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
CEFS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences
Department of Statistics, Pennsylvania State University

Richard Paap
Econometric Institute, Erasmus University Rotterdam

Abstract: This paper considers estimating the slope parameters and forecasting in potentially heterogeneous panel data regressions with a long time dimension. 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 and robust in extreme cases. Our results provide practical guidance for the best estimator and forecast depending on features of data and models. We apply our method to examine the determinants of sovereign credit default swap spreads and forecast future spreads.

Keywords: Credit default swap spreads; Heterogeneous panel; Model screening; Panel data forecasting; Pooling averaging

JEL Classification: C23, C52, G15

*Address for correspondence: X. Zhang, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, 55 Zhong-Guan-Cun East Road, Beijing, China, e-mail: xinyu@amss.ac.cn. We thank the editor, Jonathan Wright, and the two referees for helpful comments. We are indebted to Pavel Ciezek, Herman van Dijk, Qingfeng Liu, Jan Magnus, Ryo Okui, Alan Wan, and participants at seminars at Renmin University, Xiamen University, University of Sydney, the 10th NESG at Maastricht, the 21st international panel data conference at Budapest, Econometric Society 2016 China Meeting at Chengdu and Australasia Meeting at Sydney, and the 2nd Ecosta conference at Hong Kong, for constructive comments. Wang acknowledges the support of an EUR fellowship. Zhang acknowledges the support of the NNSFC (71522004, 71463012, 71631008). 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 a country's 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, namely, whether the sovereign credit risk is driven by local economic performance or by forces from the global market. Furthermore, it is of equal significance and interest to forecast future sovereign credit risks. A good forecast could not only serve as a crucial basis of an effective policy but also directly influence the ability of financial market participants to diversify risks. 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 (Longstaff et al., 2011; Dieckmann and Plank, 2012). The literature suggests that although some common patterns could be found within certain sub-samples, the potential determinants exhibit significant heterogeneity across countries. Hence, to examine the determinants of sovereign CDS spreads and forecast the future spreads, it is important to model heterogeneity in the best way.

Existing studies on sovereign CDS spreads are either based on individual time series regressions 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 models (Remolona et al., 2008) ignore the heterogeneity and assumes homogeneous coefficients for all countries. The question is 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. 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 produce 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, many empirical 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 then 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 using panel data models.

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. 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 the number of pooling specifications can be large even for a moderate number of individuals and regressors. Averaging over the entire model space could be computationally intensive and inefficient. Consequently, we introduce a model screening procedure to address this issue. Instead of estimating and averaging over all candidate models, we propose to first cluster individual units based on the similarity of parameter estimates and average over group estimators obtained from different numbers of groups. This would result in a much smaller post-screened model space, where averaging is computationally more efficient and accurate. In addition, we find that averaging 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.

There are of course several ways to address heterogeneity in the literature, for example, the random coefficient model (RCM, Swamy, 1970), the pooled mean group estimator (Pesaran et al., 1999), various group estimators (e.g. Bonhomme and Manresa, 2015; 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 the homogeneity assumption. The estimator obtained after a preliminary test is called the pretest estimator. Our proposed pooling averaging method 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, so parameters can be heterogeneous in any pattern. (iii) Our approach avoids the problems caused by pretesting since it is continuous, unconditional, and takes model uncertainty into account (Danilov and Magnus, 2004).

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. From this analysis, we provide empirical researchers with guidelines on how to handle parameter heterogeneity in panel data models. Given 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 show that there is no single method that performs best in all situations, and the performance depends on the features of data and models, including 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 simulation studies. We conclude that the pooling averaging estimator/forecast, especially Mallows pooling averaging, is recommended when the panel is heterogeneous and the signal-to-noise ratio is moderate or large, while the pooled estimator/forecast is recommended when individual units have homogeneous slope parameters and a small signal-to-noise ratio.

After showing how to deal with slope heterogeneity in an optimal way, we turn back to our empirical question, examining and forecasting sovereign CDS spreads for a panel of countries. We extend the time dimension of the data of Longstaff et al. (2011) 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.

The remainder of this paper is organized as follows. Section 2 briefly reviews parameter estimation and testing strategies for heterogeneous panel data models. Section 3 discusses the model setup and introduces the general pooling averaging estimator. Section 4 presents 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. Section 6 introduces a model screening procedure. A simulation study is provided in Section 7. Section 8 provides an empirical study of sovereign CDS spreads. Finally, Section 9 offers some practical suggestions on how to handle slope heterogeneity based on the theoretical, simulation and empirical results. Technical proofs of all theorems and additional numerical results are provided in the online appendix.

2 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. Estimating heterogeneous slope parameters in panel models can be dated back to Swamy (1970), who proposed a RCM and estimated the mean of the heterogeneous coefficient (average effect) using generalized least squares (GLS). 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 first aggregates the data over individuals and then estimates aggregated time-series regressions. A seminal and comprehensive study of aggregation estimation is given by Pesaran et al. (1989). See Hsiao and Pesaran (2008) for a comprehensive review of methods for RCMs.

If the average effect is of interest, the GLS estimator is shown to provide an optimal trade-off between bias and efficiency (Swamy, 1970). However, researchers sometimes are more interested in the individual parameters (individual-specific effect), especially when providing heterogeneous policy implications and decisions or forecasting individual time series is the primary goal of research. This paper focuses on estimating the individual-specific effect and making individual forecast. 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 however, is not required in our set-up. See Baltagi et al. (2008) for an excellent survey of estimating heterogeneous coefficients in panel models.

Recent developments in individual-specific-effect estimation involve a latent class/group specification, such as finite mixture models and various grouped estimators. For example, Ando and Bai (2016) estimated the panel data models with unobserved group factors. Like the pooled mean group estimator, a heterogeneity structure is imposed where only the factors and their loadings are heterogeneous. To estimate group-specific parameters and the unknown group membership structure, various techniques have been proposed. 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. (2018) extended clustering algorithm in regression via data-driven segmentation (CARDS) to a panel framework. These group estimators have noticeable merits, especially when identifying the true grouping is of interest.

Unlike these existing approaches, our objective is not to *consistently* estimate the group membership and/or slope coefficients, but obtain the *most accurate* parameter estimator or forecast in terms of MS(F)E. We make a simple, yet largely overlooked point, namely, 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 and forecasts, most existing approaches 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 choosing appropriate weights.

An alternative way of dealing with the potential heterogeneity is to use statistical pretests for parameter homogeneity under different model specifications. To give a partial list of possible tests, Pesaran and Yamagata (2008) proposed dispersion type tests for large panels with large cross section and time dimension. Juhl and Lugovskyy (2014) focused on the typical micro-panel with large N and fixed T. Su and Chen (2013) proposed a residual-based test applicable in panel models with interactive fixed effects. If the researchers' ultimate interest lies in estimating the parameters of the models, the pretest estimator is however not completely satisfactory. The estimator is discontinuous and the testing result may depend on some arbitrarily chosen significance level. These are the problems of pretesting estimators; see Danilov and Magnus (2004) for a detailed discussion. Moreover, how to proceed after testing remains unclear. For example, if the hypothesis of homogeneity is rejected, whether we should estimate individuals separately or continue testing the subsamples? 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.

Given this result, we propose a pooling average estimator and analyse its finite sample and asymptotic properties. 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). A recent study by Hashem and Zhou (2018) provided a comparative analysis of pooled least squares versus fixed effects estimators of the slope coefficients. While they focus on pooling in standard panel data models with homogeneous slopes, we consider the pooling decision in the model with cross-sectionally heterogeneous slopes. Our study is also related with the panel forecast literature; See the review by Baltagi (2008) and references therein. Most studies are based on the best linear unbiased prediction. Instead of producing an unbiased predictor, we focus on achieving the minimum MSFE by balancing the bias and efficiency.

A limitation of our approach is that we only provide point estimates for the coefficients, and thus statistical inference is challenging. To address this issue, 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 deserves future research.

