 1970), the pooled mean group estimator (Pesaran et al., 1999), various group estimators (see, e.g., Lin and Ng, 2012; Bonhomme and Manresa, 2015; Ando and Bai, 2016; Su et al., 2016), see Section 2 for a more thorough review of literature. These estimation strategies are useful, but require correct specification of the heterogeneity structure or the number of groups. Another stream of literature focuses on testing for the homogeneity assumption. The estimator obtained after a preliminary testing step is called the pretest estimator.

Our proposed pooling averaging method includes several popular approaches as special cases, such as shrinkage estimation and pretesting estimation. It has three main advantages compared to the existing approaches. (i) Estimators from different pooling specifications have different degrees of bias and variance. Our method makes an explicit bias-variance trade-off by appropriately combining these estimators/forecasts. Hence, estimation and forecasting is directly based on the MS(F)E. Note that existing methods based on other
criteria than MS(F)E cannot guarantee that the resulting estimators/forecasts achieve the minimum MS(F)E. (ii) Our method does not require specifying the heterogeneity structure or the number of groups, and parameters can be heterogeneous in any pattern. (iii) Our approach avoids the problems caused by pretesting since it is continuous, unconditional (model uncertainty is already taken into account), and has a bounded risk (see, e.g., Danilov and Magnus, 2004). We theoretically examine both the finite sample and asymptotic properties of the pooling averaging estimator based on the often used Mallows criterion.

A practical issue for pooling averaging is that even for a moderate number of individuals and regressors, the number of pooling specifications can be large. Averaging over the entire model space is computationally intensive and inefficient. We introduce a model screening procedure to address this issue. Instead of estimating and averaging over all candidate models, we propose to average over group estimators obtained from different numbers of groups. This results in a much smaller post-screened model space, where averaging is computationally more efficient and accurate. As a by-product, we show that averaging also offers an alternative way to address the difficulty of specifying the number of groups for the group estimators, since the same trade-off between efficiency and consistency applies to choice of the number of groups here. Our simulation and application both show that averaging over different numbers of groups generally leads to better forecasts than selecting a specific number of groups.

The second contribution of this paper is that we analytically compare the finite sample MS(F)E of different pooling estimators/forecasts, and analyse how the performance of these estimators/forecasts varies over situations. Based on this analysis, we provide empirical researchers with guidelines on how to handle parameter heterogeneity in panel data models. Given the fact that the performance of panel data methods is sensitive to data properties, it is important to understand which and how data properties matter in practice. We explicitly show that there does not exist a uniformly best method, and the performance depends on the features of data and models, that is, the degree of coefficient heterogeneity, signal-to-noise ratio, time series dimension, cross-sectional dimension, number of regressors, and the choice of weights. This theoretical finding is supported by an extensive simulation study. An important conclusion from this study is that the pooling averaging estimator/forecast, especially Mallows pooling averaging, is to be preferred when the panel is heterogeneous and the signal-to-noise ratio is moderate or large, while the pooled estimator/forecast is recommended when the signal-to-noise ratio is small.
After we have shown how to deal with slope heterogeneity in an optimal way, we turn back to our empirical question to examine and forecast sovereign CDS spreads for a panel of countries. We extend the time dimension of Longstaff et al.’s (2011) data to 2016. Given the presence of financial crises in our updated sample, we consider possible structural breaks in the slope coefficients. We employ recent developments of structural break detection in heterogeneous panels to identify the change points, and investigate the effect of determinants and the forecasting performance of competing methods with and without structural breaks. In general we find that the pooling averaging provides intuitive estimates. By exploring cross-section variation in an optimal manner, pooling averaging also produces more accurate forecasts than alternative methods including separate country-specific time series forecasts.

The remainder of this paper is organized as follow. Section 2 briefly reviews parameter estimation and testing strategies for heterogeneous panel data models. In Section 3 we discuss the model setup used in this paper and introduce the pooling averaging estimator in its general form. Section 4 derives the MSFE of our pooling averaging forecast and compares it with pooled and individual forecasts. Section 5 discusses the choice of weights in our pooling averaging procedure and its theoretical properties. A model screening procedure is introduced in Section 6. A simulation study is provided in Section 7 to support our theory. Section 8 deals with the empirical study of sovereign CDS spreads. Finally, in Section 9, we offer some practical suggestions on how to handle slope heterogeneity based on the theoretical, simulation and empirical results.

2 Brief Literature Review

The literature on heterogeneous panel data models mainly focuses on how to estimate a (possibly) heterogeneous parameter and how to test the homogeneity assumption. Allowing parameters to vary across individuals can be dated back to Swamy (1970), which proposed to estimate the mean of the heterogeneous coefficient (average effect) using a generalized least squares (GLS) type estimator. 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).

If the average effect is of interest, it can be shown that the FGLS estimator can be written as a weighted average of estimators obtained from individual time series regressions (individual estimators), and that it provides a good trade-off between bias and efficiency (Swamy, 1970). However, researchers are sometimes more interested in the individual parameters (individual-specific effect), especially when heterogeneous policy implications, decisions or forecasts need to be made. Estimating the individual-specific effect and further making individual forecast are the focus of this paper. A popular individual-specific-effect estimator is the shrinkage estimator proposed by Maddala et al. (1997), which is a combination of the pooled and individual regressions. In a dynamic context, Pesaran et al. (1999) distinguished between the long-run and short-run parameters, and only allowed the short-run parameters to be heterogeneous (pooled mean group estimator). The validity of this estimator relies on a careful specification of the long-run and short-run parameters, which is however not required in our set-up. See Baltagi et al. (2008) for an excellent survey.

Recent developments in individual-specific-effect estimation involves a latent class/group specification, such as finite mixture models and various types of grouped estimators. For example, Bonhomme and Manresa (2015) and Lin and Ng (2012) suggested a $k$-means approach, Su et al. (2016) proposed a lasso-type estimator, Wang et al. (2016) extended clustering algorithm in regression via data-driven segmentation (CARDS) to a panel framework, and Su and Wang (2017) classified individual units via a sequential binary segmentation algorithm, all of which can simultaneously determine the unknown group membership and slope coefficients. These group estimators have many attractive features and make good sense, especially when identifying the true grouping is of interest.

Compared to these existing approaches, our objective here is not to consistently estimate the group membership and/or slope coefficients. Instead, we aim at obtaining the most accurate parameter estimator or forecast in terms of MS(F)E. We make a simple, yet largely overlooked point, namely that when the objective is to minimize MS(F)E the true grouping is not necessary optimal. Using controlled incorrect grouping may lead to a substantial gain in efficiency that offsets potential bias. Although MSE or MSFE criteria are usually applied in simulations and applications to evaluate panel estimators, most existing estimators are based on minimizing other criteria. Our pooling averaging method explicitly aims at minimizing the MS(F)E. Furthermore, pooling averaging also offers an
alternative way of dealing with uncertainty in the number of latent groups in classification methods by combining estimates obtained from different numbers of groups. Given the trade-off between consistency and efficiency for different choices in the number of groups, one can make an optimal trade-off by averaging and appropriately choosing weights.

Alternatively, the literature has provided various tests for parameter homogeneity under different model specifications. To give a partial list of such tests, Pesaran and Yamagata (2008) proposed dispersion type tests for large panels with large cross section and time dimension. Su and Chen (2013) proposed a residual-based test applicable in panel models with interactive fixed effects. Jin and Su (2013) considered the case with cross-sectional dependence and provided a non-parametric test. If the ultimate interest lies in estimating the parameters of the models, the pretest estimator is however not completely satisfactory. It is discontinuous and has unbounded risk. These are the general problems of pretesting estimators, see Danilov and Magnus (2004) for a detailed discussion. Besides, in the context of heterogeneous panel, even if the true model is selected, it does not necessarily produce the best estimator in terms of MSE nor the best forecast in terms of MSFE. We analyse the finite sample and asymptotic properties of our pooled average estimator. To our best knowledge, no finite sample properties of existing heterogeneous panel estimators are theoretically studied except for the shrinkage estimator (Maddala et al., 2001). One limitation of our approach is that it only provides point estimates for the coefficients, and thus statistical inference is challenging. We provide a practical way to estimate the variance and confidence interval of the estimator based on bootstrap. The theoretical justification of using bootstrap statistics in the model averaging framework is beyond the scope of this paper and considered to be an interesting topic for future research.

3 Pooling Averaging Estimation

3.1 Model Setup

Consider the linear panel data model with heterogeneous slopes

\[ y_i = X_i \beta_i + u_i \quad i = 1, \ldots, N, \]

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, that is, \( X_{it1} = 1 \) for all \( i, t \), and where the series \( \{y_{it}, X_{it}\} \)
is assumed to be stationary. The coefficient $\beta_i = (\beta_{i1}, \ldots, \beta_{ik})'$ is assumed to be *fixed* but allowed to differ across individuals, that is, some or all of the elements in $\beta_i$ can be different from the elements in $\beta_j$ for $i \neq j$. If pooling of the individual-specific intercept is not desired, one can eliminate the fixed effect using a within transformation. For notation convenience, we first assume that the error term of each individual $u_i$ is independently and identically distributed (IID) across time, but different individuals can have heteroskedastic errors with mean zero and variance $\sigma^2_i I_T$ (between-individual heteroskedasticity). Later we shall relax this assumption by allowing conditional heteroskedastic errors both between and within individuals (completely heteroskedasticity). The $u_1, \ldots, u_N$ terms are assumed to be uncorrelated conditional on $X_i$ for all $i$. In some cases we use the matrix form of (1) which is given by

$$y = X\beta + u,$$

where $y = (y'_1, \ldots, y'_N)'$, $X = \text{diag}(X_1, \ldots, X_N)$, $\beta = (\beta'_1, \ldots, \beta'_N)'$, and $u = (u'_1, \ldots, u'_N)'$.

To demonstrate our new method we focus on the case of strictly exogenous explanatory variables and assume that model (1) is correctly specified in regressors. This assumption ensures the unbiasedness of the individual estimator, but it rules out the dynamic model where the lagged dependent variable is included as explanatory variables. In the dynamic panel and the presence of omitted variables, least square estimation of individual time series is biased, and thus theorems derived later in Section 5.1 do not hold. Nevertheless, the bias-variance trade-off remains relevant. In these two cases, the bias of pooling estimators is caused by heterogeneity and endogeneity. Pooling may increase or decrease the endogeneity bias depending on the bias of individual estimators, and it is not clear how the overall bias will be changed by pooling.

### 3.2 Average Pooling Strategies

Our goal is to estimate each individual’s coefficient $\beta_i$ or forecast individual’s outcome variable. The value of individual coefficients is of particular interest when individualized policies or decisions need to be made. This goal is different from the random coefficient model where one wants to estimate a common average effect, say $E(\beta_i)$.

To estimate the $\beta_i$ parameters in (1), one can consider separate least-square (LS) estimators for each time series as long as $T > k$, $\hat{\beta}_i = (X'_iX_i)^{-1}X'_iy_i$, called the individual estimator. This individual estimator $\hat{\beta}_i$ is unbiased given that individual $i$’s regression is
correctly specified, but it is not efficient as it is a limited information estimator, not making use of any cross-section variation at all. In the other extreme, one can ignore the slope heterogeneity and estimate the pooled model, obtaining a common estimator for all individuals, that is, 
\[ b = (\sum_{i=1}^{N} X_i'X_i)^{-1} \sum_{i=1}^{N} X_i'X_i \hat{\beta}_i. \]
The pooled estimator 
\[ \hat{\beta}_{\text{pool}} = (b', \ldots, b')' \]
is more efficient than the individual estimator, but can be severely biased due to incorrect pooling of heterogeneous coefficients. The comparison between these two estimators suggests a typical bias-variance trade-off in choosing which estimator to use. The forecast of individual’s outcome variable that associates with \( \hat{\beta}_i \) and \( \hat{\beta}_{\text{pool}} \) can be obtained by 
\[ \hat{y}_i = X \hat{\beta}_i \]
and 
\[ \hat{y}_{\text{pool}} = X \hat{\beta}_{\text{pool}}, \]
respectively, and they face precisely the same bias-variance trade-off.

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 (2), that is,
\[ R_m \beta = 0, \]
where \( R_m \) is the restriction matrix under the \( m \)-th pooling strategy. For instance, if the restriction is \( \beta_i = \beta_j \) for \( j > i \), then 
\[ R_m = (0_{k \times (i-1)k}, I_k, 0_{k \times (j-i-1)k}, -I_k, 0_{k \times (N-j)k}). \]
For each \( R_m \), we can construct the projection matrix \( P_m \)
\[ P_m = I_N - (X'X)^{-1} R_m' (R_m(X'X)^{-1} R_m')^{-1} R_m, \]
such that the OLS estimator under the \( m \)-th pooling strategy is
\[ \hat{\beta}_{(m)} = P_m \hat{\beta}, \]
where \( \hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_N)' \) is the vector of individual OLS estimators. The estimator \( \hat{\beta}_{(m)} \) 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 to identical coefficients over individuals. The question is now 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. However, in practice, it is difficult to determine the true model because it is hard to distinguish efficiency loss resulting from inefficient pooling or resulting form estimation noise. Even if one can select the correct parameter restrictions, the true restriction specification does not always produce
the best estimator or forecast in terms of MSE and MSFE. This happens, for example, when the heterogeneity in coefficients is small while the signal-to-noise ratio is small. In this case incorrectly pooling heterogeneous individuals may lead to lower MS(F)E, because the efficiency gains from pooling dominate the heterogeneity bias. Therefore, if the MSE of the coefficient estimates or the forecasts of the \( y_i \) variables are of central interest, it seems less plausible to focus on testing or selecting the right pooling pattern.

To make an optimal trade-off between bias and efficiency, we propose to average estimators or forecasts from different pooling strategies and appropriately choose the weights. Our pooling averaging estimator is given by

\[
\hat{\beta}(w) = \sum_{m=1}^{M} w_m \hat{\beta}_{(m)} = \sum_{m=1}^{M} w_m P_m \hat{\beta} = P(w)\hat{\beta},
\]

where \( M \) is the number of candidate pooling strategies, \( P(w) = \sum_{m=1}^{M} w_m P_m \) is an \( Nk \times Nk \) matrix, and \( w = (w_1, \ldots, w_M)' \) belongs to the set \( \mathcal{W} = \{w \in [0,1]^M : \sum_{m=1}^{M} w_m = 1\} \). Its associated combined forecast is \( \hat{y}(w) = X\hat{\beta}(w) \). In practice, the number of pooling strategies can be substantial, much larger than \( N \), due to the lack of natural ordering over the individual units. In the case of large \( M \), we propose to “screen out” poor pooling strategies as a preliminary step based on efficient clustering. We discuss this approach in more detail in Section 6.

Our pooling averaging estimator (6) clearly includes the pretesting estimator as a special case that assigns all weights to a single candidate estimator. It also includes the popular shrinkage estimator of Maddala et al. (1997) as a special case. The shrinkage estimator is defined as

\[
\hat{\beta}_{\text{shrinkage}} = \left(1 - \frac{\nu}{F}\right)\hat{\beta} + \frac{\nu}{F}\hat{\beta}_{\text{pool}},
\]

where \( \nu = \left[(N - 1)k - 2\right]/\left[NT - Nk + 2\right] \) and \( F \) is the test statistic for null hypothesis \( H_0 : \beta_1 = \ldots = \beta_N \). This estimator can be regarded as the pooling average of only the pooled and individual estimators.

4 Theoretical MSFE 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. To save space, the discussion focuses on forecasting and we theoretically compare
the mean square forecast error of pooling averaging with the pooled and individual time series models. The comparison of slope coefficient estimators can be done in a similar way.

The purpose of this analysis is twofold. Although there have been many empirical studies showing that the performance of forecasts/estimators differs significantly in applications, there is lack of theoretical explanation, and no consensus is reached on which method to use in different practical situations. Hence, the first purpose is to provide theoretical explanations for the diverging performance of forecasts/estimators. Second, the theoretical comparison also sheds some light on how data and model features, e.g. the degree of coefficient heterogeneity and level of noise, affect the performance of alternative forecasts. This further provides guidance on which method to choose in practice. To sharpen the focus and highlight the role of different quantities on the forecasts, we assume that the weights are non-random.

