Forecasting with panel data: estimation uncertainty versus parameter heterogeneity*

M. Hashem Pesaran† Andreas Pick‡ Allan Timmermann§

February 21, 2022

Abstract

We develop forecasting methods for panel data with heterogeneous parameters and conduct a comprehensive comparison of their predictive accuracy in settings with different cross-sectional (N) and time (T) dimensions and varying degrees of parameter heterogeneity. We investigate conditions under which panel forecasting methods can perform better than unit-specific individual forecasts and demonstrate how gains in predictive accuracy depend on the degree of parameter heterogeneity, whether heterogeneity is correlated with the regressors, the goodness of fit of the model, and, particularly, the time dimension of the data set. We propose optimal combination weights for pooled and individual forecasts and a forecast poolability test that can be used as a pretesting tool. Through a set of Monte Carlo simulations and three empirical applications to house prices, CPI inflation, and stock returns, we show that no single forecasting approach dominates uniformly. However, forecast combination and shrinkage methods provide better overall forecasting performance and offer more attractive risk profiles compared to individual, pooled, and random effects methods.

JEL codes: C33, C53
Keywords: Forecasting, Panel data, Heterogeneity, Forecast evaluation, Forecast combination, Shrinkage, Pooling.

*We would like to thank Ron Smith, Cynthia Yang and Liying Yang for helpful comments.

†University of Southern California and Trinity College, Cambridge. Email: pesaran@usc.edu

‡Erasmus University Rotterdam, Erasmus School of Economics, Burgemeester Oudlaan 50, 3000DR Rotterdam, Tinbergen Institute, and De Nederlandsche Bank. Email: andreas.pick@cantab.net

§UC San Diego, Rady School of Management, 9500 Gilman Drive, La Jolla CA 92093-0553. Email: atimmermann@ucsd.edu.

1 Introduction

Panel data are widely available at the level of individuals, firms and industries, as well as at country and regional granularities and have been extensively used for estimation and inference. Yet, panel data methods have had little impact on common practices in economic forecasting, which remain dominated by single-equation forecasting models or low-dimensional multivariate models such as vector autoregressions. The relative shortage of panel applications in the economic forecasting literature is, in part, a result of the paucity of studies on forecasting techniques for panel data and the absence of guidelines on which methods work well in different settings.

In this paper, we examine existing approaches and develop novel forecasting methods for panel data with heterogeneous parameters and conduct a systematic comparison of their predictive accuracy in settings with different cross-sectional (N) and time (T) dimensions and varying degrees of parameter heterogeneity. Our analysis provides a deeper understanding of the determinants of the performance of these methods in different settings. This includes the important choice of whether to use pooled versus individual estimates, with a focus on forecasting rather than parameter estimation and inference.

We begin by exploring the bias-variance trade-off between individual and pooled estimation when the target variables of interest are individual forecasts. We then develop a novel test that compares forecasts based on individual versus pooled estimation using a mean squared forecast error (MSFE) loss function. Our test does not address whether parameter heterogeneity is significant but instead examines whether forecasts from a model with pooled parameters are expected to be significantly more accurate than forecasts from individual estimation. The proposed test of forecast poolability accounts for the effect of predictor and parameter heterogeneity on forecast errors.

The literature on pre-testing stresses that combining forecasts can be a valid alternative to pre-testing. We therefore also consider combining the pooled and individual forecasts and develop a novel bias-adjusted combination scheme that minimizes the expected square forecast error. Our combination weights are related to shrinkage forecasts along the lines of Lee and Griffith (1979) and Maddala et al. (1997). Unlike the shrinkage forecasts, however, our combination forecasts do not rely on iteratively estimated parameters, which makes them computationally considerably less burdensome.

We apply the alternative panel forecasting methods to three empirical

applications selected to represent varying degrees of heterogeneity and predictive power of the underlying forecasting models. We characterize the center of the cross-sectional loss distribution of the forecasts through the median of the ratio of their MSFE values divided by the MSFE of the unit-specific benchmark. We also study the tail features of the loss distributions through the proportion of units for which the predictive accuracy of each approach is either best or worst.