3 POOLING AVERAGING ESTIMATION

3.1 Model setup

Consider the linear panel data model with heterogeneous slopes

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

where $y_i = (y_{i1}, \ldots, y_{iT})'$ and $X_i = (X'_{i1}, \ldots, X'_{iT})'$ is a $T \times k$ matrix of explanatory variables including the intercept, that is, $X_{it1} = 1$ for all i, t. We assume that the series $\{y_{it}, X_{it}\}$ are both stationary for all $i = 1, \ldots, N$, and thus we rule out the possibility of time-varying slope coefficients. 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 β_i can be different from the elements in β_j for $i \neq j$.[1] 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 $\sigma_i^2 I_T$ (between-individual heteroskedasticity). Later we shall relax this assumption by allowing conditional heteroskedastic errors both between and within individuals. 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

$$y = X\beta + u, \tag{2}$$

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 derive our pooling estimator 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.[2]

3.2 Average pooling strategies

Our goal is to estimate each individual coefficient β_i or forecast individual 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 RCM where one wants to estimate a common average effect, say $\mathrm{E}(\beta_i)$.

To estimate the β_i parameters in (1), one can consider separate least-square (LS) estimators for each time series as long as $T > k$, $\widehat{\beta}_i = (X'_i X_i)^{-1} X'_i y_i$, called the individual estimator. The

[1] Note that we do not assume that β_is share a common mean and a common variance as in RCMs, nor do we require a group pattern of coefficient heterogeneity as in panel (group) structure models (Su et al., 2016).

[2] In the dynamic panel and the presence of omitted variables, least square estimation of individual time series is biased, and theorems derived later in Section 5.1 do not hold. Nevertheless, the bias-variance trade-off remains relevant.

individual estimator $\widehat{\beta}_i$ is unbiased given that individual i's regression is correctly specified. However, this estimator does not make use of any cross-section variation at all, and thus its variance can be larger than that of a pooled LS estimators.

In the other extreme, one could ignore the slope heterogeneity and estimate the pooled model, obtaining a common estimator for all individuals, that is,

$$b = (\sum_{i=1}^{N} X_i' X_i)^{-1} \sum_{i=1}^{N} X_i' X_i \widehat{\beta}_i.$$

The pooled estimator $\widehat{\beta}_{\text{pool}} = (b', \ldots, b')'$ has smaller variances 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 outcome variable that associates with $\widehat{\beta}_i$ and $\widehat{\beta}_{\text{pool}}$ can be obtained by $\widehat{y}_i = X\widehat{\beta}_i$ and $\widehat{y}_{\text{pool}} = X\widehat{\beta}_{\text{pool}}$, respectively, and they face precisely the same bias-variance trade-off.

An intermediate estimator (between the individual and pooled estimators) restricts some of the coefficients to be identical, which is obtained by imposing equality restrictions to a set of coefficients when estimating (2), that is,

$$R_m \beta = 0, \tag{3}$$

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 $R_m = \left(0_{k\times(i-1)k}, I_k, 0_{k\times(j-i-1)k}, -I_k, 0_{k\times(N-j)k}\right)$. For each R_m, we can construct the projection matrix P_m

$$P_m = I_{Nk} - (X'X)^{-1} R_m' (R_m (X'X)^{-1} R_m')^{-1} R_m, \tag{4}$$

so that the OLS estimator under the m-th pooling strategy is

$$\widehat{\beta}_{(m)} = P_m \widehat{\beta}, \tag{5}$$

where $\widehat{\beta} = (\widehat{\beta}_1', \ldots, \widehat{\beta}_N')'$ is the vector of individual OLS estimators. The estimator $\widehat{\beta}_{(m)}$ allows estimated coefficients to vary over individuals while restricting some of them to be the same. Different pooling strategies are characterized by different restrictions R_m, and the resulting estimators have different degrees of bias and variance. The question is then 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, the true model is difficult to select because it is hard to distinguish whether the efficiency loss is from inefficient pooling or 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 and the signal-to-noise ratio are both small. In this case pooling heterogeneous individuals incorrectly may lead to lower

MS(F)E, because the efficiency gains from pooling dominate the heterogeneity bias. Therefore, if the MS(F)E of the coefficient estimates or the forecasts of the y_i variables are of central interest, it is less plausible to test or select the right pooling pattern.

To achieve 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

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

where 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 $\mathcal{W} = \{w \in [0,1]^M : \sum_{m=1}^{M} w_m = 1\}$.[3] Its associated combined forecast is $\widehat{y}(w) = X\widehat{\beta}(w)$. In practice, the number of pooling strategies M can be substantial, and in this case we propose to "screen out" poor pooling strategies as a preliminary step based on efficient clustering in Section 6.

Our pooling averaging estimator is in sharp contrast to the average-effect estimators. Since each pooling estimator $\widehat{\beta}_{(m)}$ provides estimates of slope coefficients for each individual, averaging over M pooling estimators leads to potentially different *individual-specific* estimates. This differs from the average-effect estimator (e.g. Pesaran and Smith, 1995; Swamy, 1970) that produce a common coefficient estimate for all units. We shall compare $\widehat{\beta}(w)$ with average-effect estimators in the Monte Carlo simulation and applications.

Our pooling averaging estimator (6) includes the pretesting estimator as a special case that assigns all weights to a single candidate estimator. It is also closely related with the popular shrinkage estimator provided by Maddala et al. (1997), defined as

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

where $\nu = [(N-1)k-2]/[NT-Nk+2]$ and F is the test statistic for null hypothesis $H_0 : \beta_1 = \ldots = \beta_N$. When the "weight" ν/F is between 0 and 1, $\widehat{\beta}_{\text{shrinkage}}$ can be regarded as a special case of the pooling average estimator that combines only the pooled and individual estimators (see the online appendix for further discussion on their relationship).

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.

[3] The pooling averaging estimator can be written as a weighted average of the individual estimator $\widehat{\beta}$ as in (6), where the associated weight $w_m P_m$ is a matrix. Instead of optimizing the scalar weight w_m, direct optimization of the whole matrix $w_m P_m$ (assuming P_m is unknown) is possible. This is not only computationally more difficult but also not efficient given that P_m is observed.

To save space, the discussion will focus 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 first assume that the weights are non-random. In Section 4.3 we consider random weights which corresponds to the situation where the weights have to be estimated from the data.

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 θ, where $A = \mathrm{diag}(A_1, \ldots, A_N) = X'X$ and $A_i = X_i'X_i$ for $i = 1, \ldots, N$.[4] 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 $V = \mathrm{diag}(V_1, \ldots, V_N)$. The analysis can easily be extended to (completely) conditional heteroskedastic errors but with more notational complexity.

4.1 MSFE of pooled and individual forecasts

The pooled forecast can be obtained by $\widehat{y}_{\text{pool}} = X\widehat{\beta}_{\text{pool}}$, where $\widehat{\beta}_{\text{pool}} = (b', \ldots, b')'$ and $b = Q^{-1}\sum_{i=1}^N Q_i\widehat{\beta}_i$. The individual forecast is based on individual estimators, that is, $\widehat{y}_{\text{ind}} = (\widehat{y}_1', \ldots, \widehat{y}_N')'$ with $\widehat{y}_i = X\widehat{\beta}_i$, and these individual estimators $\widehat{\beta}_i$'s are uncorrelated with $\widehat{\beta}_i \sim (\beta_i, \sigma_i^2 Q_i^{-1}/T)$.[5] Hence, the MSFEs of the individual and pooled forecasts can be obtained by

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

and

$$\begin{aligned} \mathrm{MSFE}_{\text{pool}} &\equiv \mathrm{MSFE}(\widehat{y}_{\text{pool}}) = \sum_{i=1}^N \mathrm{E}\|b - \beta_i\|_{A_i}^2 \\ &= \sum_{i=1}^N \|Q^{-1}\sum_{i=1}^N Q_i\beta_i - \beta_i\|_{A_i}^2 + \frac{N}{T}\sum_{i=1}^N \mathrm{tr}(\sigma_i^2 Q^{-1}Q_iQ^{-1}A_i). \end{aligned} \tag{9}$$

[4] The comparison of slope coefficient estimates can be made by setting $A = I_{Nk}$.