For notation simplicity, we denote $Q_i = X_i'X_i/T$, $Q = \sum_{i=1}^{N} Q_i$, and $\|\theta\|_A^2 = \theta' A \theta$ for any vector $\theta$, where $A = \text{diag}(A_1, \ldots, A_N) = X'X$ and $A_i = X_i'X_i$ for $i = 1, \ldots, N$.\(^1\) We perform the MSFE comparison under between-individual heteroskedastic errors. We denote the variance of the individual coefficient estimator as $V_i = \sigma_i^2 Q_i^{-1}/T$ and let $V = \text{diag}(V_1, \ldots, V_N)$. The analysis can easily be extended to completely (conditional) heteroskedastic errors but with more notational complexity.

The pooled forecast can be obtained by $\hat{y}_{\text{pool}} = X \hat{\beta}_{\text{pool}}$, where $\hat{\beta}_{\text{pool}} = (b', \ldots, b')'$ and $b = Q^{-1} \sum_{i=1}^{N} Q_i \hat{\beta}_i$. The individual forecast is based on individual estimators, that is, $\hat{y}_{\text{ind}} = (\hat{y}_1', \ldots, \hat{y}_N')$ with $\hat{y}_i = X \hat{\beta}_i$, and these individual estimators $\hat{\beta}_i$’s are uncorrelated with $\hat{\beta}_i \sim (\beta_i, \sigma_i^2 Q_i^{-1}/T)$.\(^2\) Hence, the MSFEs of the individual and pooled forecasts can be obtained by

$$\text{MSFE}_{\text{ind}} \equiv \text{MSFE}(\hat{y}_{\text{ind}}) = \sum_{i=1}^{N} E \|\hat{\beta}_i - \beta_i\|_A^2 = \frac{1}{T} \sum_{i=1}^{N} \sigma_i^2 \text{tr}(Q_i^{-1} A_i)$$  \hspace{1cm} (8)

and

$$\text{MSFE}_{\text{pool}} \equiv \text{MSFE}(\hat{y}_{\text{pool}}) = \sum_{i=1}^{N} E \|b - \beta_i\|_A^2 = \sum_{i=1}^{N} Q_i \beta_i - \beta_i\|^2_{A_i} + \frac{N}{T} \sum_{i=1}^{N} \text{tr}(\sigma_i^2 Q_i Q_i^{-1} A_i).$$  \hspace{1cm} (9)

\(^1\)The comparison of slope coefficient estimates can be made by setting $A = I_{Nk}$.

\(^2\)In the dynamic panel the OLS estimator is biased. Comparing the MSFEs of biased estimators is still possible, but it complicates the analysis since the degree of bias differs across model specifications.
The first term in (9) captures the bias caused by pooling heterogeneous coefficients, and the second term measures the variance. Note that for fixed $N$, as $T$ goes to infinity, MSFE$_{ind}$ is generally lower order than the first term in (9), suggesting that individual forecast is always better than the pooled forecast under fixed $N$ and large $T$ asymptotics. However, if both $N$ and $T$ go to infinity, there can exist trade-off if (8) and the first term of (9) are of comparable scale. This happens, Furthermore, there is no guarantee that MSFE$_{ind}$ is less than MSFE$_{pool}$ in finite samples. The relation between the finite sample MSFE$_{ind}$ and MSFE$_{pool}$ depends on the magnitude of the bias term $\sum_{i=1}^{N} \|Q^{-1} \sum_{i=1}^{N} Q_i \beta_i - \beta_i \|_A^2$ and the difference between two scaled variance terms $\frac{1}{T} \sum_{i=1}^{N} \sigma_i^2 \text{tr}(Q^{-1_i} A_i) - \frac{N}{T} \sum_{i=1}^{N} \text{tr}(\sigma_i^2 Q^{-1} Q^{-1} A_i)$. In practice, error variances may be quite large in which case the variance term dominates. Hence, this explains, to some extent, why individual time series forecasts are less preferred in most empirical research.

Next, we derive the MSFE of the pooling averaging forecast $\hat{y}(w) = X\hat{\beta}(w)$ assuming weights are given. In this case, we have

$$\text{MSFE}(\hat{y}(w)) = E\|\hat{\beta}(w) - \beta\|_A^2 = \|P(w)\hat{\beta} - \beta\|_A^2$$

$$= \|P(w)\beta - \beta\|_A^2 + \text{tr}[P(w)VP'(w)A].$$

(10)

We see that the comparison between the MSFEs depends on the degree of heterogeneity in the true coefficients $\beta$, the error variances of individual regressions $\sigma_i^2$‘s contained in $V$, $A$, and of course the weight choice. To shed light on this comparison, we consider below several special cases.

First, if the pooling averaging estimator $\hat{\beta}(w)$ only averages over the pooled and individual estimators, namely $\hat{\beta}_i(w) = w_1 b + w_2 \hat{\beta}_i$, we can write MSFE($\hat{y}(w)$) in terms of MSFE$_{pool}$ and MSFE$_{ind}$ as

$$\text{MSFE}(\hat{y}(w)) = \sum_{i=1}^{N} E\|w_1 b + w_2 \hat{\beta}_i - \beta_i\|_A^2$$

$$= \sum_{i=1}^{N} E\|w_1 Q^{-1} \sum_{i=1}^{N} Q_i \hat{\beta}_i + w_2 \hat{\beta}_i - \beta_i\|_A^2$$

$$= \sum_{i=1}^{N} \|w_1 Q^{-1} \sum_{i=1}^{N} Q_i \beta_i + w_2 \beta_i - \beta_i\|_A^2 + \sum_{j=1}^{N} \text{var}(w_1 A_i^{1/2} Q^{-1} \sum_{i=1}^{N} Q_i \hat{\beta}_i + w_2 A_i^{1/2} \hat{\beta}_j)$$

$$= w_1^2 \text{MSFE}_{pool} + w_2^2 \text{MSFE}_{ind} + 2w_1 w_2 \frac{1}{T} \sum_{i=1}^{N} \sigma_i^2 \text{tr}(Q^{-1} A_i).$$

(11)
The comparison will be even more clear if all regressors are normalized, such that \( Q_i = I_k \) and thus \( Q = NI_k \). In this case, we have

\[
\text{MSFE}_{\text{ind}} = k \sum_{i=1}^{N} \sigma_i^2, \quad \text{MSFE}_{\text{pool}} = \sum_{i=1}^{N} \| \bar{\beta} - \beta_i \|_{A_i}^2 + \frac{k}{N} \sum_{i=1}^{N} \sigma_i^2, \quad (12)
\]

and

\[
\text{MSFE}(\hat{y}(w)) = w_1^2 \sum_{i=1}^{N} \| \bar{\beta} - \beta_i \|_{A_i}^2 + w_2^2 \frac{(N-1)k}{N} \sum_{i=1}^{N} \sigma_i^2 + \frac{k}{N} \sum_{i=1}^{N} \sigma_i^2, \quad (13)
\]

where \( \bar{\beta} = N^{-1} \sum_{i=1}^{N} \beta_i \). Comparing the pooling averaging and pooled forecast, we see that \( \text{MSFE}(\hat{y}(w)) < \text{MSFE}_{\text{pool}} \) if and only if

\[
\sum_{i=1}^{N} \| \bar{\beta} - \beta_i \|_{A_i}^2 > \frac{w_2^2}{1-w_1^2} \cdot \frac{(N-1)k}{N} \sum_{i=1}^{N} \sigma_i^2. \quad (14)
\]

This suggests that the pooling averaging forecast is superior to the pooled if the difference between individual coefficients is large enough. In the extreme case of a completely homogeneous panel \( \sum_{i=1}^{N} \| \bar{\beta} - \beta_i \|_{A_i} = 0 \leq w_2^2 (N-1)k \sum_{i=1}^{N} \sigma_i^2/[N(1-w_1^2)] \), it always holds that \( \text{MSFE}(\hat{y}(w)) \geq \text{MSFE}_{\text{pool}} \) as expected. It can also be seen from (14) that the pooled forecast is more likely to outperform the pooling averaging when the variance of the errors \( \sigma_i^2 \) and/or the number of regressors \( k \) increase.

When we compare the pooling averaging with the individual forecasts, we have that \( \text{MSFE}(\hat{y}(w)) < \text{MSFE}_{\text{ind}} \) if and only if

\[
\sum_{i=1}^{N} \| \bar{\beta} - \beta_i \|_{A_i}^2 < \frac{1-w_2^2}{w_1^2} \cdot \frac{(N-1)k}{N} \sum_{i=1}^{N} \sigma_i^2. \quad (15)
\]

Inequality (15) shows that pooling averaging is advantageous over the individual time series forecast when coefficient heterogeneity is bounded by the product of \((N-1)k \sum_{i=1}^{N} \sigma_i^2/N\) (since \((1-w_2^2)/w_1^2 > 1\)). Even if the panel is completely heterogeneous with all coefficients different across individuals, pooling averaging can still outperform the individual forecast when the variance of the errors is large or when there are too many explanatory variables in the model. Or in other words, large error variances favor the pooling averaging approach as the inequality (15) is more likely to hold. These arguments will be confirmed by our simulation study.
Choosing Pooling Averaging Weights

We have seen in the previous section that the pooling averaging can make an appropriate trade-off between bias and variance, depending on how the weights are chosen. In this section, we discuss how to choose the appropriate pooling averaging weights.

5.1 Mallows Pooling Averaging

We propose to choose the weights based on the Mallows criterion. Using Mallows criterion to average the models is initiated by Hansen (2007), which is asymptotically optimal in the sense of achieving the lowest possible squared error. This method is further justified by Wan et al. (2010). To derive the Mallows pooling averaging (MPA) criterion in estimating a heterogeneous panel, we define \( \| \theta \|^2_A = \theta' A \theta \) for any vector \( \theta \) and non-negative definite matrix \( A \). The choice of \( A \) depends on whether the interest is in forecasting or coefficient estimates. If forecasting is of the main interest, we set \( A = X'X \); otherwise, we set \( A = I_{Nk} \).

We generalize the heteroskedastic structure of the error terms and now allow for conditional heteroskedasticity between and within individuals (completely heteroskedasticity). Hence, if we define \( \Omega_i = \text{var}(u_i) \) and \( \Xi_i = X'_i \Omega_i X_i / T \), the variance of the individual coefficient estimator \( \hat{\beta}_i \) can be written as \( V_i = Q_i^{-1} \Xi_i Q_i^{-1} / T \) with \( Q_i = X'_i X_i / T \). When the errors are between-individual heteroskedastic, \( V_i \) reduces to \( \sigma^2_i Q_i^{-1} / T \) as used before in Section 4.

Under squared loss \( L_A(w) = \| \hat{\beta}(w) - \beta \|^2_A \) and squared risk \( R_A(w) = E\{L_A(w)\} \), the Mallows criterion can be written as

\[
C_A(w) = \| P(w) \hat{\beta} - \beta \|^2_A + 2 \text{tr}(P'(w)AV) - \| \hat{\beta} - \beta \|^2_A,
\]

where \( V = \text{diag}(V_1, \ldots, V_N) \) as defined in Section 4. This criterion is a good approximate for the MS(F)E, in the sense that it is an unbiased estimator of the squared risk under regular conditions.\(^3\) We choose weights by minimizing the above criterion, that is

\[
\hat{w} = (\hat{w}_1, \ldots, \hat{w}_M)' = \arg \min_{w \in \mathcal{W}} C_A(w).
\]

Note that although \( \| \hat{\beta} - \beta \|^2_A \) appears in \( C_A(w) \), it is not affected by the choice of \( w \). If we denote \( D = (\hat{\beta}_1(1) - \hat{\beta}, \ldots, \hat{\beta}_M(1) - \hat{\beta}) \) and \( v = \{ \text{tr}(P'_1 Av), \ldots, \text{tr}(P'_M Av) \} \)', then we can rewrite (16) as

\[
C_A(w) = w'D'ADw + 2w'v - \| \hat{\beta} - \beta \|^2_A,
\]

\(^3\)See the supplementary file for the proof.
which is clearly a quadratic function in \( w \).

The criterion \( C_A(w) \) is a generalization of the Mallows model averaging criterion defined by Hansen (2007). When \( \Omega_i = \sigma^2 I_T \) for all \( i \) and \( A = X'X, C_A(w) \) simplifies to Hansen’s criterion (Equation (11) in Hansen (2007)), which focuses on the average (forecasting) squared error loss \( (X\hat{\beta}(w) - E(y))'(X\hat{\beta}(w) - E(y)) \). When we set \( A = I_{Nk} \), (16) extends Hansen’s (2007) criterion to concentrate on the accuracy of the estimated coefficients, and \( C_A(w) \) aims at minimizing the average squared error of coefficient estimates. It is worth noting that if we only average the pooled and individual estimators, then the Mallows pooling averaging estimator is essentially a Stein-rule estimator (see Equation (2) of Maddala et al. (1997)). The weights of Mallows pooling averaging estimator and Maddala et al.’s (1997) shrinkage estimator are proportional to each other.\(^4\)

In practice, the covariance matrix \( V \) is unknown, and has to be replaced by its estimate \( \hat{V} \). A feasible version of (16) is

\[
C_A^*(w) = \|P(w)\hat{\beta} - \hat{\beta}\|^2_A + 2\text{tr}[P'(w)A\hat{V}] - \|\hat{\beta} - \beta\|^2_A,
\]

and the feasible weight vector is obtained by

\[
\hat{w}^* = \arg\min_{w \in W} C_A^*(w).
\]

Depending on the assumptions of the error structure, the covariance matrix \( V \) can be estimated as follows:

1. Homoscedasticity: If we assume that \( \text{var}(u_i) = \sigma^2 I_T \) for all \( i \), we estimate \( V \) by

\[
\hat{V}_{\text{homo}} = \hat{\sigma}^2(X'X)^{-1},
\]

where \( \hat{\sigma}^2 \) is the variance of residuals associated with from the individual OLS estimator, i.e. \( \hat{\sigma}^2 = (Y - X\hat{\beta})'(Y - X\hat{\beta})/(NT - Nk) \).

2. Between-individual heteroskedasticity: If we assume that \( \text{var}(u_i) = \sigma^2_i I_T \), we consider

\[
\hat{V}_{\text{bh}} = \text{diag}(\hat{\sigma}^2_1 Q_1^{-1}, \ldots, \hat{\sigma}^2_N Q_N^{-1})/T,
\]

where \( \hat{\sigma}^2_i = \hat{u}_i^2/(T-k) \) and \( \hat{u}_i \) is the OLS residual of the \( i \)-th individual regression.

3. Completely (conditional) heteroskedasticity: If we assume that \( u_i \) is (conditional) heteroskedastic for each \( i = 1, \ldots, N \), the most general situation, we use

\[
\hat{V}_{\text{ch}} = \frac{1}{T(T-k)} \text{diag} \left( Q_1^{-1} \sum_{t=1}^T \hat{u}_{1t}^2 X'_{1t} X_{1t} Q_1^{-1}, \ldots, Q_N^{-1} \sum_{t=1}^T \hat{u}_{Nt}^2 X'_{Nt} X_{Nt} Q_N^{-1} \right),
\]

where \( \hat{u}_{it} \) is the \( t \)-th element of \( \hat{u}_i \) for \( i = 1, \ldots, N \) and \( t = 1, \ldots, T \).

\(^4\)See the supplementary file for the details and proof.
5.2 Finite sample and asymptotic properties

Given the fact that the variance of individual estimators vanishes under fixed $N$ and large $T$ resulting in no bias-variance trade-off as discussed in Section 4, we mainly discuss the properties of the Mallows pooling averaging estimator under finite sample and large $N$, $T$ asymptotics. We first examine the finite sample property of the MPA estimator by deriving its risk bounds. The risk bound is widely used as an important theoretical property (or justification) of an estimation procedure (see, e.g., Yuan and Yang, 2005). It tells us how the Mallows pooling averaging performs in the worst situation, and we can examine how this bound depends on various features of data.