Our first application considers predictability of house prices across 362 US metropolitan statistical areas (MSAs). The forecasting models for this application have a high R^2 value above 0.8. The pooled model's forecasts succeed in reducing the median MSFE value by up to 3% relative to the individual forecasts. Combination forecasts work even better in this application, beating the individual forecasts for more than 90% of MSAs. Overall, shrinkage forecasts produce the most accurate forecasts, reducing the median MSFE value by up to 5% relative to the individual forecasts. Among the shrinkage forecasts, a Bayesian scheme works particularly well.

Our second application considers forecasts for a panel containing 202 subcategories of CPI inflation. Our forecasting models for this application have a substantially lower R^2 in the range of 0.1 to 0.3. In this application, forecasts from the individual regressions are more accurate than those from the pooled models. Combination forecasts are, however, even more accurate than either of these methods and beat the benchmark individual forecasts for up to 96% of the series. They produce the largest MSFE for at most 0.5% of the series and have the smallest MSFE for up to 32% of series. The shrinkage forecasts also perform quite well, with a Bayes method never producing the largest MSFE for any of the variables while generating the most accurate forecasts for up to 14% of the price categories.

Finally, we examine forecasts of monthly stock returns on more than 23,000 individual firms. Stock returns are well known to be very difficult to forecast, so this application represents an environment with an extremely low predictive R^2, less than 0.01 for many stocks. Given this low predictive power, forecast bias is of secondary importance compared to estimation error variance and so it is not surprising that pooling produces more accurate forecasts than the individual models. Pre-test forecasts, Bayesian and empirical Bayes shrinkage forecasts are among the small set of models capable of producing lower average MSFE values for a majority of stocks than a simple prevailing mean forecast for these series.

Overall, forecasts that use only the information on a given unit tend to have loss distributions with wide dispersions across units. Their associated forecasts are therefore often the best but also often the worst, and their

MSFE values are often shifted to the right, implying slightly larger losses on average. Shrinkage forecasts tend to have much narrower MSFE distributions across units, often shifted to the left as they are centered around a smaller average loss. Combination methods produce MSFE distributions that are narrower still and rarely produce the largest squared forecast error among all methods that we consider. Finally, pooled forecasts often fall in the middle as their loss distributions are narrower than the individual forecasts, but wider than those from the combination forecasts.

The literature on forecasting with panel data has mainly focussed on panel data models developed for inference rather than forecasting, but there are some exceptions. Most notably, the review articles by Baltagi (2008, 2013) consider the forecasting performance of the best linear unbiased predictor (BLUP) of Goldberger (1962) in models with either fixed effects or random effects. The BLUP estimator gives rise to a generalized least squares (GLS) predictor which Baltagi compares to models that allow for autoregressive-moving average (ARMA) dynamics in innovations as well as models with spatial dependencies in the errors. In our empirical applications, we also consider the random effects BLUP and find that the combination, pre-test, and shrinkage forecasts provide more precise forecasts.

Trapani and Urga (2009) use Monte Carlo simulations to assess the forecasting performance of pooled, individual, and shrinkage estimators and find that the degree of heterogeneity is the most important determinant of the accuracy of different forecasts. Brückner and Siliverstovs (2006) have assessed a similar group of methods for the prediction of migration, where fixed effects and shrinkage estimators perform best.

Wang et al. (2019) also propose forecast combination methods. However, their combination weights are determined from in-sample test statistics rather than the expected out-of-sample performance that we propose. In this sense, our test is closer related to the forecast based test for a structural break of Boot and Pick (2020), where the target is also significant improvements in forecast accuracy rather than a significant change in parameters.

Liu, Moon and Schorfheide (2020) consider forecasting in dynamic panel data models with very short time dimension, so individual parameters cannot be estimated for each individual separately. Like Lee and Griffith (1979), they adopt a Bayesian approach to shrink the heterogeneous parameters to their mean. Their approach is complementary to ours as we assume that enough observations are available per unit to allow the individual parameters to be estimated.

The outline of the rest of the paper is as follows. Section 2 introduces the model setup and presents theoretical results that compare the predictive

accuracy of individual and pooled estimation along with that of forecast combination, pre-test and shrinkage methods. Section 3 presents a set of Monte Carlo experiments designed to shed light on the determinants of the (relative) forecasting performance of the methods introduced in Section 2. Section 4 presents the results for our three empirical applications, while Section 5 concludes. Technical details are provided in appendices at the end of the paper.