[5] 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, $\mathrm{MSFE}_{\mathrm{ind}}$ is generally of 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 a trade-off if (8) and the first term of (9) are of comparable scale. Furthermore, there is no guarantee that $\mathrm{MSFE}_{\mathrm{ind}}$ is less than $\mathrm{MSFE}_{\mathrm{pool}}$ in finite samples. The relation between the finite sample $\mathrm{MSFE}_{\mathrm{ind}}$ and $\mathrm{MSFE}_{\mathrm{pool}}$ depends on the magnitude of the bias term $\sum_{i=1}^{N}\|Q^{-1}\sum_{i=1}^{N}Q_i\beta_i-\beta_i\|_{A_i}^2$ and the difference between two scaled variance terms $\frac{1}{T}\sum_{i=1}^{N}\sigma_i^2\mathrm{tr}(Q_i^{-1}A_i)-\frac{N}{T}\sum_{i=1}^{N}\mathrm{tr}(\sigma_i^2Q^{-1}Q_iQ^{-1}A_i)$. 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.

4.2 MSFE of pooling averaging forecast with fixed weights

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

$$\begin{aligned}\mathrm{MSFE}_{\mathrm{fw}}(\widehat{y}(w)) &= \mathrm{E}\|\widehat{\beta}(w)-\beta\|_A^2=\mathrm{E}\|P(w)\widehat{\beta}-\beta\|_A^2\\ &= \|P(w)\beta-\beta\|_A^2+\mathrm{tr}[P(w)VP'(w)A]. \qquad (10)\end{aligned}$$

We see that the comparison between the MSFEs depends on the degree of heterogeneity in the *true* coefficients β, the error variances of individual regressions σ_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 $\widehat{\beta}(w)$ only averages over the pooled and individual estimators, namely $\widehat{\beta}(w)=w_1\widehat{\beta}_{\mathrm{pool}}+w_2\widehat{\beta}$, we can write $\mathrm{MSFE}_{\mathrm{fw}}(\widehat{y}(w))$ in terms of $\mathrm{MSFE}_{\mathrm{pool}}$ and $\mathrm{MSFE}_{\mathrm{ind}}$ as

$$\begin{aligned}\mathrm{MSFE}_{\mathrm{fw}}(\widehat{y}(w)) &= \sum_{i=1}^{N}E\|w_1b+w_2\widehat{\beta}_i-\beta_i\|_{A_i}^2\\ &= \sum_{i=1}^{N}E\|w_1Q^{-1}\sum_{i=1}^{N}Q_i\widehat{\beta}_i+w_2\widehat{\beta}_i-\beta_i\|_{A_i}^2\\ &= \sum_{i=1}^{N}\|w_1Q^{-1}\sum_{i=1}^{N}Q_i\beta_i+w_2\beta_i-\beta_i\|_{A_i}^2+\sum_{j=1}^{N}\mathrm{var}(w_1A_i^{1/2}Q^{-1}\sum_{i=1}^{N}Q_i\widehat{\beta}_i+w_2A_i^{1/2}\widehat{\beta}_j)\\ &= w_1^2\mathrm{MSFE}_{\mathrm{pool}}+w_2^2\mathrm{MSFE}_{\mathrm{ind}}+2w_1w_2\frac{1}{T}\sum_{i=1}^{N}\sigma_i^2\mathrm{tr}(Q^{-1}A_i). \qquad (11)\end{aligned}$$

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

$$\mathrm{MSFE}_{\mathrm{ind}} = k\sum_{i=1}^{N}\sigma_i^2, \qquad \mathrm{MSFE}_{\mathrm{pool}} = \sum_{i=1}^{N}\|\bar{\beta}-\beta_i\|_{A_i}^2 + \frac{k}{N}\sum_{i=1}^{N}\sigma_i^2, \tag{12}$$

and

$$\mathrm{MSFE}_{\mathrm{fw}}(\widehat{y}(w)) = w_1^2\sum_{i=1}^{N}\|\bar{\beta}-\beta_i\|_{A_i}^2 + w_2^2\frac{(N-1)k}{N}\sum_{i=1}^{N}\sigma_i^2 + \frac{k}{N}\sum_{i=1}^{N}\sigma_i^2, \tag{13}$$

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

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

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 $\sum_{i=1}^{N}\|\bar{\beta}-\beta_i\|_{A_i}^2 = 0 \leq w_2^2(N-1)k\sum_{i=1}^{N}\sigma_i^2/[N(1-w_1^2)]$, it always holds that $\mathrm{MSFE}_{\mathrm{fw}}(\widehat{y}(w)) \geq \mathrm{MSFE}_{\mathrm{pool}}$ 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 σ_i^2 and/or the number of regressors k increase.

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

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

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.

4.3 MSFE of pooling averaging forecast with random weights

Next, we consider the case where weights are functions of the data. These weights are random and correlated with the estimated coefficients as they are estimated from the same data. Furthermore, the weights can also be correlated with each other. We define $\rho_m = \mathrm{cov}(w_m, \widehat{\beta})$,

$\kappa_{m,l} = \mathrm{cov}(w_m, w_l)$, and $\delta_{m,l} = \mathrm{cov}(w_m w_l, \widehat{\beta}' P_m' P_l \widehat{\beta})$ for $m, l \in \{1, \ldots, M\}$. Let $\bar{w} = \mathrm{E}(w)$. Now the MSFE of the pooling averaging forecast with random weights can be written as

$$\begin{aligned} \mathrm{MSFE}_{\mathrm{rw}}(\widehat{y}(w)) &= \mathrm{E}\|P(w)\widehat{\beta} - \beta\|^2 = \mathrm{E}\|P(w)\widehat{\beta}\|^2 - 2\mathrm{E}\left\{\beta' P(w)\widehat{\beta}\right\} + \|\beta\|^2 \\ &= \|P(\bar{w})\beta - \beta\|^2 + \mathrm{tr}\left[P(\bar{w})VP'(\bar{w})\right] + \iota'\Phi\iota + \Gamma_1 - 2\Gamma_2, \end{aligned}$$

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

$$\Gamma_1 = \sum_m \sum_l \kappa_{m,l}\left[\beta' P_m' P_l \beta + \mathrm{tr}\left(P_m' P_l V\right)\right], \quad \text{and} \quad \Gamma_2 = \sum_m \beta' P_m \rho_m.$$

The first two terms of $\mathrm{MSFE}_{\mathrm{rw}}(\widehat{y}(w))$ are similar to the terms of $\mathrm{MSFE}_{\mathrm{fw}}(\widehat{y}(w))$ in (10). Hence, the degree of coefficient heterogeneity and the size of noise play a similar role in MSFE comparison as in the fixed-weight case. Estimating the weights from the data however introduces three extra covariance terms, which make the evaluation of the the $\mathrm{MSFE}_{\mathrm{rw}}(\widehat{y}(w))$ more complicated. To examine how the similarity of models affects the MSFE, we can rewrite the $\mathrm{MSFE}_{\mathrm{rw}}(\widehat{y}(w))$ as

$$\begin{aligned} \mathrm{MSFE}_{\mathrm{rw}}(\widehat{y}(w)) = & \sum_m \sum_l \mathrm{E}(w_m w_l)\left[\mathrm{E}\left(\widehat{\beta}_{(m)}\right)' \mathrm{E}\left(\widehat{\beta}_{(l)}\right) + \mathrm{tr}(\mathrm{cov}(\widehat{\beta}_{(m)}, \widehat{\beta}_{(l)}))\right] \\ & + \sum_m \sum_l \delta_{m,l} - 2\sum_m \bar{w}_m \mathrm{E}\left(\beta' P_m \widehat{\beta}\right) - 2\sum_m \beta' P_m \rho_m + \|\beta\|^2. \end{aligned}$$

When candidate models are well-differentiated, resulting in a small $\mathrm{tr}(\mathrm{cov}(\widehat{\beta}_{(m)}, \widehat{\beta}_{(l)}))$, $\mathrm{MSFE}_{\mathrm{rw}}(\widehat{y}(w))$ is likely to be small and pooling averaging using random weights is often more desirable. This is in line with the conventional wisdom in the forecast combination literature that the diversification gains from combination tend to be larger if candidate forecasts are strongly dissimilar (see, e.g., Timmermann (2006) and Claeskens et al. (2016) for theoretical discussions). The influence of the correlation between the weights and estimated slope coefficient estimates is less clear as it depends on the explanatory variables and true values of the coefficients.