**Theorem 1.** The upper bound of the risk of MPA estimator is

$$E\{L_A(\hat{w})\} \leq \frac{1}{1 - c} \inf_{w \in \mathcal{W}} R_A(w) + \frac{1}{1 - c} \left( \frac{1}{c} \text{tr}(AV) - 2E(\text{tr}\{P'(\hat{w})AV\}) \right),$$  \hspace{1cm} (21)

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

**Proof:** See Appendix A.1.

It shows that up to the constant $(1 - c)^{-1}$ and the additive penalty $(1 - c)^{-1}[c^{-1}\text{tr}(AV) - 2E(\text{tr}\{P'(\hat{w})AV\})]$, the pooling averaging estimator $\hat{\beta}(\hat{w})$ achieves the performance of the optimal-weight pooling averaging estimator $\inf_{w \in \mathcal{W}} R_A(w)$. The result of (21) does not depend on sample size. To further examine how risk bounds depend on various quantities that characterize the data, let $\mathcal{I}_1(\cdot)$ and $\mathcal{I}_2(\cdot)$ denote the minimum and maximum eigenvalues of a symmetric matrix. The following corollary provides the specific risk bounds for coefficient estimation and forecasting.

**Corollary 1.** If there exist constants $\bar{\Omega}$ and $c_1$ such that $\max_{i=1,\ldots,N} \mathcal{I}_2(\Omega_i) \leq \bar{\Omega}$ and $\min_{i=1,\ldots,N} \mathcal{I}_1(Q_i) \geq c_1$, then there exists $c \in (0, 0.5)$ such that when $A = I_{Nk}$,

$$E\{L_A(\hat{w})\} \leq \frac{1}{1 - c} \inf_{w \in \mathcal{W}} R_A(w) + \frac{1 - 2c}{c(1 - c)} \frac{Nk\bar{\Omega}}{Tc_1} + \frac{4}{1 - c} \frac{Nk\bar{\Omega}}{Tc_1},$$  \hspace{1cm} (22)

and when $A = X'X$,

$$E\{L_A(\hat{w})\} \leq \frac{1}{1 - c} \inf_{w \in \mathcal{W}} R_A(w) + \frac{1 - 2c}{c(1 - c)} Nk\bar{\Omega} + \frac{4}{1 - c} Nk\bar{\Omega}.$$  \hspace{1cm} (23)
The implied risk bounds are particularly informative, as they demonstrate how the performance of the Mallows pooling averaging estimator and forecast is determined by\(\inf_{w \in W} R_A(w)\) and a set of constants \(\{T, N, k, \bar{\Omega}, c_1, c\}\) in the worst situation. As expected, the risk bounds (in both cases of \(A = I_{Nk}\) and \(A = X'X\)) are 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 when we focus on coefficient estimation \((A = I_{Nk})\). Our simulation studies in Section 7 will provide numerical evidence of the effect of these constants.

Next, we study the asymptotic property of MPA estimator following the model averaging literature. We assume that the following conditions hold when \(T, N \to \infty\).

\begin{align*}
C.1: & \ X'_i u_i = O_p(T^{1/2}) \text{ uniformly for } i = 1, \ldots, N. \\
C.2: & \ 0 < c_1 \leq \min_{i \in \{1, \ldots, N\}} I_1(T^{-1}X'_iX_i) \leq \max_{i \in \{1, \ldots, N\}} I_2(T^{-1}X'_iX_i) \leq c_2 < \infty. \\
C.3: & \ MNT^{-1/2}\xi_{NT}^{-1}\mathcal{I}_2(A) \to 0 \text{ where } \xi_{NT} = \inf_{w \in W} R_A(w).
\end{align*}

Condition C.1 ensures that each individual estimation is consistent. Condition C.3 requires that candidate models are approximations. For \(A = X'X\), we know that a necessary condition of C.3 is \(\xi_{NT}^{-1} = o(M^{-1}N^{-1}T^{-1/2})\), which is similar to the condition (7) of Ando and Li (2014). For \(A = I_{Nk}\), C.3 simplifies to \(MNT^{-1/2}\xi_{NT}^{-1} \to 0\), which constrains the rate of \(\xi_{NT} \to 0\). C.1 and C.3 are not contradictory, because they require that candidate models are correctly specified on the regressors, but misspecified on the pooling. Condition C.3 is of particular relevant when a preliminary model screening step is taken to shrink the model space. We will clarify this point in the next section.

**Theorem 2.** As \(T \to \infty\) and \(N \to \infty\), if Conditions C.1–C.3 are satisfied, then

\[
\frac{L_A(\hat{w}^*)}{\inf_{w \in W} L_A(w)} \to 1,
\]

in probability, regardless of \(\hat{V} = \hat{V}_{\text{homo}}, \hat{V} = \hat{V}_{\text{bh}}\) or \(\hat{V} = \hat{V}_{\text{ch}}\).

**Proof:** See Appendix A.3.

The result (24) means that MPA estimator \(\hat{\beta}(\hat{w})\) is asymptotically optimal in the sense that
its squared loss is asymptotically identical to that of the infeasible best possible model-
averaging estimator. This optimality statement is conditional on the given set of estimators as in Hansen (2007).

One issue of this method is that it only provides a point estimate for the coefficients. Conducting inference is generally challenging for the model-averaging estimators. One possible way is to adopt the local asymptotic framework. However, it is not immediately obvious what “local misspecification” means here. It seems less relevant and also difficult, if not impossible, to analyse local misspecification in explanatory variables, because one cannot separate the biased caused by heterogeneity and omitted variables. In this paper, we propose an alternative way to conduct statistical inference in practice. We calculate the variance and confidence interval of the Mallows pooling averaging estimator by bootstrap. In particular, we implement cross-sectional resampling following Kapetanios (2008) for $B$ times, and compute the coefficient estimates for each resampling. The empirical variance and confidence interval can then be calculated based on $B$ coefficient estimates.

6 Shrinking Model Space

In practice, the number of ways of imposing restrictions on the regression parameters can be numerous for moderate and large $N$. 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. This creates a huge model space, and selecting or averaging over the entire model space is computationally difficult. In this case, a preliminary step of model screening is desirable. Model screening rules out the poor models that are obviously inferior, thus significantly reducing the number of candidate models to estimate and average over. This not only facilitates computation, and also improves estimation efficiency by reducing the number of (weight) parameters and by not averaging estimators from poor models. We first provide theoretical justification for the use of model screening in general, and then propose a specific approach and discuss its properties.

First, to justify model screening, we let $\mathcal{M}^*$ be a subset of $\{1, \ldots, M\}$ and $\mathcal{W}^* = \{w \in [0,1]^M : \sum_{m \in \mathcal{M}^*} w_m = 1$ and $\sum_{m \notin \mathcal{M}^*} w_m = 0\}$ be a subset of $\mathcal{W}$. The model-averaging estimator based on the subset $\mathcal{M}^*$ is obtained by using the weight vector $\hat{w}^* = \arg \min_{w \in \mathcal{W}^*} C_A(w)$. We make the following assumption:

C.4: There exist a non-negative series of $\nu_{NT}$ and a weight series of $w_{NT} \in \mathcal{W}$ such
that \( \xi_{NT}^{-1} \nu_{NT} \to 0 \), \( \inf_{w \in \mathcal{W}} C_A^*(w) = C_A^*(w_{NT}) - \nu_{NT} \), and \( P(w_{NT} \in \mathcal{W}^*) \to 1 \) as \( N, T \to \infty \).

Under Condition C.1–C.4, we can follow the proof of Theorem 3 of Zhang et al. (2016) and show that the post-screened model-averaging estimator based on the candidate model set \( \mathcal{M}^* \) still achieves the asymptotic optimality, namely

\[
\frac{L_A(\hat{w}^*)}{\inf_{w \in \mathcal{W}} L_A(w)} \to 1.
\]

Since the individual estimator is typically screened out by this procedure, this optimality theorem provides particular theoretical support for post-screened model-averaging estimators because of Condition C.3.

Next, for practical purpose, we need a procedure that can rule out the “poor” models that incorrectly impose equality restrictions on far different coefficients, while at the same time maintaining the “good” models. We propose to implement model screening based on estimating panel structural models with different choices of the number of groups. A panel structural model assumes that individual units are classified into groups, and individuals in the same group share a common slope coefficient vector (see, e.g., Su et al., 2016). To estimate this model, we employ classifier-Lasso (C-Lasso) proposed by Su et al. (2016). Obviously, each way of classifying individual units corresponds to a specific pooling strategy. If the number of groups is correctly chosen, one can consistently estimate group membership and slope coefficients. Even when the number is misspecified, C-Lasso provides good estimates of group membership and slope coefficients under this misspecified number by minimizing the penalized least squared objective function. Hence, by only considering estimates obtained from C-Lasso with different numbers of groups, we rule out the poor classification that pools far different individuals together.

There are several advantages of using C-Lasso for model screening. First, it is less arbitrary compared to clustering based on artificially chosen thresholds. Second, it produces estimates with well-understood and desired statistical properties. In particular, when the number of groups is correctly or over-specified, the estimated slope coefficients are consistent but less efficient, while underspecification of the number of groups gains more efficiency but leads to inconsistent estimates (Bonhomme and Manresa, 2015; Su et al., 2016). Therefore, the consistency-efficiency trade-off remains valid for the post-screening model space. Third, it is computationally more attractive than \( k\text{-means} \) since it can be transformed into
a sequence of convex problems and does not depend on the initial values. To avoid the danger of making arguments that are sensitive to our choice of screening procedures, we consider a variety of clustering methods, for example agglomerative hierarchical clustering and *k-means*. Unreported simulation results suggests that our main results are not affected.\(^5\)

Interestingly, model averaging also offers an effective way of addressing uncertainty in determining the number of groups when estimating panel structural models, especially when forecasting is of central interest. Although one can consistently estimate the slope coefficients under the correctly postulated number of groups, these consistent estimates are not necessary the most accurate in terms of minimum MSE. Nor do they necessarily result in the best forecast in terms of minimum MSFE. Instead of selecting the number of groups based on information criteria or testing procedures, one can average estimates obtained from different numbers of groups. Given the trade-off between consistency and efficiency for different choices of the number, one can make an optimal trade-off by averaging and appropriately choosing weights. The optimal weight choice depends on whether the focus is on parameter estimation or forecasting. In the next section, we shall compare the estimates/forecasts obtained from the selected best number of groups and those from averaging over different numbers of groups.

7 Monte Carlo Study

To support our theoretical claims and to shed more light on the performance of screening and pooling strategies, we consider in this section several Monte Carlo experiments.

7.1 Simulation Designs

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

\[
y_{it} = \sum_{l=1}^{3} x_{l,it} \beta_{il} + \epsilon_{it}, \quad i = 1, \ldots, N; \quad t = 1, \ldots, T, \tag{25}
\]

where \(x_{i1t} = 1\) and the remaining regressors are independently generated from the standard normal distributions. To mimic the empirical data of sovereign CDS spreads, we also

\(^5\)See the supplementary file for details.
consider regressors to follow an autoregressive (AR) process, i.e. \( x_{it} = 0.6x_{i,t-1} + v_{it} \), where \( v_{it} \) is i.i.d. normal. The idiosyncratic errors \( \epsilon_{it} \) are uncorrelated with regressors and independently normally distributed with mean zero and variance \( \sigma_{\epsilon_i}^2 \). We consider conditional heteroskedasticity, such that the variance of errors vary across individuals and its size depends on a pre-specified value of \( R^2 \). The slope coefficients are allowed to have different grouping patterns. In particular, we consider four cases with different degrees of heterogeneity in coefficients

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

DGP 2 (Weakly heterogeneous):
\[
\beta_{i1}, \beta_{i2} = \begin{cases} 
1, & i = 1, \ldots, \lfloor N/2 \rfloor \\
3, & i = \lfloor N/2 \rfloor + 1, \ldots, N 
\end{cases} \quad \beta_{i3} = \begin{cases} 
1, & i = 1, \ldots, \lfloor N/3 \rfloor \\
3, & i = \lfloor N/3 \rfloor + 1, \ldots, N 
\end{cases}
\]
where \( \lfloor N/2 \rfloor \) denotes the nearest integer value that is smaller than \( N/2 \).

DGP 3 (Strongly heterogeneous):
\[
\beta_{i1}, \beta_{i2} = \begin{cases} 
1, & i = 1, \ldots, \lfloor N/4 \rfloor \\
2, & i = \lfloor N/4 \rfloor + 1, \ldots, \lfloor N/2 \rfloor \\
3, & i = \lfloor N/2 \rfloor + 1, \ldots, \lfloor 3N/4 \rfloor \\
4, & i = \lfloor 3N/4 \rfloor + 1, \ldots, N 
\end{cases} \quad \beta_{i3} = \begin{cases} 
1, & i = 1, \ldots, \lfloor N/5 \rfloor \\
2, & i = \lfloor N/5 \rfloor + 1, \ldots, \lfloor 2N/5 \rfloor \\
3, & i = \lfloor 2N/5 \rfloor + 1, \ldots, \lfloor 3N/5 \rfloor \\
4, & i = \lfloor 3N/5 \rfloor + 1, \ldots, N 
\end{cases}
\]

DGP 4 (Completely heterogeneous): \( \beta_{il} = 0.1 \times i \times l \) for all \( i \) and \( l \).

The sample size varies from \( N \in \{10, 30\} \) and \( T \in \{20, 40\} \), leading to four combinations of \( N \) and \( T \). To save space, presentation in this section is based on \( A = X'X \), focusing on the forecasting performance. The simulation results with \( A = I_{NK} \), focusing on the slope coefficient estimates, are similar.\(^7\)

We compare the forecasting performance of Mallows pooling averaging with the pooled model, individual time series model, shrinkage estimator (7), a single pooling model selected by AIC or BIC (pretesting), pooling averaging using relative values of AIC or BIC as

\(^6\)We also consider the case where the intercept is heterogeneous across individuals but slope coefficients are homogeneous. In this case, a preliminary within-transformation leads to the similar results, and therefore not reported.

\(^7\)See our supplementary file for details.
weights (smoothed AIC/BIC), and a C-Lasso estimator with the number of groups determined by BIC as in Su et al. (2016). All pooling averaging and information-criterion-based forecasts are constructed from the preliminary model screening method using C-Lasso as described in Section 6.\(^8\) To compute the Mallows pooling averaging forecast, we proposed three versions of variance-covariance matrix estimator in Section 5.1. Although these variance estimators are especially designed for specific error distributions, it is not guaranteed that one method would always produce lower MSFE than the other in finite sample. Therefore, we consider three versions of variance-covariance estimators for Mallows pooling averaging in each case, and report the best choice, although the results of using different estimators are highly similar in all cases except DGP 1. Our simulation results are based on 1000 replications.

We evaluate all methods based on the risk (expected squared loss) following Hansen (2007). All numbers are normalized with respect to the individual time series forecast, such that the number of individual forecast is always 1 and thus not reported. We emphasize that the purpose of the simulation studies is not to show the dominant superiority of a method in all cases. Instead, we try to demonstrate that the performance depends on several factors, thus providing evidence for the theory in Section 4. Also, we aim to understand how pooling averaging behaves in various situations. Based on simulation results, we provide applied researchers with some practical rules which methods are more likely to give reliable results in a particular situation.

### 7.2 Results

We first present the results of the benchmark case with independent regressors and \(R^2 = 0.9\). Then we consider the situation of autoregressive regressors and smaller signal-to-noise ratio.

Insert TABLE 1 about here

Table 1 presents the results with i.i.d regressors and a large \(R^2\). In DGP 1 of a homogeneous panel the pooled forecast always performs best as expected. The proposed MPA

\(^8\)The tuning parameter for C-Lasso is chosen by trying different values and selecting the best one in terms of risk.
forecast is the second best in most of the cases, suggesting that the Mallows criterion can still assign rather good weights in homogeneous panels. When the panel is characterized by some degree of heterogeneity (DGP 2 and 3), MPA forecasts dominate others in all cases. Particularly, MPA produces the minimum risk in 7 out of 8 cases, while C-Lasso with the number of groups selected by the information criterion (IC) performs best in DGP 2 when $N = 30$ and $T = 40$. For the completely heterogeneous DGP 4, we find that MPA keeps the best in most of the cases, outperforming the individual forecast. The forecast based on the shrinkage estimator performs rather well, and is the best when $N = 10$ and $T = 40$. This observation may seem counterintuitive at the first glance, since one may expect that the individual forecast should perform well in completely heterogeneous panels. 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 heterogeneous. This is because although the individual estimators are unbiased, they are inefficient, especially under small $T$ or large noise. On the contrary, pooling averaging makes good use of cross-section variation and thus provides a more accurate forecast. Interestingly, we find that in most cases MPA produces better forecasts than C-Lasso based on a single selected number of groups using IC. An exception is DGP 2 under large $N$ and $T$, where C-Lasso slightly outperforms MPA. Hence, if the forecasting is of central interest, averaging offers a sensible alternative to handle the uncertainty of the number of groups, especially when the sample size is limited.