Notation: We denote the largest and smallest eigenvalues of the $N \times N$ matrix $\boldsymbol{A} = (a_{ij})$ by $\lambda_{\max}(\boldsymbol{A})$ and $\lambda_{\min}(\boldsymbol{A})$, respectively, its trace by $\operatorname{tr}(\boldsymbol{A}) = \sum_{i=1}^{N} a_{ii}$, its spectral radius by $\rho(\boldsymbol{A}) = |\lambda_{\max}(\boldsymbol{A})|$, and its spectral norm by $\|\boldsymbol{A}\| = \lambda_{\max}^{1/2}(\boldsymbol{A}'\boldsymbol{A})$. Furthermore, $\xrightarrow{p}$ denotes convergence in probability, $\xrightarrow{d}$ convergence in distribution, and $\overset{a}{\sim}$ asymptotic equivalence in distribution. $O(.)$ and $o(.)$ denote the Big O and Little o notations, respectively. If $\{f_N\}_{N=1}^{\infty}$ is any real sequence and $\{g_N\}_{N=1}^{\infty}$ is a sequence of positive real numbers, then $f_N = O(g_N)$ if there exists a positive finite constant C such that $|f_N|/g_N \leq C$ for all N. $f_N = o(g_N)$ if $f_N/g_N \to 0$ as $N \to \infty$. $O_p(.)$ and $o_p(.)$ are the equivalent orders in probability. C and c will be used to denote, respectively, finite large and non-zero small positive numbers that do not depend on N and T.

2 Theoretical results

We begin our analysis by describing the panel regression setup used in our analysis and by characterizing theoretically the trade-offs for some popular panel regression techniques. We then introduce the forecasting methods considered in the paper.

2.1 Setup and assumptions

We consider the following linear panel regression model:

$$y_{it} = \boldsymbol{\beta}_i' \boldsymbol{x}_{it} + \varepsilon_{it}, \tag{1}$$

where $i = 1, 2, \ldots, N$ refers to the individual units and $t = 1, 2, \ldots, T$ refers to the time period, y_{it} is the outcome of individual i at time t, $\boldsymbol{x}_{it}$ is a $K \times 1$ vector of regressors—or predictors—used to forecast y_{it}, $\boldsymbol{\beta}_i$, is the associated vector of regression coefficients, and ε_{it} is the disturbances of unit i in period t.[1] Stacking the time series of outcomes, regressors and dis-

[1] Note that $\boldsymbol{x}_{it}$ is assumed to be known at the point in time where the forecast of y_{it} is generated. To keep notations simple, we do not explicitly specify the forecast horizon.

turbances, define $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iT})'$, $\boldsymbol{X}_i = (\boldsymbol{x}'_{i1}, \boldsymbol{x}'_{i2,} \ldots, \boldsymbol{x}'_{iT})'$, and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT})'$. Further, let $\boldsymbol{y} = (\boldsymbol{y}'_1, \boldsymbol{y}'_2, \ldots, \boldsymbol{y}'_N)'$, $\boldsymbol{X} = (\boldsymbol{X}'_1, \boldsymbol{X}'_2, \ldots, \boldsymbol{X}'_N)'$, and $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}'_1, \boldsymbol{\varepsilon}'_2, \ldots, \boldsymbol{\varepsilon}'_N)'$

Our theoretical analysis makes the following standard assumptions about the underlying data generating process.

Assumption 1. *$\boldsymbol{\varepsilon}_i \sim \text{iid}(\boldsymbol{0}, \sigma_i^2 \boldsymbol{I}_T)$, with $0 < \sigma_i^2 < \infty$, and $\text{E}(\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}'_j) = \boldsymbol{0}$, $\forall$ $i \neq j$.*

Assumption 2. *$\text{E}(\boldsymbol{X}'_j \boldsymbol{\varepsilon}_i) = \boldsymbol{0}$, $\forall i, j$.*

Assumption 3. *The sample covariance matrices, $\boldsymbol{Q}_{iT} = T^{-1} \boldsymbol{X}'_i \boldsymbol{X}_i = O_p(1)$, $\boldsymbol{Q}_{NT} = T^{-1} N^{-1} \boldsymbol{X}' \boldsymbol{X} = O_p(1)$, and there exists a T_0 such that for all $T > T_0$, $\boldsymbol{Q}_{iT}$ and $\boldsymbol{Q}_{NT}$ are positive definite for all i, N, and T, and*

$$0 < c < inf_i \left[\lambda_{min} \left(\boldsymbol{Q}_{iT} \right) \right] < sup_i \lambda_{max} \left(\boldsymbol{Q}_{iT} \right) < C < \infty,$$

for some positive constants $c < C$.