5 CHOOSING POOLING AVERAGING WEIGHTS

We have seen in Section 4 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 $\|\theta\|_A^2 = \theta' A\theta$ for any vector θ 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 heteroskedasticicy between and within individuals. 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 $\widehat{\beta}_i$ 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 $\sigma_i^2 Q_i^{-1}/T$ as used before in Section 4.

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

$$\mathcal{C}_A(w) = \|P(w)\widehat{\beta} - \widehat{\beta}\|_A^2 + 2\text{tr}[P'(w)AV] - \|\widehat{\beta} - \beta\|_A^2, \tag{16}$$

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.[6]

The criterion $\mathcal{C}_A(w)$ is a generalization of the Mallows model averaging criterion defined by Hansen (2007). When $\Omega_i = \sigma^2 I_T$ for all i and $A = X'X$, $\mathcal{C}_A(w)$ simplifies to Hansen's criterion (Equation (11) in Hansen (2007)), which focuses on the average (forecasting) squared error loss $(X\widehat{\beta}(w) - \text{E}(y))'(X\widehat{\beta}(w) - \text{E}(y))$. When we set $A = I_{Nk}$, (16) extends Hansen's (2007) criterion to concentrate on the accuracy of the estimated coefficients, and $\mathcal{C}_A(w)$ 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.[7]

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

$$\mathcal{C}_A^*(w) = \|P(w)\widehat{\beta} - \widehat{\beta}\|_A^2 + 2\text{tr}[P'(w)A\widehat{V}] - \|\widehat{\beta} - \beta\|_A^2, \tag{17}$$

and the feasible weight vector is obtained by

$$\widehat{w}^* = \arg\min_{w\in\mathcal{W}} \mathcal{C}_A^*(w). \tag{18}$$

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 $\widehat{V}_{\text{homo}} = \widetilde{\sigma}^2(X'X)^{-1}$, where $\widetilde{\sigma}^2$ is the variance of residuals associated with from the individual OLS estimator, i.e. $\widetilde{\sigma}^2 = (Y - X\widehat{\beta})'(Y - X\widehat{\beta})/(NT - Nk)$.

[6] See the online appendix for the proof.

[7] See the online appendix for the details and proof.

2. Between-individual heteroskedasticity: If we assume that $\text{var}(u_i) = \sigma_i^2 I_T$, we consider $\widehat{V}_{\text{bh}} = \text{diag}(\widehat{\sigma}_1^2 Q_1^{-1}, \ldots, \widehat{\sigma}_N^2 Q_N^{-1})/T$, where $\widehat{\sigma}_i^2 = \widehat{u}_i' \widehat{u}_i/(T-k)$ and $\widehat{u}_i$ is the OLS residual of the i-th individual regression.

3. Heteroskedasticity between and within individuals: If we assume that u_i is (conditional) heteroskedastic for each $i = 1, \ldots, N$, the most general situation, we use

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

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

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 obtaining its risk bound. 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 the features of data.

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

$$E\{L_A(\widehat{w})\} \ \ \leq \ \ \frac{1}{1-c} \inf_{w \in \mathcal{W}} R_A(w) + \frac{1}{1-c} \left(\frac{1}{c} tr(AV) - 2E(tr\{P'(\widehat{w})AV\}) \right), \tag{19}$$

where c is a constant belonging to $(0, 1)$.
Proof*: See Section A.1 of the online appendix.*

It shows that up to the constant $(1-c)^{-1}$ and the additive penalty $(1-c)^{-1}[c^{-1}\text{tr}(AV) - 2\text{E}(\text{tr}\{P'(\widehat{w})AV\})]$, Mallows pooling averaging estimator $\widehat{\beta}(\widehat{w})$ has the same risk performance as the averaging estimator using the optimal weights, $\inf_{w \in \mathcal{W}} R_A(w)$. The result of (19) does not depend on sample size. To further examine how this risk bound depends 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 of coefficient estimates and forecasts, respectively.

Corollary 1. *If there exist positive 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 $A = I_{Nk}$,*

$$E\{L_A(\widehat{w})\} \leq \frac{1}{1-c} \inf_{w \in \mathcal{W}} R_A(w) + \frac{1-2c}{c(1-c)} \frac{Nk\bar{\Omega}}{Tc_1} + \frac{4}{1-c} \frac{Nk\bar{\Omega}}{Tc_1}, \tag{20}$$

and when $A = X'X$,

$$E\{L_A(\widehat{w})\} \leq \frac{1}{1-c} \inf_{w \in \mathcal{W}} R_A(w) + \frac{1-2c}{c(1-c)} Nk\bar{\Omega} + \frac{4}{1-c} Nk\bar{\Omega}. \tag{21}$$

Proof*: See Section A.1 of the online appendix.*

The implied risk bounds are particularly informative, as they demonstrate how the performance of the Mallows pooling averaging estimator and forecast is determined by $\inf_{w \in \mathcal{W}} R_A(w)$ 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 $T, N \to \infty$.

C.1: $X_i'u_i = O_p(T^{1/2})$ uniformly for $i = 1, \ldots, N$.

C.2: $0 < c_1 \leq \min_{i \in \{1,\ldots,N\}} \mathcal{I}_1(T^{-1}X_i'X_i) \leq \max_{i \in \{1,\ldots,N\}} \mathcal{I}_2(T^{-1}X_i'X_i) \leq c_2 < \infty$.

C.3: $MNT^{-1/2}\xi_{NT}^{-1}\mathcal{I}_2(A) \to 0$ where $\xi_{NT} = \inf_{w \in \mathcal{W}} R_A(w)$.

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*

$$\frac{L_A(\widehat{w}^*)}{\inf_{w \in \mathcal{W}} L_A(w)} \to 1, \tag{22}$$

in probability, regardless of $\widehat{V} = \widehat{V}_{homo}$, $\widehat{V} = \widehat{V}_{bh}$ or $\widehat{V} = \widehat{V}_{ch}$.
Proof*: See Section A.1 of the online appendix.*

This theorem suggests that MPA estimator $\widehat{\beta}(\widehat{w})$ 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. 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 via bootstrap. Particularly, we implement cross-sectional resampling (B times) following Kapetanios (2008), and compute the coefficient estimates for each sample. The empirical variance and confidence interval are calculated using the B estimates. When a pre-screening step is involved (see next section), the resampled data are used in both the pre-screening and estimation step, and hence the bootstrap variance reflects the uncertainty of both pre-screening and pooling averaging.

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, creating a huge model space for model averaging and selection. In this case, a preliminary step of model screening is desirable. 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 $\mathcal{M}^s$ be a subset of $\{1, \ldots, M\}$ and $\mathcal{W}^s = \{w \in [0,1]^M : \sum_{m \in \mathcal{M}^s} w_m = 1 \text{ and } \sum_{m \notin \mathcal{M}^s} w_m = 0\}$ be a subset of $\mathcal{W}$. The model-averaging estimator based on the subset $\mathcal{M}^s$ is obtained by using the weight vector $\widehat{w}^s = \arg\min_{w \in \mathcal{W}^s} \mathcal{C}_A(w)$. We make the following assumption:

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

Under Conditions 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 $\mathcal{M}^s$ still achieves the asymptotic optimality, namely

$$\frac{L_A(\widehat{w}^s)}{\inf_{w \in \mathcal{W}} L_A(w)} \to 1.$$

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 impose equality restrictions incorrectly on far different coefficients. 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 (Lin and Ng, 2012; 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 slope coefficients are characterized by a group pattern of heterogeneity and 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 there exists a group pattern of heterogeneity and the number of groups is correctly specified or over-specified, the estimated slope coefficients are consistent, while underspecification of the number of groups leads to inconsistent estimates but gains more efficiency (Bonhomme and Manresa, 2015; Liu et al., 2018). Therefore, the consistency-efficiency trade-off remains valid for the post-screening model space. Third, it is computationally more attractive than *k-means* since it can be transformed into a sequence of convex problems and does not depend on the initial values (Su et al., 2016).