Table 2 presents the results under AR(1) regressors with a large $R^2$, and this setup allows us to analyse the effect of autocorrelation. The results are generally similar to the case of independent regressors. The pooled forecast is still the best in homogeneous panels. MPA keeps on performing best in heterogeneous panels. Again, MPA outperforms C-Lasso in all cases except in DGP 2 when $N$ and $T$ are large.

So far, the results are all based on DPGs with $R^2 = 0.9$. The theory in Section 4 suggests that more noise will weaken the advantages of pooling averaging estimates, but
support the use of the pooled forecast. 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 $R^2 = 0.7$ and relatively poor fitness with $R^2 = 0.5$. Representative results are reported in Table 3. The upper panel of Table 3 shows the results under $R^2 = 0.7$. In this case, MPA remains the best in heterogeneous panels (DGP 2 to 4), but the difference between MPA and the pooled forecast is much smaller compared to the cases with $R^2 = 0.9$, especially in partially heterogeneous panels with small sample size. This is because the estimation error is particularly sizeable in small samples, and also efficiency loss is more likely to dominate the bias if the degree of heterogeneity is weak. This result firmly supports the theoretical argument in Section 4. If we further decrease the $R^2$ to $0.5$ as in the bottom panel of Table 3, we see that performance of the pooled and MPA are further closer, both of which dominantly outperform other rivals. The pooled forecast sometimes even performs the best in heterogeneous panels. This is expected, because when data are highly noisy, efficiency becomes more important. Thus the efficiency gain of the pooled forecast dominates the bias in this situation. Besides, we also see that when the signal-to-noise ratio is low, suggesting a larger degree of uncertainty in deciding the number of groups, the superiority of MPA to C-Lasso based on a single selected number of groups is even more obvious.

In general, we find MPA performs robustly well in panels with various degrees of heterogeneity. When the signal-to-noise ratio is moderate or high, MPA dominates other methods. When the signal-to-noise ratio is relatively low (reflected by a small $R^2$ and/or a small $T$), MPA tends to assign most weights to the pooled model, gaining most efficiency, and thus still remains one of the best choices.

8 Explain and Forecast Sovereign Credit Risk

Now we have discussed the optimal pooling strategy in heterogeneous panels, we utilize cross-country panel data to explain and forecast the sovereign credit risk using the proposed method. Since the breakout of a wide range of financial crises, many countries have experienced a dramatic increase in their government debts. The large size of government debt has attracted extensive attention to the sovereign credit risk. It is of key importance for both policy makers and financial market agents to understand the nature of sovereign credit risk and to forecast the future risk. Acknowledging the determinants of sovereign
credit risk is helpful to control the risk, and a good forecast is crucial for effective policy making and diversifying the risk.

We focus on the sovereign credit default swap (CDS) spreads as a proxy of sovereign credit risk. A CDS contract is an insurance contract that protects the buyer from the credit event, e.g. a loan default. Its spread, expressed in basis points, is the insurance premium that buyers have to pay, and thus reflects the credit risk. To examine the determinants of sovereign CDS spreads and forecast its future values, we collect the most recent cross-country data on sovereign CDS spreads and financial indicators of macroeconomic fundamentals. In particular, we follow Longstaff et al. (2011) to focus on spreads of five-year sovereign credit default swaps, and associate the CDS spreads with a set of local and global variables. The local variables include local stock market returns ($lstock$), 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 ($trsy$), high-yield corporate bond spreads ($hy$), equity premium ($eqp$), volatility risk premium ($volp$), equity flows ($ef$), and bond flows ($bf$). The data set is an updated version of Longstaff et al. (2011) (see Longstaff et al. (2011) for a detailed definition of the variables). To have a balanced panel, the updated data set contains 14 countries, i.e. Brazil, Bulgaria, Chile, China, Hungary, Japan, Korea, Malaysia, Philippines, Poland, Romania, Slovak, South Africa, and Thailand, and we use the monthly data starting from January 2003 to January 2016 resulting in 156 time observations.

Recently, an increasing number of studies have tried to associate the changes of CDS spreads with various macroeconomic fundamentals, see among others, Dieckmann and Plank (2012); Longstaff et al. (2011); Aizenman et al. (2013). Despite the availability of a cross-country panel, most of these studies analyse 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. More importantly, individual-country studies can be rather inefficient since the cross-country information is not utilized at all, especially when the time-series dimension is not extremely long. In our application, the entire time span contains 156 observations, which is not a particular large sample to estimate the effects of a relatively large number of determinants for each individual country separately, especially when breaks are likely to be present. Furthermore, given the prevailing financial crises, there are likely structural breaks in the effect
of determinants (Dieckmann and Plank, 2012; Qian et al., 2017). In the presence of time instability, ignoring the breaks and estimating slope coefficients or making forecasts using the whole sample of time series observations may not be the best strategy. Instead, it may require subsample analysis to better understand the time-varying nature of the CDS spreads and perhaps provide a more accurate forecast. This implies that the length of the estimation window span could be even shorter, resulting in a larger degree of efficiency loss for individual time series estimation. Hence, the bias-efficiency trade-off is especially important and appropriate pooling is highly desired.

To examine the determinants of sovereign CDS spreads and forecast its future values, we consider the following model

$$\Delta CDS_{it} = \alpha_i + X'_{i,t-1}\beta_i + \varepsilon_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T,$$

(26)

where $\Delta CDS_{it}$ is the first-differenced sovereign CDS spread of country $i$ at time $t$, $X_{it}$ is a $10 \times 1$ vector of covariates, and $\varepsilon_{it}$ is the error term. The change variable is used as the dependent variable, following Longstaff et al. (2011). Preliminary unit root tests show that the CDS spread changes of all countries are generally stable. The lagged covariates are used mainly for the forecasting purpose, and they also reduce possible reverse causality from CDS spreads to macroeconomic fundamentals, if not completely remove it. Note that the slope coefficients are allowed to be heterogeneous across countries. We normalize all covariates to make the slope coefficients of individuals comparable.

### 8.1 Structural Break Detection

We first examine whether there exist obvious structural breaks. To detect and date possible structural breaks, it is important to incorporate individual heterogeneity in slope coefficients. We employ the recently developed break detection method by Baltagi et al. (2016) that explicitly allows for heterogeneous slope coefficients. If there is one break, the estimated break point $\hat{k}$ can be obtained by minimizing the following sum of squared

9If the interest lies on forecasting, how to deal with structural breaks is more complicated, since it is not guaranteed that the forecast based on the post-break subsample always outperforms the one using the whole sample period, see also Pesaran et al. (2013). Nevertheless, completely ignoring structural breaks and using the whole sample of periods without careful analysis is not an appropriate approach.
residual function\(^{10}\) as
\[
\hat{k} = \arg \min_{1 < k < T - 1} \sum_{i=1}^{N} \sum_{t=1}^{T} \left[ \Delta CDS_{it} - X'_{it} \hat{\beta}_i(k) - Z'_{it} \hat{\delta}_i(k) \right] \left[ \Delta CDS_{it} - X'_{it} \hat{\beta}_i(k) - Z'_{it} \hat{\delta}_i(k) \right],
\]
where \(Z_{it}(k) = X_{it} I(t \geq k)\), \(I(\cdot)\) is an indicator function, and \(\hat{\beta}_i(k)\) and \(\hat{\delta}_i(k)\) are the least-squared estimator of regressing \(\Delta CDS_{it}\) on \((X_{it}, Z_{it})\). For multiple structural breaks, we follow Baltagi et al. (2016) to use the sequential break detection technique that estimate break points one by one. To determine the number of breaks, we employ Bayesian information criterion, and it reports 2 structural breaks. The two estimated break points corresponds to January 2009 and December 2009, which closely match important events during the global financial crisis in 2009. Particularly, although US had already entered recession in the end of 2007, a wide range of global crisis broke out in late 2008 when Lehman Brothers declared bankruptcy and a number of European countries fell into banking crisis. After several negative signals on the financial markets released in the last three months of 2008, e.g. a prediction of a deep recession in the UK by The Times and S&P’s sovereign credit rating cut for a number of countries, the global economies became highly unstable from January 2009, and it lasted for roughly a complete year. The two break points result in 72 time observations in first regime, 11 in the second regime, and 73 in the final regime. Given the moderate size of sample in each regime, efficiency is an important concern, and it is particularly useful to make use of cross-section similarity and consider appropriate pooling for estimation and forecasting purposes.

8.2 Effects of Local and Global Variables

We first examine the effect of determinants on sovereign CDS spreads using Mallows pooling averaging. We estimate (26), respectively, for the three regimes segmented according to the two estimated structural break points. To reduce the model space and facilitate computation, we employ the pre-screening approach described in Section 6. The maximum number of groups \(G_{\max}\) is set to be \(N/2\), and we average grouped estimators obtained from \(G = 1, \ldots, G_{\max}\). Robustness analysis suggests that results are stable for a reasonably wide range of \(G_{\max}\). The implementation of C-Lasso requires a tuning parameter. We follow Su et al. (2016) to consider the tuning parameter \(\lambda = c_\lambda s^2_T T^{1/3}\), where \(c_\lambda = \{0.0625, 0.125, 0.25, 0.5, 1\}\) is a geometrically increasing sequence and \(s^2_T\) is the sample

\(^{10}\)No weighted average of individuals is needed since variables are all normalized.
variance of $y_{it}$. From the set of $\lambda$s, we pick up the value that minimizes the risk like in the simulation. To compute the Mallows criterion, the completely heteroskedastic variance structure is assumed. Finally, like other model-averaging estimators, MPA does not directly provide a variance estimator. To make an inference, we employ the bootstrap method to obtain standard errors as suggested in Section 5.1.

We first examine the overall time-varying pattern and the degree of heterogeneity of MPA estimated slope coefficients, and then study how the effect of determinants varies across country in details. Figure 1 presents the boxplot of the MPA estimates of slope coefficients of six most significant and salient determinants, i.e. local and global stock returns, treasury yields, high-yield corporate bond spreads, equity premium, and equity flows. In each panel of the figure, we plot the estimates in three regimes separated by the two break points. The height of the boxes shows the extent of cross-country heterogeneity, and thus we can see how the average effect and the extent of heterogeneity vary over regimes. In general, the effect of these determinants demonstrates a strong time-varying pattern, and the cross-country heterogeneity is more prominent after the onset of the crisis, especially in the second regime of the unstable crisis period. In particular, local stock returns seem to play a relatively weak role for all countries before the crisis, but it has a much stronger but also more heterogeneous impact after the breakout of the crisis. The effect of global stock returns is strongly negative for all countries before the crisis, which is in line with most literature. However, this effect again becomes more heterogeneous after the first break, and interestingly, turns positive for a proportion of countries. The switching role of global stock returns implies that the global market performance can be negatively or positively related to a sovereign’s CDS spreads. The negative association is because a good global economic situation can positively influence the domestic economy and reduces the local CDS spreads, while the positive association is due to the substitution of the two markets for investors. The third important determinant is treasury yields, which has a more negative and heterogeneous effect after the crisis than before the crisis. The effect of high-yield corporate bond spreads is robustly negative for all countries before the crisis, but more heterogeneous afterwards. Equity premium generally has a positive impact
on sovereign CDS spreads both before and after the crisis, while the effect of equity flows is similar across countries in the first and third regime, but highly heterogeneous in the second regime.

Next, we are particularly interested in the featuring determinants of sovereign CDS spreads for each individual country, especially those that are characterized by large amount of government debt. Hence, in Table 4 we report the MPA estimates of regime-specific slope coefficients for individual countries whose ratio of gross government debt to GDP is more than 50% in 2016. 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.

We first examine the local variables. The effect of local stock returns is mostly insignificant in the first regime before 2009. In the unstable second regime, this effect becomes significantly negative for Brazil and China, but remains unclear for the other countries. After 2009, we see a significantly negative association between local stock returns and sovereign CDS spreads for most countries except Hungary and Slovak. 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. Nevertheless, the importance of this determinant varies across regimes and countries. The effect is especially prominent during and after the crisis, and it appeals stronger in big economies, such as Brazil and China, than in small ones, such as Hungary, Poland, and Slovak. The other two local determinants, changes in local exchange rates and changes in foreign currency reserves are both insignificant with relatively small effect.

As for the global variables, we find that the US stock return is the most salient determinant before the crisis, imposing a negative impact on all sovereigns’ credit spreads. Interestingly, during and after the crisis, this effect turns insignificant for many countries, and some sovereigns report a positive relation. This suggests that the optimistic signals from the global market and agents’ confidence in the global and domestic market play a dominant role during the tranquil period. On the contrary, in the crisis period when the domestic market is volatile, the substitution effect between domestic and global markets would play a more important role, leading to a positive relationship. Treasury yield is
another significant determinant that has a negative effect for almost all countries in all regimes. This effect is particularly strong in most Latin American countries, e.g. Brazil and Chile, but relatively weak in Southeast Asian countries, e.g. Korea, Malaysia and Philippines. The effect of high-yield corporate bond spreads is particularly sizeable and it shows a clear time-varying pattern. Both equity and volatility premium has a significantly positive impact on CDS spreads for all countries in the first regime. However, the financial crisis reshapes this impact to a large extent. The role of both determinants becomes rather unclear during the unstable second regime. In the third regime, the effect of equity premium returns significantly positive for most countries, while that of volatility premium turns negative in general. The effect of bond flows is generally weaker than the that of equity flows, but both show a large degree of heterogeneity across countries.

In general, we observe time-varying feature of the effects of determinants due to financial crisis. Countries do demonstrate heterogeneity, but also possess similarity to some extent. MPA incorporates both heterogeneity and similarity simultaneously, which is not achievable for either the pooled or individual time series estimators that are usually employed in this literature.

8.3 Out-of-sample Forecasting

Next, we employ our MPA to forecast the sovereign CDS spreads, and compare with alternative methods listed in simulation. Given the existence of two structural breaks, we consider forecasting based on three different samples, the full sample ignoring the structural breaks, the subsample after the first break, and that after the second break. It is not guaranteed that post-break subsamples always lead to better forecasts compared to the full sample due to bias-efficiency trade-off.\footnote{There exist various methods to deal with the window-selection issue in time series literature, (see, e.g., Pesaran and Timmermann, 2007; Pesaran et al., 2013, among others). Optimal window selection in panel forecasting is an interesting topic that deserves future research.} The forecasts are constructed using both fixed window and expanding window. To evaluate the forecasting performance, in each case we divide the available time periods into two sub-samples: the first $\tau\%$ of the sample is used to estimate the coefficients and weights, and the remaining are used for forecasting and evaluation. We let the forecasting proportion $1 - \tau\%$ vary among 0.01, 0.05, and 0.1. We evaluate the forecasts using the root mean square forecasting error (RMSFE), which
is averaged across 14 countries. To facilitate comparison, all numbers are normalized with respect to the individual time series forecasts using the full sample.

Table 5 presents the out-of-sample forecasting performance of competing methods based on the fixed window. It indeed shows that accommodating structural breaks does not always lead to more accurate forecasts. With a relatively large and moderate out-of-sample proportion $1 - \tau\% = 0.1$ and $0.05$, using the full sample results in more accurate forecasts than using the post-break samples. With out-of-sample proportion 0.01, the forecasts using the sample after the first break (but ignoring the second break) lead to the most accurate forecast. This suggests that the second break is not significant enough for the forecast purpose, and using the pre-break observations helps to gain more efficiency that offsets the bias. In general, MPA provides the most accurate forecast, regardless of the sample in use and the out-of-sample proportion. If only the post-break sample is used, the signal-to-noise ratio is typically low, and the forecast based on the pooled model is most reliable. In this case, MPA adaptively assigns most weights to the pooled model, and it produces equally good forecasts as the pooled model. This result is in line with our simulation findings when the signal-to-noise ratio is low and the sample size is small. In the case of expanding window, one-step-ahead forecasts are constructed at each time as the window expands. The forecasting performance of competing methods is exactly the same as in the fixed window case, namely that MPA outperforms other methods when the full sample is used, and it performs as well as the pooled model when only subsamples are used. Detailed results are omitted but available upon request from the authors.

9 Implications and Discussions

In this paper we have proposed a novel optimal pooling averaging method to analyse the determinants of sovereign CDS spreads in a potentially heterogeneous cross-country panel and to forecast the future values of the spreads for each sovereign. The forecasting performance of our method generally outperforms the alternatives and is rather robust.

Based on our theoretical and numerical results, we conclude that the performance of
different pooling estimators depends on the situation at hand. There is however a clear pattern in the results. We therefore end this paper with a practical procedure, including rules of thumb based on checkable conditions, to determine which estimator/forecast to use in different situations.

As the first step, estimate individual time series separately, and compute the coefficient estimates and (adjusted) \( R^2 \) 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 \( R^2 \), say \( R^2 < 0.5 \), the pooled model seems to be a safe choice (as suggested by the Monte Carlo study). When large \( R^2 \) values are observed in most regressions and coefficient estimates exhibit heterogeneity, Mallows pooling averaging appears to be a good choice. Sample size also plays an important role when choosing pooling methods. Mallows pooling averaging typically performs well when \( T \) is finite. If the time dimension is particularly large, the C-Lasso approaches can however produce accurate results. Although our simulation studies consider a variety of DGPs, we emphasize that these rules of thumb are concluded based our given setup. There are still several cases that we do not cover, such as dynamic panels and panels with cross-section dependence. Hence, one needs to be cautious when applying these suggestions to aforementioned extensions.

References


Su L, Wang W. 2017. Identifying latent group structures in nonlinear panels. working paper.


Technical Appendices

A.1 Proof of Theorem 1

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

\[
R_A(\tilde{w}) = \inf_{w \in W} R_A(w),
\]

which means that \( \tilde{w} \) is theoretically optimal weight vector.

We rewrite \( C_A(w) \) as

\[
C_A(w) = \|P(w)\hat{\beta} - \hat{\beta} \|^2_A + 2 \text{tr}\{P'(w)AV\} - \|\hat{\beta} - \beta \|^2_A
\]

Letting

\[
a(w) = 2 \text{tr}\{P'(w)AV\} - 2 \left\{ P(w)\hat{\beta} - \beta \right\}' A(\hat{\beta} - \beta),
\]

we have

\[
C_A(w) = L_A(w) + a(w).
\]

It is straightforward to show that for any weight vector \( w \),

\[
E\{a(w)\} = 2 \text{tr}\{P'(w)AV\} - 2E \left[ \left\{ P(w)\hat{\beta} - \beta \right\}' A(\hat{\beta} - \beta) \right] = 0.
\]

It follows from (17) and (A.3) that

\[
L_A(\hat{w}) = C_A(\hat{w}) - a(\hat{w}) \leq C_A(\tilde{w}) - a(\tilde{w}) = L_A(\tilde{w}) + a(\tilde{w}) - a(\hat{w}).
\]

Taking expectations of both sides of above formula, by (A.1) and (A.4) we have

\[
E\{L_A(\hat{w})\} \leq E\{L_A(\tilde{w})\} + E\{a(\tilde{w})\} - E\{a(\hat{w})\}
\]

\[
= \inf_{w \in W} R_A(w) - E\{a(\tilde{w})\}
\]

\[
= \inf_{w \in W} R_A(w) - 2E\text{tr}\{P'(\hat{w})AV\} + E \left[ 2 \left\{ P(\hat{w})\hat{\beta} - \beta \right\}' A(\hat{\beta} - \beta) \right]
\]

\[
\leq \inf_{w \in W} R_A(w) - 2E\text{tr}\{P'(\hat{w})AV\} + E \left\{ c\|P(\hat{w})\hat{\beta} - \beta \|^2_A + c^{-1}\|\hat{\beta} - \beta \|^2_A \right\}
\]
\[
\begin{align*}
\inf_{w \in \mathbb{W}} R_A(w) &= 2 \text{Etr} \left\{ P'(\hat{w})AV \right\} + c E \{ L_A(\hat{w}) \} + c^{-1} E \| \hat{\beta} - \beta \|_A^2 \\
\inf_{w \in \mathbb{W}} R_A(w) &= 2 \text{Etr} \left\{ P'(\hat{w})AV \right\} + c E \{ L_A(\hat{w}) \} + c^{-1} \text{tr}(AV),
\end{align*}
\]

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

\[
E \{ L_A(\hat{w}) \} \leq \frac{1}{1 - c} \inf_{w \in \mathbb{W}} R_A(w) + \frac{1}{c - 1} \left\{ \frac{1}{c} \text{tr}(AV) - 2 \text{Etr} \left\{ P'(\hat{w})AV \right\} \right\}.
\]

(A.7)

\section*{A.2 Proof of Corollary 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

\[
B_m = (X'X)^{-1/2} R_m' \left( R_m (X'X)^{-1} R_m' \right)^{-1} R_m (X'X)^{-1/2},
\]

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

\[
I_2(B_m) \leq 1, \quad \mathcal{I}_2(B(w)) \leq 1, \quad \text{rank}(B_m) \leq Nk.
\]

(A.8)

Then, we can rewrite the risk bound of (21) in terms of \( B(\hat{w}) \) as

\[
E \{ L_A(\hat{w}) \} \\
\leq \frac{1}{1 - c} \inf_{w \in \mathbb{W}} R_A(w) + \frac{1}{1 - c} \left\{ \frac{1}{c} \text{tr}(AV) - 2 \text{Etr} \left\{ \left[ I_{Nk} - (X'X)^{1/2} B(\hat{w}) (X'X)^{-1/2} \right] AV \right\} \right\}
\]

\[
= \frac{1}{1 - c} \inf_{w \in \mathbb{W}} R_A(w) + \frac{1 - 2c}{c(1 - c)} \text{tr}(AV) + \frac{2}{c - 1} \text{Etr} \left\{ B(\hat{w}) (X'X)^{-1/2} AV (X'X)^{1/2} \right\}.
\]

(A.9)

We first consider the case of \( A = I_{Nk} \). Using (A.8), the second term of (A.9) satisfies

\[
\text{tr}(AV) = \text{tr}(V) = T^{-1} \sum_{i=1}^{N} \text{tr} \left( Q_i^{-1} X_i' \Omega_i X_i / TQ_i^{-1} \right)
\]

\[
\leq T^{-1} \Omega \sum_{i=1}^{N} \text{tr}(Q_i^{-1})
\]

\[
\leq T^{-1} \Omega \sum_{i=1}^{N} \text{rank}(X_i) \mathcal{I}_2(Q_i^{-1})
\]

\[
\leq T^{-1} Nk \Omega c_i^{-1}
\]

(A.10)

and the third term of (A.9) satisfies

\[
\text{Etr} \left\{ B(\hat{w}) (X'X)^{-1/2} AV (X'X)^{1/2} \right\}
\]
A.3 Proof of Theorem 2

Plugging the inequalities (A.10) and (A.11) into (21), we can obtain the risk bound for \( A = I_{Nk} \) as in (22). Similarly, for the case of \( A = X'X \), we have

\[
\text{tr}(AV) = \text{tr}(X'XV) \leq Nk\Omega, \quad (A.12)
\]

and

\[
\text{Etr} \left\{ B(\hat{w})(X'X)^{-1/2}AV(X'X)^{1/2} \right\} = \text{Etr} \left\{ B(\hat{w})(X'X)^{1/2}V(X'X)^{1/2} \right\} \\
\leq 2Nk\Omega \max_{i \in \{1, \ldots, N\}} I_2(Q_i^{-1}) \leq 2Nk\Omega. \quad (A.13)
\]

Plugging (A.12) and (A.13) into (A.9) leads to the risk bound given in (23).

\[
= \text{Etr} \left\{ B(\hat{w})(X'X)^{-1/2}V(X'X)^{1/2} \right\} \\
= 2^{-1}\text{Etr} \left\{ B(\hat{w})(X'X)^{-1/2}V(X'X)^{1/2} + (X'X)^{1/2}V(X'X)^{-1/2}B(\hat{w}) \right\} \\
\leq 2^{-1}\text{E} \left[ \text{rank} \left\{ B(\hat{w})(X'X)^{-1/2}V(X'X)^{1/2} + (X'X)^{1/2}V(X'X)^{-1/2}B(\hat{w}) \right\} \times I_2 \left\{ B(\hat{w})(X'X)^{-1/2}V(X'X)^{1/2} + (X'X)^{1/2}V(X'X)^{-1/2}B(\hat{w}) \right\} \right] \\
\leq 2\text{E} \left[ \text{rank} \left\{ B(\hat{w})(X'X)^{-1/2}V(X'X)^{1/2} \right\} \right] I_2 \left\{ B(\hat{w})(X'X)^{-1/2}V(X'X)^{1/2} \right\} \\
\leq 2Nk\Omega \max_{i \in \{1, \ldots, N\}} I_2(Q_i^{-1}) \leq 2^{-1}Nk\Omega \Omega_\ast^{-1}. \quad (A.11)
\]

Observe that

\[
R_A(w) = E \{ L_A(w) \} = E \| \hat{\beta}(w) - \beta \|^2_A = E \| P(w)\hat{\beta} - \beta \|^2_A \\
= \| P(w)\beta - \beta \|^2_A + \text{tr} \left\{ P(w)AP'(w)V \right\} \\
= \| P(w)\hat{\beta} - \beta - P(w)(\hat{\beta} - \beta) \|^2_A + \text{tr} \left\{ P(w)AP'(w)V \right\} \\
= L_A(w) + \| P(w)(\hat{\beta} - \beta) \|^2_A + 2\beta'AP(w)(\hat{\beta} - \beta) \\
- 2\hat{\beta}'P(w)AP'(w)\hat{\beta} + \text{tr} \left\{ P(w)AP'(w)V \right\} \\
\equiv L_A(w) + \Xi(w). \quad (A.14)
\]
From (A.3), (A.14), and the proof of Theorem 1 of Hansen (2007), to prove (24), we need only verify that
\[
\sup_{w \in W} |a(w)| R_A(w) = o_p(1),
\] (A.15)
where \(a(w)\) is defined in (A.2), and
\[
\sup_{w \in W} |\Xi(w)| R_A(w) = o_p(1).
\] (A.16)

From Conditions C.1 and C.2, we know that no matter \(\hat{V} = \hat{V}_{\text{homo}}, \hat{V} = \hat{V}_{\text{bh}}\) or \(\hat{V} = \hat{V}_{\text{ch}},\)
\[
\text{tr}(\hat{V}) = O_p(NT^{-1}), \quad \mathcal{I}_2(\hat{V}) = O_p(T^{-1}).
\] (A.17)

Using (A.8), C.2 and the proving steps of (A.10), we have
\[
\sup_{w \in W} \left| \text{tr} \left\{ P'(w) A \hat{V} \right\} \right| \leq \sum_{m=1}^{M} \left| \text{tr}(P'_m A \hat{V}) \right|
\]
\[
= \sum_{m=1}^{M} \left[ \left| \text{tr}(A \hat{V}) \right| + \left| \text{tr} \left\{ (X'X)^{1/2} B_m (X'X)^{-1/2} A \hat{V} \right\} \right| \right]
\]
\[
\leq \sum_{m=1}^{M} \left\{ \mathcal{I}_2(A) \text{tr}(\hat{V}) + 2 \mathcal{I}_2(A) \mathcal{I}_2((X'X)^{1/2}) \mathcal{I}_2((X'X)^{-1/2}) \text{rank}(B_m) \mathcal{I}_2(\hat{V}) \right\}
\]
\[
= \mathcal{I}_2(A) O(MNT^{-1}).
\] (A.18)

From Conditions C.1 and C.2, we have \(\hat{\beta}_i - \beta_i = O_p(T^{-1/2})\) uniformly for \(i = 1, \ldots, N,\) which implies
\[
\sup_{w \in W} \left| \left\{ P(w) \hat{\beta} - \beta \right\}' A(\hat{\beta} - \beta) \right| \leq \|\hat{\beta} - \beta\| \mathcal{I}_2(A) \sup_{w \in W} \| P(w) \hat{\beta} - \beta \|
\]
\[
\leq \|\hat{\beta} - \beta\| \mathcal{I}_2(A) \sum_{m=1}^{M} \| P_m \hat{\beta} - \beta \|
\]
\[
= O_p(MNT^{-1/2}) \mathcal{I}_2(A).
\] (A.19)

Now, from (A.18), (A.19) and C.3, we can obtain (A.15). Similarly, we can also obtain (A.16). This completes the proof.
Figure 1: Time-varying pattern and heterogeneity of slope coefficients
Table 1: Risk comparison: Independent regressors with $R^2 = 0.9$

<table>
<thead>
<tr>
<th>DGP</th>
<th>MPA</th>
<th>C-Lasso</th>
<th>SAIC</th>
<th>SBIC</th>
<th>AIC</th>
<th>BIC</th>
<th>Pool</th>
<th>SHK</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.130</td>
<td>0.452</td>
<td>0.477</td>
<td>0.142</td>
<td>0.567</td>
<td>0.154</td>
<td>0.099</td>
<td>0.739</td>
</tr>
<tr>
<td>$N = 10$</td>
<td>2</td>
<td><strong>0.343</strong></td>
<td>0.607</td>
<td>0.465</td>
<td>0.435</td>
<td>0.505</td>
<td>0.491</td>
<td>4.323</td>
</tr>
<tr>
<td>$T = 20$</td>
<td>3</td>
<td><strong>0.537</strong></td>
<td>0.834</td>
<td>0.723</td>
<td>0.777</td>
<td>0.799</td>
<td>0.825</td>
<td>2.629</td>
</tr>
<tr>
<td>4</td>
<td><strong>0.953</strong></td>
<td>1.972</td>
<td>1.232</td>
<td>1.423</td>
<td>1.296</td>
<td>1.598</td>
<td>14.92</td>
<td>0.984</td>
</tr>
<tr>
<td>1</td>
<td>0.172</td>
<td>0.312</td>
<td>0.364</td>
<td>0.265</td>
<td>0.405</td>
<td>0.302</td>
<td>0.099</td>
<td>0.739</td>
</tr>
<tr>
<td>$N = 30$</td>
<td>2</td>
<td><strong>0.160</strong></td>
<td>0.192</td>
<td>0.260</td>
<td>0.214</td>
<td>0.270</td>
<td>0.225</td>
<td>4.192</td>
</tr>
<tr>
<td>$T = 20$</td>
<td>3</td>
<td><strong>0.371</strong></td>
<td>0.640</td>
<td>0.496</td>
<td>0.520</td>
<td>0.509</td>
<td>0.551</td>
<td>2.299</td>
</tr>
<tr>
<td>4</td>
<td><strong>0.406</strong></td>
<td>0.662</td>
<td>0.634</td>
<td>0.661</td>
<td>0.650</td>
<td>0.662</td>
<td>1.766</td>
<td>0.892</td>
</tr>
<tr>
<td>1</td>
<td>0.135</td>
<td>0.466</td>
<td>0.480</td>
<td>0.128</td>
<td>0.578</td>
<td>0.133</td>
<td>0.103</td>
<td>0.875</td>
</tr>
<tr>
<td>$N = 40$</td>
<td>2</td>
<td><strong>0.272</strong></td>
<td>0.303</td>
<td>0.466</td>
<td>0.279</td>
<td>0.511</td>
<td>0.284</td>
<td>8.790</td>
</tr>
<tr>
<td>$T = 40$</td>
<td>3</td>
<td><strong>0.643</strong></td>
<td>1.249</td>
<td>0.894</td>
<td>1.060</td>
<td>0.989</td>
<td>1.163</td>
<td>5.074</td>
</tr>
<tr>
<td>4</td>
<td>1.250</td>
<td>2.707</td>
<td>1.728</td>
<td>1.791</td>
<td>1.755</td>
<td>1.953</td>
<td>29.68</td>
<td><strong>0.997</strong></td>
</tr>
<tr>
<td>1</td>
<td>0.185</td>
<td>0.324</td>
<td>0.373</td>
<td>0.264</td>
<td>0.418</td>
<td>0.306</td>
<td>0.034</td>
<td>0.857</td>
</tr>
<tr>
<td>$N = 30$</td>
<td>2</td>
<td><strong>0.100</strong></td>
<td>0.082</td>
<td>0.271</td>
<td>0.156</td>
<td>0.282</td>
<td>0.172</td>
<td>8.493</td>
</tr>
<tr>
<td>$T = 40$</td>
<td>3</td>
<td><strong>0.493</strong></td>
<td>1.030</td>
<td>0.653</td>
<td>0.747</td>
<td>0.693</td>
<td>0.768</td>
<td>4.666</td>
</tr>
<tr>
<td>4</td>
<td><strong>0.637</strong></td>
<td>1.056</td>
<td>0.874</td>
<td>0.985</td>
<td>0.909</td>
<td>1.009</td>
<td>3.450</td>
<td>0.968</td>
</tr>
</tbody>
</table>