There also exist various others methods to shrink the model space, such as top m model screening and ordering model screening (e.g., Claeskens et al., 2006; Zhang et al., 2016), where the screening is mainly based on the values of information criteria. The bias-variance properties of these IC-based screening approaches are however less explicit compared to the proposed screening approach. To avoid the danger of making arguments sensitive to our choice of screening procedures, we will also consider in our Monte Carlo studies alternative methods, for example, the mixture-like iterative (M-Estimation) method proposed by Liu et al. (2018) and the agglomerative hierarchical clustering. Unreported results show that our main results are not affected.[8]

Interestingly, model averaging also offers an effective way of addressing uncertainty when determining the number of groups in 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 necessarily the most accurate in terms of minimum MSE, nor do they guarantee 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 the weights. The optimal weight choice depends on whether the focus is on parameter estimation or forecasting. In the next section, we shall

[8] Detailed studies of alternative pre-screening methods are provided in the online appendix.

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

$$y_{it} = \sum_{l=1}^{3} x_{l,it}\beta_{il} + \epsilon_{it}, \quad i = 1, \ldots, N; \quad t = 1, \ldots, T, \tag{23}$$

where $x_{it1} = 1$ and the remaining regressors are independently generated from the standard normal distributions. To mimic the empirical data of sovereign CDS spreads, we also consider regressors to follow an autoregressive process, and the details are provided in the online appendix. The idiosyncratic errors ϵ_{it} are uncorrelated with regressors and independently normally distributed with mean zero and variance $\sigma^2_{\epsilon i}$. We consider conditional heteroskedasticity, such that the variance of errors varies across individuals and its size depends on a pre-specified value of R^2. The slope coefficients are cross-sectional heterogeneous. 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):

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

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

DGP 3 (Strongly heterogeneous):

$$\beta_{i1}, \beta_{i2} = \begin{cases} 1, & i = 1, \ldots, [N/4] \\ 2, & i = [N/4] + 1, \ldots, [N/2], \\ 3, & i = [N/2] + 1, \ldots, [3N/4], \\ 4, & i = [3N/4] + 1, \ldots, N, \end{cases} \qquad \beta_{i3} = \begin{cases} 1, & i = 1, \ldots, [N/5] \\ 2, & i = [N/5] + 1, \ldots, [2N/5], \\ 3, & i = [2N/5] + 1, \ldots, [3N/5], \\ 4, & i = [3N/5] + 1, \ldots, N. \end{cases}$$

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 $T \in \{20, 40, 80\}$, leading to six 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 very similar.

We compare the forecasting performance of Mallows pooling averaging with the pooled model, Swamy's feasible generalized least squared estimator (FGLS), individual time series model, shrinkage estimator (7), a single pooling model selected by AIC or BIC, pooling averaging using relative values of AIC or BIC as weights (smoothed AIC/BIC), Bayesian pooling averaging (BPA, see online appendix for the computation), and a C-Lasso estimator with the number of groups determined by the information criterion (IC) defined by Su et al. (2016).[9] All pooling averaging and information-criterion-based forecasts are constructed from the preliminary model screening method using C-Lasso as described in Section 6.[10] 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, and report the best choice, despite high similarity of the results of using different estimators 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, so that the number of the individual forecast is always 1 and thus not reported. We emphasize that the purpose of the simulation studies is not to show the superiority of one 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. According to simulation results, we provide applied researchers with practical rules as to which methods are more likely to produce reliable results in a particular situation.

7.2 Results

We first present the results of the benchmark case when $R^2 = 0.9$, and then we consider smaller signal-to-noise ratios.

Insert TABLE 1 here

Table 1 presents the results with i.i.d regressors and a large R^2. In DGP 1 of a homogeneous

[9] The tuning parameter for C-Lasso is chosen by trying different values and selecting the best one in terms of risk. Comparison with more alternative methods can be found in the online appendix.

[10] We also compare with an M-estimation method for the latent group structure proposed by Liu et al. (2018). This method is used to directly provide a forecast and to shrink the model space as an alternative to C-Lasso. The forecast produced by M-estimation is close to that of C-Lasso, and pre-screening performance is also similar to C-Lasso. Detailed results are provided in our online appendix.

panel the pooled forecast always performs best as expected, almost identical to the FGLS forecast, both of which are based on the average-effect estimators. The proposed MPA forecast is the third best in most cases, closely following the pooled and FGLS forecasts, 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 11 out of 12 cases, while C-Lasso with the number of groups selected by the information criterion (IC) performs best in DGP 2 when $N = 30$ and $T = 40$. BPA appears to be a close competitor to MPA, especially when the time span is short. It also seems more favourable than SBIC when N is large or T is small. For the completely heterogeneous DGP 4, we find that MPA remains the best in most cases, outperforming the individual forecast. The forecast based on the shrinkage estimator performs rather well, and is the best when N is small and T is moderate or large ($T = 40$ or 80). This observation may seem counterintuitive at the first glance, since one may expect the individual forecast to 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. In contrast, 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. Hence, if the forecasting is of central interest, averaging offers a sensible alternative to handle the uncertainty in the number of groups, especially when the sample size is limited.

Insert TABLE 2 here

So far, the results are all based on DPGs with $R^2 = 0.9$. The theory in Section 4 suggests that greater noise weakens the advantages of pooling averaging estimates, but supports 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 R^2 ranging from 0.4 to 0.6, and the results are presented in Table 2. When $R^2 = 0.6$ (upper panel of Table 2), MPA remains the best in heterogeneous panels (DGP 2 to 4), but the advantages of MPA over the pooled and FGLS forecast is much smaller compared to the cases with $R^2 = 0.9$. This is because when the estimation error is more sizeable, the efficiency gains from pooling become more important. This result firmly supports the theoretical argument in Section 4. If we decrease R^2 to 0.5, we see even closer performance of the pooled, FGLS, and MPA forecast, all of which dominantly outperform other rivals. The pooled and FGLS forecast sometimes even performs the best in heterogeneous panels since the efficiency gain of the pooled forecast dominates the bias when data are highly noisy. As R^2 further decreases to 0.4, the advantage of the pooled and FGLS forecasts becomes more prominent. Furthermore, 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 BPA and C-Lasso based on a single selected number of groups becomes even more obvious.

DGPs 1–4 consider various degrees of heterogeneity by changing the number of groups. An equally important determinant of the degree of heterogeneity is the discrepancy between the coefficient values across groups. When the coefficient values get closer, the degree of heterogeneity is reduced even though the number of groups is fixed, and this also favours the pooled or FGLS methods. More detailed discussion and numerical evidence are provided in the online appendix.

In general, we find that MPA performs robustly well in panels with various degrees of heterogeneity. When the signal-to-noise ratio is moderate or high, MPA prevails. When the signal-to-noise ratio is relatively low (reflected by a small R^2 and/or a small T) and the degree of heterogeneity is weak, MPA tends to assign the most weights to the pooled model, gaining the 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 apply the proposed method to explain and forecast cross-country sovereign credit risk. Since the breakout of a wide range of financial crises, many countries have experienced a dramatic increase in government debts, which 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 future risks.

We focus on the sovereign credit default swap (CDS) spreads as a proxy of sovereign credit risks. A CDS contract is an insurance contract that protects the buyer from credit events, 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 their 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 detailed definition of 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 (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, which shows that there is indeed some common pattern in the processes of sovereign CDS spreads across countries. It thus raises a question whether a determinant has similar impacts on CDS spreads in different countries. 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. 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.[11] 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 desirable.

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

$$\Delta CDS_{it} = \alpha_i + X'_{i,t-1}\beta_i + \varepsilon_{it}, \quad i = 1, \ldots, N,\ t = 1, \ldots, T, \tag{24}$$

where ΔCDS_{it} is the first-differenced sovereign CDS spread of country i at time t, X_{it} is a 10×1 vector of covariates, and ε_{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 remove possible reverse causality from CDS spreads to macroeconomic fundamentals to a certain degree, if not completely. 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

Since our pooling averaging techniques require that the data be stationary, 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.

[11] If 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 would always outperform the one using the whole sample period (Pesaran et al., 2013). Nevertheless, completely ignoring structural breaks and using the whole sample of periods without careful analysis is not an appropriate approach.