Notes:
1. Forecasts constructed using: MPA: Mallows pooling averaging estimator; C-Lasso: C-Lasso estimator with the number of groups determined by BIC; SAIC/SBIC: pooling averaging estimator based on relative values of AIC/BIC; AIC/BIC: estimator selected based on minimum value information criterion; Pool: pooled estimator; SHK: shrinkage estimator (7).
2. All numbers are divided by the risk of the individual time series forecast.
Table 2: Risk comparison: Autoregressive regressors with $R^2 = 0.9$

<table>
<thead>
<tr>
<th></th>
<th>DGP</th>
<th>MPA</th>
<th>C-Lasso</th>
<th>SAIC</th>
<th>SBIC</th>
<th>AIC</th>
<th>BIC</th>
<th>Pool</th>
<th>SHK</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N = 10$</td>
<td>1</td>
<td>0.117</td>
<td>0.424</td>
<td>0.467</td>
<td>0.138</td>
<td>0.567</td>
<td>0.149</td>
<td><strong>0.099</strong></td>
<td>0.743</td>
</tr>
<tr>
<td>$T = 20$</td>
<td>2</td>
<td><strong>0.358</strong></td>
<td>0.675</td>
<td>0.437</td>
<td>0.469</td>
<td>0.473</td>
<td>0.517</td>
<td>4.345</td>
<td>0.955</td>
</tr>
<tr>
<td>$N = 30$</td>
<td>3</td>
<td><strong>0.509</strong></td>
<td>0.826</td>
<td>0.669</td>
<td>0.774</td>
<td>0.750</td>
<td>0.813</td>
<td>2.523</td>
<td>0.928</td>
</tr>
<tr>
<td>$T = 40$</td>
<td>4</td>
<td><strong>0.897</strong></td>
<td>1.795</td>
<td>1.134</td>
<td>1.317</td>
<td>1.200</td>
<td>1.461</td>
<td>14.32</td>
<td>0.985</td>
</tr>
<tr>
<td>$N = 10$</td>
<td>1</td>
<td>0.133</td>
<td>0.276</td>
<td>0.337</td>
<td>0.202</td>
<td>0.378</td>
<td>0.233</td>
<td><strong>0.034</strong></td>
<td>0.698</td>
</tr>
<tr>
<td>$T = 20$</td>
<td>2</td>
<td><strong>0.172</strong></td>
<td>0.225</td>
<td>0.232</td>
<td>0.215</td>
<td>0.241</td>
<td>0.222</td>
<td>4.174</td>
<td>0.943</td>
</tr>
<tr>
<td>$N = 30$</td>
<td>3</td>
<td><strong>0.371</strong></td>
<td>0.660</td>
<td>0.481</td>
<td>0.523</td>
<td>0.493</td>
<td>0.555</td>
<td>2.295</td>
<td>0.909</td>
</tr>
<tr>
<td>$T = 40$</td>
<td>4</td>
<td><strong>0.406</strong></td>
<td>0.621</td>
<td>0.588</td>
<td>0.619</td>
<td>0.606</td>
<td>0.620</td>
<td>1.808</td>
<td>0.890</td>
</tr>
<tr>
<td>$N = 10$</td>
<td>1</td>
<td>0.124</td>
<td>0.449</td>
<td>0.469</td>
<td>0.120</td>
<td>0.561</td>
<td>0.123</td>
<td><strong>0.103</strong></td>
<td>0.875</td>
</tr>
<tr>
<td>$T = 40$</td>
<td>2</td>
<td><strong>0.264</strong></td>
<td>0.351</td>
<td>0.457</td>
<td>0.288</td>
<td>0.492</td>
<td>0.297</td>
<td>8.776</td>
<td>0.987</td>
</tr>
<tr>
<td>$N = 30$</td>
<td>3</td>
<td><strong>0.657</strong></td>
<td>1.295</td>
<td>0.880</td>
<td>1.062</td>
<td>0.967</td>
<td>1.166</td>
<td>5.314</td>
<td>0.981</td>
</tr>
<tr>
<td>$T = 40$</td>
<td>4</td>
<td>1.182</td>
<td>2.546</td>
<td>1.613</td>
<td>1.729</td>
<td>1.646</td>
<td>1.873</td>
<td>28.06</td>
<td><strong>0.996</strong></td>
</tr>
<tr>
<td>$N = 10$</td>
<td>1</td>
<td>0.154</td>
<td>0.300</td>
<td>0.353</td>
<td>0.209</td>
<td>0.393</td>
<td>0.249</td>
<td><strong>0.032</strong></td>
<td>0.853</td>
</tr>
<tr>
<td>$T = 40$</td>
<td>2</td>
<td><strong>0.095</strong></td>
<td>0.258</td>
<td>0.153</td>
<td>0.270</td>
<td>0.165</td>
<td>0.165</td>
<td>8.165</td>
<td>0.984</td>
</tr>
<tr>
<td>$N = 30$</td>
<td>3</td>
<td><strong>0.501</strong></td>
<td>1.014</td>
<td>0.658</td>
<td>0.745</td>
<td>0.696</td>
<td>0.763</td>
<td>4.676</td>
<td>0.974</td>
</tr>
<tr>
<td>$T = 40$</td>
<td>4</td>
<td><strong>0.652</strong></td>
<td>1.064</td>
<td>0.872</td>
<td>0.991</td>
<td>0.907</td>
<td>1.014</td>
<td>3.516</td>
<td>0.968</td>
</tr>
</tbody>
</table>

Notes: See footnote of Table 1.
<table>
<thead>
<tr>
<th>DGP</th>
<th>MPA</th>
<th>C-Lasso</th>
<th>SAIC</th>
<th>SBIC</th>
<th>AIC</th>
<th>BIC</th>
<th>Pool</th>
<th>SHK</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.133</td>
<td>0.454</td>
<td>0.494</td>
<td>0.146</td>
<td>0.588</td>
<td>0.163</td>
<td><strong>0.104</strong></td>
<td>0.742</td>
</tr>
<tr>
<td>2</td>
<td><strong>0.321</strong></td>
<td>0.479</td>
<td>0.471</td>
<td>0.406</td>
<td>0.509</td>
<td>0.461</td>
<td>0.389</td>
<td>0.801</td>
</tr>
<tr>
<td>3</td>
<td>0.237</td>
<td>0.259</td>
<td>0.298</td>
<td>0.273</td>
<td>0.357</td>
<td>0.280</td>
<td>0.273</td>
<td>0.775</td>
</tr>
<tr>
<td>4</td>
<td><strong>0.500</strong></td>
<td>0.619</td>
<td>0.769</td>
<td>0.716</td>
<td>0.831</td>
<td>0.801</td>
<td>1.071</td>
<td>0.874</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.129</td>
<td>0.459</td>
<td>0.466</td>
<td>0.120</td>
<td>0.553</td>
<td>0.122</td>
<td><strong>0.101</strong></td>
<td>0.872</td>
</tr>
<tr>
<td>2</td>
<td><strong>0.435</strong></td>
<td>0.627</td>
<td>0.610</td>
<td>0.618</td>
<td>0.639</td>
<td>0.683</td>
<td>0.710</td>
<td>0.922</td>
</tr>
<tr>
<td>3</td>
<td>0.315</td>
<td>0.349</td>
<td>0.413</td>
<td>0.416</td>
<td>0.475</td>
<td>0.443</td>
<td>0.435</td>
<td>0.908</td>
</tr>
<tr>
<td>4</td>
<td><strong>0.567</strong></td>
<td>0.785</td>
<td>0.843</td>
<td>0.809</td>
<td>0.931</td>
<td>0.892</td>
<td>2.125</td>
<td>0.959</td>
</tr>
</tbody>
</table>

|     |     |         |      |      |      |      |      |     |
| 1   | 0.131 | 0.440 | 0.506 | 0.138 | 0.608 | 0.146 | **0.104** | 0.739 |
| 2   | 0.154 | 0.217 | 0.221 | 0.161 | 0.242 | 0.163 | **0.153** | 0.751 |
| 3   | 0.133 | 0.231 | 0.217 | 0.148 | 0.242 | 0.152 | **0.131** | 0.737 |
| 4   | 0.266 | 0.585 | 0.554 | 0.422 | 0.588 | 0.477 | 0.278 | 0.774 |
|     |     |         |      |      |      |      |      |     |
| 1   | 0.130 | 0.459 | 0.487 | 0.119 | 0.580 | 0.123 | **0.103** | 0.873 |
| 2   | **0.206** | 0.305 | 0.298 | 0.217 | 0.348 | 0.218 | 0.211 | 0.882 |
| 3   | 0.165 | 0.364 | 0.322 | 0.187 | 0.352 | 0.192 | **0.164** | 0.875 |
| 4   | **0.376** | 0.614 | 0.596 | 0.541 | 0.626 | 0.604 | 0.468 | 0.903 |

Notes: See footnote of Table 1.
Table 4: Effects of CDS spreads determinants for selected countries: Mallows pooling averaging estimates

<table>
<thead>
<tr>
<th></th>
<th>Brazil</th>
<th></th>
<th>China</th>
<th></th>
<th>Hungary</th>
<th></th>
<th>Japan</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reg. 1</td>
<td>Reg. 2</td>
<td>Reg. 3</td>
<td>Reg. 1</td>
<td>Reg. 2</td>
<td>Reg. 3</td>
<td>Reg. 1</td>
<td>Reg. 2</td>
</tr>
<tr>
<td><code>lstock</code></td>
<td>-0.047</td>
<td>-1.772</td>
<td>-2.43</td>
<td>0.059</td>
<td>-4.202</td>
<td>-0.161</td>
<td>0.059</td>
<td>-0.391</td>
</tr>
<tr>
<td></td>
<td>(0.062)</td>
<td>(0.847)</td>
<td>(0.039)</td>
<td>(0.090)</td>
<td>(1.887)</td>
<td>(0.044)</td>
<td>(0.108)</td>
<td>(1.560)</td>
</tr>
<tr>
<td><code>fxrates</code></td>
<td>0.044</td>
<td>-2.213</td>
<td>-0.119</td>
<td>0.156</td>
<td>30.182</td>
<td>-0.018</td>
<td>0.156</td>
<td>-0.383</td>
</tr>
<tr>
<td></td>
<td>(0.058)</td>
<td>(2.841)</td>
<td>(0.034)</td>
<td>(0.086)</td>
<td>(13.591)</td>
<td>(0.038)</td>
<td>(0.102)</td>
<td>(10.460)</td>
</tr>
<tr>
<td><code>fxxes</code></td>
<td>-0.083</td>
<td>-0.942</td>
<td>-0.183</td>
<td>0.030</td>
<td>7.693</td>
<td>0.108</td>
<td>0.030</td>
<td>0.185</td>
</tr>
<tr>
<td></td>
<td>(0.060)</td>
<td>(0.951)</td>
<td>(0.059)</td>
<td>(0.052)</td>
<td>(3.509)</td>
<td>(0.082)</td>
<td>(0.050)</td>
<td>(2.807)</td>
</tr>
<tr>
<td><code>gstock</code></td>
<td>-0.212</td>
<td>0.582</td>
<td>0.277</td>
<td>-0.292</td>
<td>3.001</td>
<td>0.152</td>
<td>-0.292</td>
<td>0.630</td>
</tr>
<tr>
<td></td>
<td>(0.044)</td>
<td>(0.285)</td>
<td>(0.046)</td>
<td>(0.127)</td>
<td>(1.415)</td>
<td>(0.051)</td>
<td>(0.132)</td>
<td>(1.291)</td>
</tr>
<tr>
<td><code>trsy</code></td>
<td>-0.078</td>
<td>-0.629</td>
<td>-0.158</td>
<td>-0.042</td>
<td>-0.493</td>
<td>-0.143</td>
<td>-0.042</td>
<td>-0.343</td>
</tr>
<tr>
<td></td>
<td>(0.030)</td>
<td>(0.229)</td>
<td>(0.023)</td>
<td>(0.036)</td>
<td>(0.314)</td>
<td>(0.020)</td>
<td>(0.039)</td>
<td>(0.342)</td>
</tr>
<tr>
<td><code>by</code></td>
<td>-0.178</td>
<td>0.609</td>
<td>-0.062</td>
<td>-0.183</td>
<td>-2.297</td>
<td>0.088</td>
<td>-0.183</td>
<td>0.299</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td>(0.385)</td>
<td>(0.042)</td>
<td>(0.042)</td>
<td>(1.327)</td>
<td>(0.040)</td>
<td>(0.040)</td>
<td>(1.114)</td>
</tr>
<tr>
<td><code>eqp</code></td>
<td>0.138</td>
<td>0.739</td>
<td>0.060</td>
<td>0.121</td>
<td>-0.100</td>
<td>-0.041</td>
<td>0.121</td>
<td>0.658</td>
</tr>
<tr>
<td></td>
<td>(0.054)</td>
<td>(0.163)</td>
<td>(0.038)</td>
<td>(0.151)</td>
<td>(0.424)</td>
<td>(0.041)</td>
<td>(0.152)</td>
<td>(0.392)</td>
</tr>
<tr>
<td><code>volp</code></td>
<td>0.121</td>
<td>1.713</td>
<td>-0.071</td>
<td>0.118</td>
<td>43.058</td>
<td>-0.147</td>
<td>0.118</td>
<td>7.639</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td>(4.642)</td>
<td>(0.026)</td>
<td>(0.027)</td>
<td>(16.678)</td>
<td>(0.025)</td>
<td>(0.024)</td>
<td>(13.175)</td>
</tr>
<tr>
<td><code>ef</code></td>
<td>-0.032</td>
<td>0.625</td>
<td>-0.119</td>
<td>-0.137</td>
<td>-5.845</td>
<td>-0.247</td>
<td>-0.137</td>
<td>1.018</td>
</tr>
<tr>
<td></td>
<td>(0.061)</td>
<td>(0.615)</td>
<td>(0.020)</td>
<td>(0.042)</td>
<td>(3.188)</td>
<td>(0.020)</td>
<td>(0.041)</td>
<td>(2.592)</td>
</tr>
<tr>
<td><code>bf</code></td>
<td>0.110</td>
<td>0.362</td>
<td>-0.058</td>
<td>0.092</td>
<td>-3.561</td>
<td>-0.095</td>
<td>0.092</td>
<td>0.378</td>
</tr>
<tr>
<td></td>
<td>(0.057)</td>
<td>(0.345)</td>
<td>(0.017)</td>
<td>(0.063)</td>
<td>(1.758)</td>
<td>(0.017)</td>
<td>(0.077)</td>
<td>(1.368)</td>
</tr>
</tbody>
</table>

Notes: Bootstrap standard errors given in the parentheses.
Table 4: Effects of CDS spreads determinants for selected countries: Mallows pooling averaging estimates

<table>
<thead>
<tr>
<th></th>
<th>Malaysia</th>
<th></th>
<th></th>
<th>Poland</th>
<th></th>
<th></th>
<th>Slovak</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reg. 1</td>
<td>Reg. 2</td>
<td>Reg. 3</td>
<td>Reg. 1</td>
<td>Reg. 2</td>
<td>Reg. 3</td>
<td>Reg. 1</td>
<td>Reg. 2</td>
<td>Reg. 3</td>
</tr>
<tr>
<td>lstock</td>
<td>0.145</td>
<td>−0.391</td>
<td>−0.300</td>
<td>−0.028</td>
<td>−0.391</td>
<td>−0.192</td>
<td>−0.047</td>
<td>−0.391</td>
<td>0.242</td>
</tr>
<tr>
<td></td>
<td>(0.121)</td>
<td>(1.165)</td>
<td>(0.037)</td>
<td>(0.100)</td>
<td>(1.275)</td>
<td>(0.038)</td>
<td>(0.064)</td>
<td>(1.556)</td>
<td>(0.039)</td>
</tr>
<tr>
<td>fxrates</td>
<td>0.210</td>
<td>−0.383</td>
<td>−0.007</td>
<td>0.067</td>
<td>−0.383</td>
<td>−0.090</td>
<td>0.044</td>
<td>−0.383</td>
<td>−0.016</td>
</tr>
<tr>
<td></td>
<td>(0.104)</td>
<td>(3.275)</td>
<td>(0.044)</td>
<td>(0.097)</td>
<td>(3.308)</td>
<td>(0.041)</td>
<td>(0.096)</td>
<td>(1.022)</td>
<td>(0.038)</td>
</tr>
<tr>
<td>fxres</td>
<td>0.039</td>
<td>0.185</td>
<td>0.091</td>
<td>−0.036</td>
<td>0.185</td>
<td>0.202</td>
<td>−0.083</td>
<td>0.185</td>
<td>−0.068</td>
</tr>
<tr>
<td></td>
<td>(0.107)</td>
<td>(1.967)</td>
<td>(0.097)</td>
<td>(0.356)</td>
<td>(2.320)</td>
<td>(0.091)</td>
<td>(0.475)</td>
<td>(2.059)</td>
<td>(0.074)</td>
</tr>
<tr>
<td>gstock</td>
<td>−0.307</td>
<td>0.630</td>
<td>0.462</td>
<td>−0.299</td>
<td>0.630</td>
<td>0.245</td>
<td>−0.212</td>
<td>0.630</td>
<td>0.316</td>
</tr>
<tr>
<td></td>
<td>(0.109)</td>
<td>(1.071)</td>
<td>(0.051)</td>
<td>(0.075)</td>
<td>(0.833)</td>
<td>(0.048)</td>
<td>(0.080)</td>
<td>(0.654)</td>
<td>(0.054)</td>
</tr>
<tr>
<td>trsy</td>
<td>−0.033</td>
<td>−0.343</td>
<td>−0.124</td>
<td>−0.056</td>
<td>−0.343</td>
<td>−0.182</td>
<td>−0.078</td>
<td>−0.343</td>
<td>−0.342</td>
</tr>
<tr>
<td></td>
<td>(0.033)</td>
<td>(0.405)</td>
<td>(0.019)</td>
<td>(0.040)</td>
<td>(0.609)</td>
<td>(0.021)</td>
<td>(0.063)</td>
<td>(1.224)</td>
<td>(0.024)</td>
</tr>
<tr>
<td>hy</td>
<td>−0.193</td>
<td>0.299</td>
<td>0.265</td>
<td>−0.170</td>
<td>0.299</td>
<td>0.144</td>
<td>−0.178</td>
<td>0.299</td>
<td>0.273</td>
</tr>
<tr>
<td></td>
<td>(0.040)</td>
<td>(1.055)</td>
<td>(0.036)</td>
<td>(0.034)</td>
<td>(0.988)</td>
<td>(0.037)</td>
<td>(0.032)</td>
<td>(0.849)</td>
<td>(0.045)</td>
</tr>
<tr>
<td>eqp</td>
<td>0.111</td>
<td>0.658</td>
<td>0.240</td>
<td>0.191</td>
<td>0.658</td>
<td>0.168</td>
<td>0.138</td>
<td>0.658</td>
<td>0.269</td>
</tr>
<tr>
<td></td>
<td>(0.127)</td>
<td>(0.349)</td>
<td>(0.046)</td>
<td>(0.102)</td>
<td>(0.338)</td>
<td>(0.045)</td>
<td>(0.114)</td>
<td>(0.435)</td>
<td>(0.045)</td>
</tr>
<tr>
<td>volf</td>
<td>0.121</td>
<td>7.639</td>
<td>−0.165</td>
<td>0.119</td>
<td>7.639</td>
<td>−0.177</td>
<td>0.121</td>
<td>7.639</td>
<td>−0.160</td>
</tr>
<tr>
<td></td>
<td>(0.031)</td>
<td>(5.933)</td>
<td>(0.020)</td>
<td>(0.030)</td>
<td>(5.361)</td>
<td>(0.019)</td>
<td>(0.025)</td>
<td>(3.588)</td>
<td>(0.022)</td>
</tr>
<tr>
<td>ef</td>
<td>−0.146</td>
<td>1.018</td>
<td>−0.229</td>
<td>−0.082</td>
<td>1.018</td>
<td>−0.147</td>
<td>−0.032</td>
<td>1.018</td>
<td>−0.203</td>
</tr>
<tr>
<td></td>
<td>(0.035)</td>
<td>(1.583)</td>
<td>(0.020)</td>
<td>(0.037)</td>
<td>(1.343)</td>
<td>(0.019)</td>
<td>(0.051)</td>
<td>(0.714)</td>
<td>(0.019)</td>
</tr>
<tr>
<td>bf</td>
<td>0.157</td>
<td>0.378</td>
<td>−0.153</td>
<td>0.064</td>
<td>0.378</td>
<td>−0.128</td>
<td>0.110</td>
<td>0.378</td>
<td>−0.243</td>
</tr>
<tr>
<td></td>
<td>(0.085)</td>
<td>(0.562)</td>
<td>(0.017)</td>
<td>(0.071)</td>
<td>(0.471)</td>
<td>(0.018)</td>
<td>(0.061)</td>
<td>(0.173)</td>
<td>(0.019)</td>
</tr>
</tbody>
</table>

Notes: Bootstrap standard errors given in the parentheses.
Table 5: Out-of-sample forecasting comparison with fixed window

<table>
<thead>
<tr>
<th>Method</th>
<th>Full sample</th>
<th>Post-first-break sample</th>
<th>Post-second-break sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 − τ% 0.1</td>
<td>0.05 0.01</td>
<td>0.1 0.05 0.01</td>
</tr>
<tr>
<td>MPA</td>
<td>0.8680 0.890 1.020</td>
<td>1.469 0.923 0.476</td>
<td>1.578 1.248 0.571</td>
</tr>
<tr>
<td>C-Lasso</td>
<td>0.8750 0.892 1.024</td>
<td>1.469 0.924 0.476</td>
<td>1.580 1.248 0.571</td>
</tr>
<tr>
<td>SAIC</td>
<td>0.8700 0.891 1.022</td>
<td>1.470 0.924 0.477</td>
<td>1.582 1.264 0.730</td>
</tr>
<tr>
<td>SBIC</td>
<td>0.8750 0.892 1.024</td>
<td>1.469 0.924 0.476</td>
<td>1.580 1.248 0.571</td>
</tr>
<tr>
<td>AIC</td>
<td>0.8750 0.892 1.024</td>
<td>1.469 0.924 0.476</td>
<td>1.580 1.271 0.786</td>
</tr>
<tr>
<td>BIC</td>
<td>0.8750 0.892 1.024</td>
<td>1.469 0.924 0.476</td>
<td>1.580 1.248 0.571</td>
</tr>
<tr>
<td>Pool</td>
<td>0.8750 0.892 1.024</td>
<td>1.469 0.924 0.476</td>
<td>1.580 1.248 0.571</td>
</tr>
<tr>
<td>SHK</td>
<td>0.9800 0.982 0.992</td>
<td>1.552 0.967 0.508</td>
<td>1.636 1.273 0.687</td>
</tr>
<tr>
<td>Indiv</td>
<td>1.0000 1.000 1.000</td>
<td>1.622 1.012 0.562</td>
<td>1.715 1.324 0.763</td>
</tr>
</tbody>
</table>

Notes:
1. τ denotes percentage of sample used for parameter estimation. Abbreviations explained in footnote 1 of Table 1.
2. RMSFE are divided by the RMSFE of the individual time series forecast using the full sample.
3. Numbers in bold are the unique minimum values in the corresponding column.
Supplementary file to
“To pool or not to pool: What is a good strategy for parameter estimation and forecasting in panel regressions?”

This supplementary file contains further theoretical properties of pooling averaging approaches and additional simulation studies that are not reported in the paper.

S.1 Unbiasedness of Mallows criterion

This section provides the theorem of the unbiasedness of Mallows criterion as an estimator or the squared risk.

**Theorem S.1.** Under model (1), the Mallows criterion defined in Equation (16) is an unbiased estimator of the squared risk $R_A(w)$.

**Proof.** From (6), it is straightforward to show that

\[
R_A(w) = \mathbb{E}\{L_A(w)\} = \mathbb{E}\|\hat{\beta}(w) - \beta\|^2_A = \mathbb{E}\|P(w)\hat{\beta} - \beta\|^2_A
\]

and

\[
\mathbb{E}\{C_A(w)\} = \mathbb{E}\|P(w)\hat{\beta} - \beta\|^2_A + 2\text{tr}\{P'(w)A\beta\} - \text{tr}(AV)
\]

\[
= \mathbb{E}\|P(w)\hat{\beta}\|^2_A + \mathbb{E}\|\beta\|^2_A - 2\text{tr}(AV)
\]

\[
-2\text{tr}\left\{\hat{\beta}'P'(w)A\beta\right\} + 2\text{tr}\{P'(w)AV\}
\]

\[
= \mathbb{E}\|P(w)\hat{\beta}\|^2_A + \|\beta\|^2_A - 2\beta'P'(w)A\beta.
\]

(S.1)

So $C_A(w)$ is an unbiased estimator of $R_A(w)$.

\[\square\]

S.2 Equivalence of MPA and Stein-rule estimators

This section provides the details on the relation between the MPA and the Stein-rule shrinkage estimator. Our pooling averaging estimator includes the shrinkage estimator of
Maddala et al. (1997) as a special case (see Section 5.1). The shrinkage estimator is defined as

$$\hat{\beta}_{\text{shrinkage}} = \left(1 - \frac{\nu}{F}\right) \hat{\beta} + \frac{\nu}{F} \hat{\beta}_{\text{pool}},$$  \hspace{1cm} (S.2)

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 \(\tilde{R}\) as the restriction matrix associated with \(H_0\) and

$$\tilde{\sigma}^2 = (Y - X\hat{\beta})' (Y - X\hat{\beta})/(NT - N),$$  \hspace{1cm} (S.3)

the rank of \(\tilde{R}\) is \(k(N - 1)\) and the \(F\) statistic is

$$F = (\tilde{R}\hat{\beta})' (\tilde{R}(X'X)^{-1}\tilde{R}')^{-1}(\tilde{R}\hat{\beta})/[(N - 1)k\tilde{\sigma}^2].$$ \hspace{1cm} (S.4)

The shrinkage estimator can be regarded as the pooling average of only the pooled and individual estimators.

Now 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, \(\hat{\beta}\) and \(\hat{\beta}_{\text{pool}}.\) In this case, the averaged estimator is \(\hat{\beta}(w) = w\hat{\beta} + w_2\hat{\beta}_{\text{pool}}.\) Following Maddala et al. (1997), we assume \(\sigma_1^2 = \ldots = \sigma_N^2 = \sigma^2,\) and \(\sigma^2\) can be estimated by \(\tilde{\sigma}^2\) as in (S.3). We consider the case with \(A = X'X.\) If the \(F\) statistic given by (S.4) is larger than 1, such that \(1/F \in [0, 1],\) then by minimizing \(C_A(w)\) we can obtain

$$\hat{w}_2 = \frac{1}{F}.$$  \hspace{1cm} (S.5)

This result suggests that if we only average the pooled and individual estimators and \(1/F \in [0, 1],\) then the Mallows pooling averaging estimator is essentially a Stein-rule estimator (see Equation (2) of Maddala et al. (1997)). The weights of the Mallows pooling averaging estimator and the shrinkage estimator defined by (S.2) are proportional to each other (see also Hansen, 2014). We provide the proof of (S.5) below.

**Proof.** Let \(P_{\text{pool}} = I_{Nk} - (X'X)^{-1}\tilde{R}'(\tilde{R}(X'X)^{-1}\tilde{R}')^{-1}\tilde{R},\) where \(\tilde{R}\) is defined below (S.2). When \(A = X'X,\) we have

\[
C_A(w) = \|w_1\hat{\beta} + w_2\hat{\beta}_{\text{pool}} - \hat{\beta}\|^2_A + 2\text{tr}[(w_1I_{Nk} + w_2P'_{\text{pool}})A\hat{V}_{\text{homo}}] - \|\hat{\beta} - \beta\|^2_A \\
= \|(1 - w_2)\hat{\beta} + w_2\hat{\beta}_{\text{pool}} - \hat{\beta}\|^2_A + 2\text{tr}[(1 - w_2)I_{Nk} + w_2P'_{\text{pool}})A\hat{V}_{\text{homo}}] - \|\hat{\beta} - \beta\|^2_A \\
= w_2^2\|\hat{\beta} - \hat{\beta}_{\text{pool}}\|^2_A + 2w_2\text{tr}(P'_{\text{pool}}A\hat{V}_{\text{homo}}) + 2(1 - w_2)\text{tr}(A\hat{V}_{\text{homo}}) - \|\hat{\beta} - \beta\|^2_A
\]
\[ w_2^2 \| \hat{\beta} - \hat{\beta}_{\text{pool}} \|^2_A + 2w_2 \tilde{\sigma}^2 \{ \text{tr}(P_{\text{pool}}) - Nk \} + 2 \tilde{\sigma}^2 Nk - \| \hat{\beta} - \beta \|^2_A, \]  

(S.6)

where the last two terms have nothing to do with \( w \). From (S.4) and 

\[
\text{tr}((X'X)^{-1}\tilde{R}'(X'X)^{-1}\tilde{R}) \]
\[
= \text{tr}((X'X)^{-1/2}\tilde{R}'(X'X)^{-1}\tilde{R})^{-1}(X'X)^{-1/2}) \]
\[
= \text{rank}(\tilde{R}) \]
\[
= (N - 1)k, \]

we have 

\[
\frac{\tilde{\sigma}^2 \{ Nk - \text{tr}(P_{\text{pool}}) \}}{\| \hat{\beta} - \hat{\beta}_{\text{pool}} \|^2_A} = \frac{\tilde{\sigma}^2 (N - 1)k}{\beta' R' (X'X)^{-1} R \beta} = \frac{1}{F}. \]  

(S.7)

So when \( 1/F \in [0, 1] \), we can obtain (S.5). \( \square \)

## S.3 Additional Monte Carlo Simulation

This section provides various extensions of simulation, including extensions in the data generation process and methods.

### Extensions on Data Generating Process

In the data generation process, we examine how the degree of coefficient heterogeneity affects the performance. Particularly, we consider four data generation processes of heterogeneous panels as in the paper, but the heterogeneous coefficients can be drawn from two sets of coefficient vector \( q \), a set with large degree of heterogeneity \( q_L = [1.0, 1.5, 3.3, 3.0]' \) and a set with smaller size coefficients \( q_S = [1.0, 1.2, 2.3, 2.0]' \) resulting in smaller in-between difference. The choice of \( q \) affects the DGP of weakly and strongly heterogeneous panels (DGP 2 and 3), but not the homogeneous and completely heterogeneous ones (DGP 1 and 4). We also consider two error structures. First, the variance of errors \( \sigma^2 \) is homoskedastic within each individual but varies across individuals (between heteroskedasticity). In this case we could use theoretical \( R^2 \) to determine \( \sigma^2_{ei} \) for individual \( i \). Second, the variance of errors is conditional heteroskedastic for each individual, i.e. \( \sigma^2_{ei} = 0.5|X_{2, it}| \) (completely heteroskedasticity). We also consider smaller sample sizes since we are particularly interested in the finite sample performance. The sample size varies from \( N \in \{5, 10\} \) and \( T \in \{10, 40\} \), leading to four combinations of \( N \) and \( T \).
An Alternative Pre-screening Method

To examine whether the performance of competing methods is sensitive to the choice of pre-screening method, here we consider a different type of clustering method from C-Lasso. The new screening procedure starts with normalizing the estimated coefficients

$$\hat{\beta}_i = \hat{\beta}_i/\max\{||\hat{\beta}_1||, \ldots, ||\hat{\beta}_N||\}$$

for each $l = 1, \ldots, k$, so that the coefficients of the 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. In the second step, we group the normalized coefficient estimates based on their differences.

To incorporate estimation uncertainty in the coefficient estimates, we employ the Bhattacharyya distance. If we assume that the individual estimators are normally distributed, then the Bhattacharyya distance between two coefficient estimates can be obtained by

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

(S.8)

where $\hat{\sigma}_{i,l}^2$ is the estimated variance of $\hat{\beta}_{i,l}$, and $\hat{\sigma}_l^2 = (\hat{\sigma}_{i,l}^2 + \hat{\sigma}_{j,l}^2)/2$. In the third step, we employ agglomerative hierarchical clustering (AHC). In the AHC procedure, 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. A significant advantage of using AHC for clustering is its low computational cost. This algorithm leads to slightly different clustering results from C-Lasso\(^{12}\), so that we can examine the sensitivity of competing methods with respective to pre-screening.