We employ the recently developed break detection method by Baltagi et al. (2016) that explicitly allows for heterogeneous slope coefficients, and conclude that there are two structural breaks. The two estimated break points, corresponding to January 2009 and December 2009, closely match important events during the global financial crisis in 2009. Particularly, although US had already entered recession at the end of 2007, a wide range of global crises broke out in late 2008 when Lehman Brothers declared bankruptcy and a number of European countries slipped 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, a situation that lasted for roughly a whole year. The two break points result in 72 observations in the first regime, 11 in the second, and 73 in the last. Given the moderate size of samples 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. Hence, we conditional on the breaks and we apply our pooling averaging approach to each regime separately. A joint framework to deal with potential breaks and pooling averaging is considered to be a topic for future research.

8.2 Effects of local and global variables

We examine the effect of determinants on sovereign CDS spreads using Mallows pooling averaging. Existing studies on the determinants of sovereign CDS spreads typically estimate each country separately (Longstaff et al., 2011; Dieckmann and Plank, 2012). This approach cannot capture any common pattern nor the correlation between the countries, and it may be rather inefficient if some countries share similar features. Alternatively, one can also estimate the average effect of determinants across all countries (Remolona et al., 2008). Our pooling averaging approach can capture both heterogeneity and similarity across countries, and it outperforms the individual estimators and average-effect estimator when the data are characterized by heterogeneity and moderate noise. Compared to the standard approaches, the implementation of MPA requires a preliminary step of shrinking model space to facilitate computation, when the number of individual units is large. This step may introduce extra uncertainty in the estimation and forecast.

We estimate (24) using MPA, respectively, for the three regimes segmented according to the two estimated structural break points. We employ the pre-screening approach described in Section 6. We set the maximum number of groups $G_{\max}$ to be $N/2$, and 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 $\lambda = c_\lambda s_Y^2 T^{1/3}$, where $c_\lambda = \{0.0625, 0.125, 0.25, 0.5, 1\}$ is a geometrically increasing sequence and s_Y^2 is the sample variance of y_{it}. From the set of λs, we pick up the value that minimizes the risk as in the simulation.

To compute the Mallows criterion, the conditional 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 5.2. Particularly, we conduct cross-sectional resampling for B times. In each replication, we shrink the model space using C-Lasso and average estimates obtained from pre-screened models. The bootstrap standard errors can then be obtained based on these B pooling averaging estimates.

Insert TABLE 3 here

Table 3 provides the MPA estimates of regime-specific slope coefficients for individual countries. To save space, we only report the results of the seven countries whose ratio of gross government debt to GDP is more than 50% in 2016, given that the credit risk of these highly leveraged countries is often of great interest for investors and policy makers.[12] 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 also compare the MPA estimates with the pooled and individual estimates (reported in the online appendix).

We first examine the local variables. In the first regime before 2009, the effects of all local determinants are insignificant for almost all countries. Nevertheless, the impact of these determinants gets stronger after the crisis. In particular, the effect of local stock returns is strengthened and becomes strongly significant and also much more heterogeneous in the third regime. This differs from the individual estimates that report a counterintuitively positive effect of local stock returns for several countries. The local exchange rate plays an important role in the third regime in Brazil, China, and Japan, while the effect of foreign currency reserves is also strong in these countries during the (post-)crisis period. Interestingly, we find that the effect of local determinants appears stronger in bigger economies, such as Brazil and China, than in smaller ones, such as Hungary, Poland, and Slovak. This suggests that the domestic economic performance imposes a greater impact on the credit risk for bigger economies, especially during the crisis, while the risk of smaller economies are more influenced by global determinants. Such heterogeneity cannot be captured by the pooled estimates that report a homogeneously significant effect of local stock returns but insignificant overall effects of foreign exchange rates and currency reserves.

Next, we examine the global variables. We find that the US stock return is the most salient determinant before the crisis, imposing a negative impact on the credit spreads of all countries. Interestingly, during and after the crisis, this effect becomes insignificant for many countries, and some countries 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. In contrast, in the crisis period when the domestic market is volatile, the substitution effect between domestic and global markets would play a more important role,

[12] The complete estimation results are available upon request.

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. In contrast, the individual estimates report an insignificant effect of treasury yield in many countries.

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

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.[13] The forecasts are constructed using both fixed and expanding windows. 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.

Insert TABLE 4 here

Table 4 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 forecasting 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. When only the post-break sample is used, the forecast based on the pooled model is most reliable. In

[13] There exist various methods to deal with the window-selection issue in time series forecasting, see, e.g. Pesaran et al. (2013) and references therein. Optimal window selection in panel forecasting is an interesting topic that deserves future research.

this case, MPA adaptively assigns the most weights to the pooled model, and it produces equally good forecasts as the pooled model. In the case of expanding windows, 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 country. The forecasting performance of our method generally outperforms the alternatives with good robustness.

Based on our theoretical and numerical results, we conclude that the performance of different pooling estimators depends on the situation at hand. If the focus is to obtain the most accurate estimator or forecast in terms of the minimum MSE, then the choice of estimator/forecast in practice involves a trade-off between efficiency gains from pooling and bias due to heterogeneity. This trade-off is jointly determined by a number of factors, such as the signal-to-noise ratio, the degree of cross-sectional heterogeneity, the length of the times series, the number of regressors. For example, it is possible that the data are characterized by a large degree of heterogeneity but a low signal-to-noise ratio, or the other way around. Hence, the practical choice of estimator/forecast should be made by considering all these factors simultaneously. We show that the Mallows pooling averaging produces favourable and fairly robust results. In heterogeneous panels with a moderate signal-to-noise ratio, MPA often performs best. Even when the panels are homogeneous or data are highly noisy, MPA is still a reliable method whose performance is only slightly worse than the best method (pooled or FGLS estimation).

We end this paper with practical recommendations on how to determine which estimator/forecast to use in different situations. The first step is to estimate individual time series separately, and compute the coefficient estimates and (adjusted) R^2 for each regression. Despite its possible inefficiency, estimation of individual regression can be used as a starting point because the coefficient estimates are consistent. In most cases, MPA is a safe choice due to its good performance and robustness. If most individual regressions produce particularly low R^2 and the coefficient estimates vary little across individual units, the pooled or FGLS estimation could be a better approach. Although our simulation studies consider a variety of DGPs, we emphasize that these rules of thumb are concluded based on 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

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

Ando T, Bai J. 2016. Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics* **31**: 163–191.

Ando T, Li KC. 2014. A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association* **109**: 254–265.

Baltagi BH. 2008. Forecasting with panel data. *Journal of Forecasting* **27**: 153–173.

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

Baltagi BH, Feng Q, Kao C. 2016. Estimation of heterogeneous panels with structural breaks. *Journal of Econometrics* **191**: 176–195.

Baltagi BH, Griffin JM. 1997. Pooled estimators vs. their heterogeneous counterparts in the context of dynamic demand for gasoline. *Journal of Econometrics* **77**: 303–327.

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

Bonhomme S, Manresa E. 2015. Grouped patterns of heterogeneity in panel data. *Econometrica* **83**: 1147–1184.

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

Claeskens G, Croux C, Van Kerckhoven J. 2006. Variable selection for logistic regression using a prediction-focused information criterion. *Biometrics* **62**: 972–979.

Claeskens G, Magnus JR, Vasnev AL, Wang W. 2016. The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting* **32**: 754–762.

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

Dieckmann S, Plank T. 2012. Default risk of advanced economies: An empirical analysis of credit default swaps during the financial crisis. *Review of Finance* **16**: 903–934.

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

Hansen BE. 2007. Least squares model averaging. *Econometrica* **75**: 1175–1189.

Hashem PM, Zhou Q. 2018. To pool or not to pool: Revisited. *Oxford Bulletin of Economics and Statistics* **80**: 185–217.

Hoogstrate AJ, Palm FC, Pfann GA. 2000. Pooling in dynamic panel-data models: An application to forecasting GDP growth rates. *Journal of Business & Economic Statistics* **18**: 274–283.

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

Juhl T, Lugovskyy O. 2014. A test for slope heterogeneity in fixed effects models. *Econometric Reviews* **33**: 906–935.

Kapetanios G. 2008. A bootstrap procedure for panel data sets with many cross-sectional units. *Econometrics Journal* **11**: 377–395.

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

Liu R, Schick A, Shang Z, Zhang Y, Zhou Q. 2018. Identification and estimation in panel models with overspecified number of groups. working paper.

Longstaff FA, Pan J, Pedersen LH, Singleton KJ. 2011. How sovereign is sovereign credit risk? *American Economic Journal: Macroeconomics* **3**: 75–103.