Evaluation criterion

Instead of forecasting, here we present the comparison focusing on the slope coefficient estimates. Hence, we set $A = I_{Nk}$, and the evaluation is based on the square loss of coefficients

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

\(^{12}\textit{k-means}\) leads to almost identical clustering results as C-Lasso, and therefore the similar ultimate estimators.
For the average effect estimators (the pooled estimator) we expand the single estimate of a regressor to an $N \times 1$ vector, so that the comparison can be made with other methods.

Results

Since there are many DGPs in the experimental design, we shall discuss the results in four parts, fixing some of the parameters in each part. We first present the results based on between-individual heteroskedastic errors ($q_L$ and $q_S$, $R^2 = 0.9$). We compare various estimation methods at different degrees of heterogeneity and different sample sizes. Two scales of coefficient parameters also shed some light on the effect of parameter heterogeneity, but from a different perspective. Next, we consider completely (conditional) heteroskedastic errors. In the third part, we add more noise by changing the $R^2$ ceteris paribus.

**Between-individual heteroskedastic errors**

The upper panel of Table S.1 presents the results under parameterization $q_L$ when the errors are only heteroskedastic between individuals. In DGP 1 of a homogeneous panel the pooled estimator always performs best as expected. The proposed MPA estimator is almost as good as the pooled one. This suggests that the Mallows criterion can assign nearly optimal weights in this case. When the panel is characterized by some degree of heterogeneity (DGP 2 and 3), the pooling averaging estimators dominate others in all cases. For the completely heterogeneous DGP 4, we find that shrinkage estimator works uniformly best.

Insert Table S.1 about here

Next, we examine how the size of the coefficients affects the performance of the estimators. The bottom panel of Table S.1 provides the MSE comparison in the similar setting as the upper panel, but under the parameterization $q_S$. Since DGP 1 and 4 are not affected by this parameterization change, we only report the results of DGP 2 and 3. The performance of alternative methods is generally consistent with the results of $q_L$. The main difference is that when the degree of heterogeneity is milder, SBIC often performs best in the weakly heterogeneous case, where MPA, as a close competitor, is always second best. One possible reason that MPA becomes less favorable is that smaller coefficient differences lead to closer candidate models which results in even more inaccurate weight estimation.
In general, we observe that pooling averaging (with different weight choices) performs best in the partially heterogeneous panel, and the pooled and shrinkage are the best choices in the homogeneous and completely heterogeneous panel, respectively. Although MPA is not the best estimator in the homogeneous or completely heterogeneous case, it always has reasonable MSE and is a close competitor with the best estimator.

**Completely heteroskedastic errors**

Table S.2 compares the MSEs in the complete heteroskedasticity case, where the variance of errors differs both within and between individuals. We see that under the performance of competing methods are the same under $q_L$, while under $q_S$ it differs from the previous case in that MPA dominates all alternative methods including SBIC.

Insert Table S.2 about here

**The effect of noise**

So far, the results are all based on DPGs with $R^2 = 0.9$. 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 $R^2 = 0.5$. We report the between-individual heteroskedasticity case in Table S.3.

Insert Table S.3 about here

The upper panel of Table S.3 shows that results under $R^2 = 0.75$ are largely similar to those under $R^2 = 0.9$. An important difference is that MPA now outperforms the shrinkage estimator and is the best method in DGP 4. If we further decrease the $R^2$ to 0.5 as in the bottom panel of Table S.3, we see that pooled estimation is especially favorable in all cases. Even in large noise cases, we see that MPA performs almost as well as pooled estimation.
Reference


Table S.1: MSE comparison: Between-individual heteroskedasticity ($R^2 = 0.9$)

<table>
<thead>
<tr>
<th>DGP</th>
<th>MPA</th>
<th>SAIC</th>
<th>SBIC</th>
<th>AIC</th>
<th>BIC</th>
<th>Pool</th>
<th>SHK</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$q = q_L$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.1684</td>
<td>0.4936</td>
<td>0.1910</td>
<td>0.5596</td>
<td>0.1988</td>
<td><strong>0.1622</strong></td>
<td>0.7570</td>
</tr>
<tr>
<td>$N = 5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td><strong>0.7543</strong></td>
<td>0.8280</td>
<td>0.8532</td>
<td>0.8771</td>
<td>0.9267</td>
<td>2.7942</td>
<td>0.9549</td>
</tr>
<tr>
<td>$T = 15$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td><strong>0.7182</strong></td>
<td>0.7901</td>
<td>0.8440</td>
<td>0.8730</td>
<td>0.9129</td>
<td>2.0200</td>
<td>0.9390</td>
</tr>
<tr>
<td>4</td>
<td>1.1393</td>
<td>1.0870</td>
<td>1.3331</td>
<td>1.1000</td>
<td>1.4027</td>
<td>23.7044</td>
<td><strong>0.9942</strong></td>
</tr>
<tr>
<td>$N = 5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.1919</td>
<td>0.4718</td>
<td>0.1911</td>
<td>0.5370</td>
<td>0.1921</td>
<td><strong>0.1868</strong></td>
<td>0.9146</td>
</tr>
<tr>
<td>$T = 15$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td><strong>0.7770</strong></td>
<td>0.8183</td>
<td>0.9651</td>
<td>0.8897</td>
<td>1.0444</td>
<td>6.0305</td>
<td>0.9914</td>
</tr>
<tr>
<td>4</td>
<td>1.0438</td>
<td>1.0025</td>
<td>1.0479</td>
<td>1.0009</td>
<td>1.0478</td>
<td>73.8860</td>
<td><strong>0.9994</strong></td>
</tr>
<tr>
<td>$N = 10$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.0830</td>
<td>0.4419</td>
<td>0.0835</td>
<td>0.5058</td>
<td>0.0848</td>
<td><strong>0.0795</strong></td>
<td>0.6635</td>
</tr>
<tr>
<td>$T = 15$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td><strong>0.5263</strong></td>
<td>0.6656</td>
<td>0.5725</td>
<td>0.7183</td>
<td>0.5990</td>
<td>1.8677</td>
<td>0.9035</td>
</tr>
<tr>
<td>4</td>
<td>0.9392</td>
<td>0.9970</td>
<td>1.1446</td>
<td>1.0382</td>
<td>1.1930</td>
<td>7.5001</td>
<td><strong>0.9715</strong></td>
</tr>
<tr>
<td>$N = 10$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.0944</td>
<td>0.4374</td>
<td>0.0913</td>
<td>0.5019</td>
<td>0.0914</td>
<td><strong>0.0908</strong></td>
<td>0.8825</td>
</tr>
<tr>
<td>$T = 40$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td><strong>0.4477</strong></td>
<td>0.7038</td>
<td>0.5156</td>
<td>0.7601</td>
<td>0.5432</td>
<td>5.7051</td>
<td>0.9853</td>
</tr>
<tr>
<td>4</td>
<td>1.1010</td>
<td>1.0957</td>
<td>1.4928</td>
<td>1.1423</td>
<td>1.5476</td>
<td>23.1843</td>
<td><strong>0.9964</strong></td>
</tr>
<tr>
<td>$q = q_S$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N = 5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td><strong>0.6961</strong></td>
<td>0.7499</td>
<td>0.7035</td>
<td>0.7918</td>
<td>0.7248</td>
<td>3.6014</td>
<td>0.9635</td>
</tr>
<tr>
<td>$T = 15$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td><strong>0.7729</strong></td>
<td>0.8415</td>
<td>0.8451</td>
<td>0.9054</td>
<td>0.8811</td>
<td>2.6861</td>
<td>0.9527</td>
</tr>
<tr>
<td>$N = 10$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.5167</td>
<td>0.6439</td>
<td><strong>0.4937</strong></td>
<td>0.6850</td>
<td>0.5038</td>
<td>10.8465</td>
<td>0.9953</td>
</tr>
<tr>
<td>$T = 40$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td><strong>0.8481</strong></td>
<td>0.9156</td>
<td>1.0326</td>
<td>1.0149</td>
<td>1.0779</td>
<td>8.0571</td>
<td>0.9934</td>
</tr>
<tr>
<td>$N = 10$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.4436</td>
<td>0.5994</td>
<td><strong>0.4301</strong></td>
<td>0.6373</td>
<td>0.4447</td>
<td>2.7782</td>
<td>0.9312</td>
</tr>
<tr>
<td>$T = 15$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td><strong>0.5826</strong></td>
<td>0.7270</td>
<td>0.6096</td>
<td>0.7743</td>
<td>0.6299</td>
<td>2.4945</td>
<td>0.9235</td>
</tr>
<tr>
<td>$N = 10$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.2328</td>
<td>0.5485</td>
<td><strong>0.2321</strong></td>
<td>0.6028</td>
<td>0.2370</td>
<td>8.6291</td>
<td>0.9900</td>
</tr>
<tr>
<td>$T = 40$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td><strong>0.6399</strong></td>
<td>0.8224</td>
<td>0.7392</td>
<td>0.8797</td>
<td>0.7693</td>
<td>7.6160</td>
<td>0.9889</td>
</tr>
</tbody>
</table>
Table S.2: MSE comparison: Complete heteroskedasticity

<table>
<thead>
<tr>
<th>DGP</th>
<th>MPA</th>
<th>SAIC</th>
<th>SBIC</th>
<th>AIC</th>
<th>BIC</th>
<th>Pool</th>
<th>SHK</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$q = q_L$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.2091</td>
<td>0.7450</td>
<td>0.3736</td>
<td>0.7983</td>
<td>0.4133</td>
<td><strong>0.1980</strong></td>
<td>0.8441</td>
</tr>
<tr>
<td>$N = 5$</td>
<td>2</td>
<td><strong>0.4179</strong></td>
<td>0.7623</td>
<td>0.5307</td>
<td>0.8156</td>
<td>0.5609</td>
<td>47.813</td>
</tr>
<tr>
<td>$T = 15$</td>
<td>3</td>
<td><strong>0.9397</strong></td>
<td>1.0031</td>
<td>1.1062</td>
<td>1.0374</td>
<td>1.1813</td>
<td>32.129</td>
</tr>
<tr>
<td>4</td>
<td>1.0785</td>
<td>1.0972</td>
<td>1.2796</td>
<td>1.1508</td>
<td>1.3439</td>
<td>3.7804</td>
<td><strong>0.9745</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$q = q_S$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N = 5$</td>
<td>2</td>
<td><strong>0.4183</strong></td>
<td>0.7781</td>
<td>0.5003</td>
<td>0.8287</td>
<td>0.5222</td>
<td>123.66</td>
</tr>
<tr>
<td>$T = 40$</td>
<td>3</td>
<td><strong>0.9456</strong></td>
<td>0.9636</td>
<td>1.0529</td>
<td>0.9805</td>
<td>1.0804</td>
<td>83.563</td>
</tr>
<tr>
<td>4</td>
<td>1.2184</td>
<td>1.1765</td>
<td>1.5973</td>
<td>1.2328</td>
<td>1.6663</td>
<td>9.3390</td>
<td><strong>0.9664</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N = 10$</td>
<td>2</td>
<td><strong>0.2135</strong></td>
<td>0.7348</td>
<td>0.3112</td>
<td>0.7753</td>
<td>0.3307</td>
<td>46.198</td>
</tr>
<tr>
<td>$T = 15$</td>
<td>3</td>
<td><strong>0.7886</strong></td>
<td>0.9094</td>
<td>0.9230</td>
<td>0.9377</td>
<td>0.9656</td>
<td>33.723</td>
</tr>
<tr>
<td>4</td>
<td>1.1919</td>
<td>1.1490</td>
<td>1.4522</td>
<td>1.1932</td>
<td>1.4989</td>
<td>14.381</td>
<td><strong>0.9884</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N = 10$</td>
<td>2</td>
<td><strong>0.2169</strong></td>
<td>0.7471</td>
<td>0.2756</td>
<td>0.7897</td>
<td>0.2905</td>
<td>122.17</td>
</tr>
<tr>
<td>$T = 40$</td>
<td>3</td>
<td><strong>0.7757</strong></td>
<td>0.8568</td>
<td>0.8060</td>
<td>0.8870</td>
<td>0.8193</td>
<td>88.164</td>
</tr>
<tr>
<td>4</td>
<td>1.3016</td>
<td>1.2198</td>
<td>1.7762</td>
<td>1.2708</td>
<td>1.8276</td>
<td>37.574</td>
<td><strong>0.9984</strong></td>
</tr>
</tbody>
</table>

55
Table S.3: More noise in the model ($N = 10$, $T = 15$)

<table>
<thead>
<tr>
<th>$q$</th>
<th>DGP</th>
<th>MPA</th>
<th>SAIC</th>
<th>SBIC</th>
<th>AIC</th>
<th>BIC</th>
<th>Pool</th>
<th>SHK</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2 = 0.75$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.0808</td>
<td>0.4518</td>
<td>0.0809</td>
<td>0.5144</td>
<td>0.0817</td>
<td><strong>0.0775</strong></td>
<td>0.6637</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>0.2998</strong></td>
<td>0.5859</td>
<td>0.3520</td>
<td>0.6532</td>
<td>0.3780</td>
<td>0.3157</td>
<td>0.7483</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>0.2619</strong></td>
<td>0.5125</td>
<td>0.3108</td>
<td>0.5655</td>
<td>0.3409</td>
<td>0.2807</td>
<td>0.7400</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>0.5063</strong></td>
<td>0.7669</td>
<td>0.8151</td>
<td>0.8249</td>
<td>0.8875</td>
<td>0.9004</td>
<td>0.8463</td>
<td></td>
</tr>
<tr>
<td>$q_L$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.0820</td>
<td>0.4455</td>
<td>0.0836</td>
<td>0.5050</td>
<td>0.0846</td>
<td><strong>0.0782</strong></td>
<td>0.6647</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>0.3489</strong></td>
<td>0.6309</td>
<td>0.4106</td>
<td>0.6800</td>
<td>0.4510</td>
<td>0.3815</td>
<td>0.7650</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>0.3125</strong></td>
<td>0.6116</td>
<td>0.3775</td>
<td>0.6609</td>
<td>0.4168</td>
<td>0.3474</td>
<td>0.7596</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>0.5077</strong></td>
<td>0.7735</td>
<td>0.8181</td>
<td>0.8293</td>
<td>0.8914</td>
<td>0.9130</td>
<td>0.8474</td>
<td></td>
</tr>
<tr>
<td>$q_S$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.0812</td>
<td>0.4294</td>
<td>0.0809</td>
<td>0.4882</td>
<td>0.0817</td>
<td><strong>0.0785</strong></td>
<td>0.6627</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>0.3489</strong></td>
<td>0.6309</td>
<td>0.4106</td>
<td>0.6800</td>
<td>0.4510</td>
<td>0.3815</td>
<td>0.7650</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>0.3125</strong></td>
<td>0.6116</td>
<td>0.3775</td>
<td>0.6609</td>
<td>0.4168</td>
<td>0.3474</td>
<td>0.7596</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>0.5077</strong></td>
<td>0.7735</td>
<td>0.8181</td>
<td>0.8293</td>
<td>0.8914</td>
<td>0.9130</td>
<td>0.8474</td>
<td></td>
</tr>
<tr>
<td>$R^2 = 0.5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.0812</td>
<td>0.4302</td>
<td>0.0817</td>
<td>0.4905</td>
<td>0.0827</td>
<td><strong>0.0777</strong></td>
<td>0.6636</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>0.1184</strong></td>
<td>0.4373</td>
<td>0.1446</td>
<td>0.4856</td>
<td>0.1552</td>
<td><strong>0.1116</strong></td>
<td>0.6831</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>0.1143</strong></td>
<td>0.4403</td>
<td>0.1346</td>
<td>0.4868</td>
<td>0.1431</td>
<td><strong>0.1074</strong></td>
<td>0.6825</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>0.1795</strong></td>
<td>0.4748</td>
<td>0.2384</td>
<td>0.5204</td>
<td>0.2533</td>
<td><strong>0.1692</strong></td>
<td>0.7079</td>
<td></td>
</tr>
</tbody>
</table>