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

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

Pesaran MH, Pick A, Pranovich M. 2013. Optimal forecasts in the presence of structural breaks. *Journal of Econometrics* **177**: 134–152.

Pesaran MH, Pierse RG, Kumar MS. 1989. Econometric analysis of aggregation in the context of linear prediction models. *Econometrica* **57**: 861–888.

Pesaran MH, Shin Y, Smith RP. 1999. Pooled mean group estimation of dynamic heterogeneous panels. *Journal of the American Statistical Association* **94**: 621–634.

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

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

Qian Z, Wang W, Ji K. 2017. Sovereign credit risk, macroeconomic dynamics, and financial contagion: Evidence from Japan. *Macroeconomic Dynamics* **21**: 2096–2120.

Remolona E, Scatigna M, Wu E. 2008. The dynamic pricing of sovereign risk in emerging markets: Fundamentals and risk aversion. *Journal of Fixed Income* **17**: 57–71.

Su L, Chen Q. 2013. Testing homogeneity in panel data models with interactive fixed effects. *Econometric Theory* **29**: 1079–1135.

Su L, Shi Z, Phillips PCB. 2016. Identifying latent structures in panel data. *Econometrica* **84**: 2215–2264.

Swamy PAVB. 1970. Efficient inference in a random coefficient regression model. *Econometrica* **38**: 311–323.

Timmermann A. 2006. *Forecast Combinations.* Handbook of Economic Forecasting. Elsevier, 135–196.

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

Wang W, Phillips PC, Su L. 2018. Homogeneity pursuit in panel data models: Theory and applications. *Journal of Applied Econometrics* **33**: 797–815.

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

Zhang X, Yu D, Zou G, Liang H. 2016. Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. *Journal of the American Statistical Association* **111**: 1775–1790.

Table 1: Risk comparison: Independent regressors with $R^2 = 0.9$

	DGP	MPA	BPA	C-Lasso	SAIC	SBIC	AIC	BIC	Pool	FGLS	SHK
	1	0.130	0.356	0.454	0.487	0.144	0.591	0.160	**0.111**	0.112	0.742
$N = 10$	2	**0.341**	0.370	0.604	0.444	0.429	0.477	0.474	4.342	4.429	0.955
$T = 20$	3	**0.536**	0.538	0.829	0.724	0.786	0.801	0.830	2.640	2.695	0.934
	4	**0.918**	1.213	1.905	1.192	1.386	1.265	1.521	14.53	14.82	0.984
	1	0.175	0.197	0.314	0.364	0.265	0.408	0.301	**0.034**	0.037	0.701
$N = 30$	2	**0.158**	0.159	0.194	0.247	0.204	0.259	0.214	4.148	4.192	0.943
$T = 20$	3	**0.376**	0.396	0.655	0.509	0.536	0.524	0.567	2.317	2.353	0.911
	4	**0.396**	0.670	0.649	0.625	0.649	0.644	0.649	1.733	1.776	0.891
	1	0.131	0.360	0.466	0.488	0.127	0.579	0.133	**0.094**	0.098	0.876
$N = 10$	2	**0.268**	0.410	0.300	0.458	0.284	0.501	0.297	8.852	8.932	0.988
$T = 40$	3	**0.657**	0.681	1.264	0.897	1.072	1.000	1.174	5.153	5.209	0.981
	4	1.264	1.854	2.679	1.757	1.884	1.789	2.062	29.94	30.23	**0.996**
	1	0.177	0.203	0.319	0.373	0.248	0.420	0.288	**0.033**	0.034	0.854
$N = 30$	2	0.105	0.141	**0.089**	0.273	0.164	0.282	0.181	8.257	8.290	0.985
$T = 40$	3	**0.490**	0.580	1.013	0.673	0.754	0.721	0.779	4.596	4.623	0.975
	4	**0.651**	1.157	1.070	0.882	1.006	0.923	1.024	3.503	3.538	0.969
	1	0.125	0.366	0.642	0.465	0.105	0.567	0.107	**0.096**	0.098	0.938
$N = 10$	2	**0.311**	0.436	0.901	0.632	0.344	0.700	0.351	17.47	17.56	0.997
$T = 80$	3	**0.959**	1.029	1.846	1.215	1.570	1.292	1.740	10.14	10.18	0.995
	4	2.223	3.644	7.613	3.252	3.253	3.300	3.376	59.96	60.23	**0.999**
	1	0.220	0.250	0.488	0.569	0.216	0.641	0.258	**0.033**	0.034	0.928
$N = 30$	2	**0.184**	0.240	0.793	0.507	0.311	0.524	0.343	16.55	16.58	0.996
$T = 80$	3	**0.723**	0.868	1.434	1.007	0.970	1.025	1.012	9.296	9.317	0.993
	4	**0.686**	0.910	1.298	1.007	0.957	1.026	1.030	7.164	7.187	0.991

Notes:

1. Forecasts constructed using: MPA: Mallows pooling averaging estimator; BPA: Bayesian pooling averaging; C-Lasso: C-Lasso estimator with the number of groups determined by IC; 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; FGLS: Swamy's estimator; SHK: shrinkage estimator.

2. All numbers are divided by the risk of the individual time series forecast.

Table 2: Risk comparison under larger noise

	DGP	MPA	BPA	C-Lasso	SAIC	SBIC	AIC	BIC	Pool	FGLS	SHK
						$R^2 = 0.6$					
	1	0.130	0.372	0.649	0.488	0.125	0.585	0.131	**0.094**	0.098	0.876
$N = 10$	2	**0.330**	0.672	0.768	0.697	0.411	0.790	0.465	0.348	0.339	0.899
$T = 40$	3	**0.245**	0.683	0.753	0.661	0.314	0.746	0.342	0.251	0.248	0.888
	4	**0.596**	0.862	0.901	0.848	0.833	0.936	0.912	0.930	0.944	0.931
	1	0.124	0.365	0.641	0.466	0.104	0.566	0.106	**0.096**	0.097	0.938
$N = 10$	2	**0.487**	0.684	0.831	0.770	0.598	0.863	0.666	0.582	0.589	0.960
$T = 80$	3	**0.366**	0.702	0.791	0.716	0.430	0.806	0.471	0.384	0.383	0.951
	4	**0.673**	0.866	0.982	0.901	0.974	0.998	1.039	1.716	1.734	0.977
						$R^2 = 0.5$					
	1	0.128	0.381	0.454	0.506	0.124	0.608	0.133	**0.094**	0.098	0.876
$N = 10$	2	0.208	0.248	0.304	0.291	0.216	0.335	0.219	0.213	**0.185**	0.885
$T = 40$	3	0.164	0.241	0.353	0.322	0.180	0.363	0.184	0.163	**0.128**	0.879
	4	**0.386**	0.406	0.610	0.599	0.550	0.634	0.604	0.470	0.463	0.905
	1	0.124	0.365	0.641	0.466	0.104	0.566	0.106	**0.096**	0.097	0.938
$N = 10$	2	**0.306**	0.604	0.758	0.679	0.361	0.770	0.399	0.314	0.308	0.949
$T = 80$	3	**0.201**	0.617	0.746	0.642	0.276	0.730	0.296	0.227	0.206	0.944
	4	**0.586**	0.866	0.907	0.847	0.796	0.935	0.875	0.818	0.826	0.964
						$R^2 = 0.4$					
	1	0.130	0.371	0.649	0.488	0.125	0.585	0.131	**0.094**	0.098	0.876
$N = 10$	2	0.161	0.539	0.715	0.597	0.204	0.689	0.224	0.153	**0.113**	0.877
$T = 40$	3	0.137	0.562	0.714	0.592	0.195	0.681	0.209	0.128	**0.085**	0.874
	4	0.265	0.824	0.805	0.696	0.401	0.771	0.442	0.265	**0.238**	0.885
	1	0.124	0.364	0.641	0.466	0.104	0.566	0.106	**0.096**	0.097	0.938
$N = 10$	2	0.200	0.525	0.718	0.613	0.232	0.710	0.252	0.195	**0.165**	0.943
$T = 80$	3	0.165	0.547	0.717	0.602	0.202	0.687	0.216	0.157	**0.121**	0.939
	4	**0.394**	0.852	0.835	0.761	0.516	0.844	0.565	0.418	0.409	0.952

Notes: See footnote of Table 1.

Table 3: Effects of CDS spreads determinants for selected countries: Mallows pooling averaging estimates

	Brazil			China			Hungary			Japan		
	Reg. 1	Reg. 2	Reg. 3	Reg. 1	Reg. 2	Reg. 3	Reg. 1	Reg. 2	Reg. 3	Reg. 1	Reg. 2	Reg. 3
lstock	−0.047	−1.772	−0.243	0.059	−4.202	−0.161	0.059	−0.391	−0.039	−0.065	1.930	−0.140
	(0.062)	(0.847)	(0.039)	(0.090)	(1.887)	(0.044)	(0.108)	(1.560)	(0.043)	(0.124)	(1.666)	(0.039)
fxrates	0.044	−2.213	−0.119	0.156	30.182	−0.018	0.156	−0.383	−0.026	0.004	−2.910	0.105
	(0.058)	(2.841)	(0.034)	(0.086)	(13.591)	(0.038)	(0.102)	(10.460)	(0.041)	(0.110)	(7.956)	(0.043)
fxres	−0.083	−0.942	−0.183	0.030	7.693	0.108	0.030	0.185	0.082	−0.037	−4.242	0.504
	(0.060)	(0.951)	(0.059)	(0.052)	(3.509)	(0.082)	(0.050)	(2.807)	(0.090)	(0.054)	(3.062)	(0.091)
gstock	−0.212	0.582	0.277	−0.292	3.001	0.152	−0.292	0.630	0.066	−0.101	−1.393	0.235
	(0.044)	(0.285)	(0.046)	(0.127)	(1.415)	(0.051)	(0.132)	(1.291)	(0.052)	(0.130)	(1.438)	(0.052)
trsy	−0.078	−0.629	−0.158	−0.042	−0.493	−0.143	−0.042	−0.343	−0.176	−0.011	−1.048	−0.009
	(0.030)	(0.229)	(0.023)	(0.036)	(0.314)	(0.020)	(0.039)	(0.342)	(0.018)	(0.038)	(0.499)	(0.018)
hy	−0.178	0.609	−0.062	−0.183	−2.297	0.088	−0.183	0.299	0.240	−0.144	0.816	0.365
	(0.029)	(0.385)	(0.042)	(0.042)	(1.327)	(0.040)	(0.040)	(1.114)	(0.039)	(0.042)	(1.164)	(0.037)
eqp	0.138	0.739	0.060	0.121	−0.100	−0.041	0.121	0.658	0.116	0.300	−0.010	−0.075
	(0.054)	(0.163)	(0.038)	(0.151)	(0.424)	(0.041)	(0.152)	(0.392)	(0.043)	(0.146)	(0.455)	(0.043)
volp	0.121	1.713	−0.071	0.118	43.058	−0.147	0.118	7.639	−0.105	0.116	−5.974	−0.273
	(0.029)	(4.642)	(0.026)	(0.027)	(16.678)	(0.025)	(0.024)	(13.175)	(0.024)	(0.026)	(11.878)	(0.021)
ef	−0.032	0.625	−0.119	−0.137	−5.845	−0.247	−0.137	1.018	−0.189	−0.078	−2.065	−0.219
	(0.061)	(0.615)	(0.020)	(0.042)	(3.188)	(0.020)	(0.041)	(2.592)	(0.020)	(0.035)	(2.416)	(0.020)
bf	0.110	0.362	−0.058	0.092	−3.561	−0.095	0.092	0.378	−0.040	0.079	−0.969	−0.139
	(0.057)	(0.345)	(0.017)	(0.063)	(1.758)	(0.017)	(0.077)	(1.368)	(0.017)	(0.086)	(1.146)	(0.017)

Notes: Bootstrap standard errors given in the parentheses.

Table 3 (con't): Effects of CDS spreads determinants for selected countries: Mallows pooling averaging estimates

	Malaysia			Poland			Slovak		
	Reg. 1	Reg. 2	Reg. 3	Reg. 1	Reg. 2	Reg. 3	Reg. 1	Reg. 2	Reg. 3
lstock	0.145	−0.391	−0.300	−0.028	−0.391	−0.192	−0.047	−0.391	0.242
	(0.121)	(1.165)	(0.037)	(0.100)	(1.275)	(0.038)	(0.064)	(1.556)	(0.039)
fxrates	0.210	−0.383	−0.007	0.067	−0.383	−0.090	0.044	−0.383	−0.016
	(0.104)	(3.275)	(0.044)	(0.097)	(3.308)	(0.041)	(0.096)	(1.022)	(0.038)
fxres	0.039	0.185	0.091	−0.036	0.185	0.202	−0.083	0.185	−0.068
	(0.107)	(1.967)	(0.097)	(0.356)	(2.320)	(0.091)	(0.475)	(2.059)	(0.074)
gstock	−0.307	0.630	0.462	−0.209	0.630	0.245	−0.212	0.630	0.316
	(0.109)	(1.071)	(0.051)	(0.075)	(0.833)	(0.048)	(0.080)	(0.654)	(0.054)
trsy	−0.033	−0.343	−0.124	−0.056	−0.343	−0.182	−0.078	−0.343	−0.342
	(0.033)	(0.405)	(0.019)	(0.040)	(0.609)	(0.021)	(0.063)	(1.224)	(0.024)
hy	−0.193	0.299	0.265	−0.170	0.299	0.144	−0.178	0.299	0.273
	(0.040)	(1.055)	(0.036)	(0.034)	(0.988)	(0.037)	(0.032)	(0.849)	(0.045)
eqp	0.111	0.658	0.240	0.191	0.658	0.168	0.138	0.658	0.269
	(0.127)	(0.349)	(0.046)	(0.102)	(0.338)	(0.045)	(0.114)	(0.435)	(0.045)
volp	0.121	7.639	−0.165	0.119	7.639	−0.177	0.121	7.639	−0.160
	(0.031)	(5.933)	(0.020)	(0.030)	(5.361)	(0.019)	(0.025)	(3.588)	(0.022)
ef	−0.146	1.018	−0.229	−0.082	1.018	−0.147	−0.032	1.018	−0.203
	(0.035)	(1.583)	(0.020)	(0.037)	(1.343)	(0.019)	(0.051)	(0.714)	(0.019)
bf	0.157	0.378	−0.153	0.064	0.378	−0.128	0.110	0.378	−0.243
	(0.085)	(0.562)	(0.017)	(0.071)	(0.471)	(0.018)	(0.061)	(0.173)	(0.019)

Notes: Bootstrap standard errors given in the parentheses.

Table 4: Out-of-sample forecasting comparison with fixed window

	Full sample			*Post-first-break sample*			*Post-second-break sample*		
$1-\tau\%$	0.1	0.05	0.01	0.1	0.05	0.01	0.1	0.05	0.01
MPA	**0.868**	**0.890**	1.020	1.469	0.923	0.476	**1.578**	1.248	0.571
BPA	0.891	0.896	1.027	1.516	**0.915**	0.505	1.624	1.257	0.726
C-Lasso	0.875	0.892	1.024	1.469	0.924	0.476	1.580	1.248	0.571
SAIC	0.870	0.891	1.022	1.470	0.924	0.477	1.582	1.264	0.730
SBIC	0.875	0.892	1.024	1.469	0.924	0.476	1.580	1.248	0.571
AIC	0.875	0.892	1.024	1.469	0.924	0.476	1.580	1.271	0.786
BIC	0.875	0.892	1.024	1.469	0.924	0.476	1.580	1.248	0.571
Pool	0.875	0.892	1.024	1.469	0.924	0.476	1.580	1.248	0.571
FGLS	0.937	0.925	1.006	1.467	1.033	0.492	1.676	1.335	0.724
SHK	0.980	0.982	**0.992**	1.552	0.967	0.508	1.636	1.273	0.687
Indiv	1.000	1.000	1.000	1.622	1.012	0.562	1.715	1.324	0.763
R^2	0.098	0.095	0.094	0.176	0.167	0.161	0.208	0.198	0.181

Notes:

1. τ denotes percentage of sample used for parameter estimation. Abbreviations explained in footnote 1 of Table1.
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.