Assumption 4. *$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\eta}_i$ with $\|\boldsymbol{\beta}\| < C$, $\text{E}(\boldsymbol{\eta}_i) = 0$ and $\text{Var}(\boldsymbol{\eta}_i) = \boldsymbol{\Omega}_{ii} = \boldsymbol{\Omega}_\eta$, $\forall i$.*

Assumption 5. *$\boldsymbol{\eta}_i$ is distributed independently of $\boldsymbol{X}_j$ and $\boldsymbol{\varepsilon}_j$, $\forall i, j$.*

Assumption 6. *$\boldsymbol{\eta}_i$ and $\boldsymbol{\eta}_j$ are weakly correlated such that $\sup_i \sum_{j=1}^N \|\boldsymbol{\Omega}_{ij}\| < C$, where $\text{E}\left(\boldsymbol{\eta}_i \boldsymbol{\eta}'_j\right) = \boldsymbol{\Omega}_{ij}$.*

This list of assumptions provides a set of sufficient conditions for establishing the results we present below. In fact, many of the assumptions can be relaxed at the cost of adding more complexity to the analysis. Assumption 1 rules out error serial correlation and imposes error cross-sectional independence. Assumption 2 requires all regressors to be strictly exogenous with respect to $\boldsymbol{\varepsilon}$. This assumption is used for forecasts based on pooled estimates and is not required when forecasts are based on the individual estimates. Assumption 3 is an identification assumption that allows consistent estimation of individual slope coefficients, $\boldsymbol{\beta}_i$, and pooled estimators of $\text{E}(\boldsymbol{\beta}_i) = \boldsymbol{\beta}$. For pooled estimation of $\boldsymbol{\beta}$, the conditions on $\boldsymbol{Q}_{iT}$ can be relaxed and it is only required that $\boldsymbol{Q}_{NT}$ is positive definite, and

$$\sup_i \text{E} \left\| \boldsymbol{Q}_{NT}^{-1} \boldsymbol{Q}_{iT} \right\| < C.$$

Assumption 4 does not rule out correlated heterogeneity and allows for non-zero values of $\text{E}(\boldsymbol{X}'_i \boldsymbol{X}_i \boldsymbol{\eta}_i)$. Correlated heterogeneity is ruled out under

Assumption 5. As we shall see later, optimality of forecasts based on pooled estimates of $\boldsymbol{\beta}$ requires Assumption 5, but this assumption is not needed for the optimality of forecasts based on the individual estimates of $\boldsymbol{\beta}_i$. Assumption 6 allows $\boldsymbol{\beta}_i$ to be cross-sectionally weakly correlated, which represents a useful generalization of the random coefficient model where it is routinely assumed that $\boldsymbol{\beta}_i$ are independent draws from the same distribution. This relaxation of the random coefficient model is relevant to pooled estimation but does not bear on individual estimates of $\boldsymbol{\beta}_i$, where $\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i$ does not depend on $\boldsymbol{\eta}_i$. Finally, $\boldsymbol{\Omega}_\eta$ is not required to be invertible. which allows for a subset of parameters to be homogeneous across i. Fixed effects can also be included in (1) by setting the first element of $\boldsymbol{x}_{it}$ to unity.

2.2 Forecasts based on individual and pooled estimates

Suppose we are interested in forecasting $y_{i,T+1}$ conditional on information known at time T, which we denote by $\boldsymbol{x}_{i,T+1}$ to clarify the correspondence to $y_{i,T+1}$. We first consider two forecasts generated from individual and pooled estimators of $\boldsymbol{\beta}_i$. Forecasts based on the individual estimators take the form

$$\hat{y}_{i,T+1} = \hat{\boldsymbol{\beta}}_i' \boldsymbol{x}_{i,T+1} \tag{2}$$

where

$$\hat{\boldsymbol{\beta}}_i = (\boldsymbol{X}_i' \boldsymbol{X}_i)^{-1} \boldsymbol{X}_i' \boldsymbol{y}_i \tag{3}$$

Similarly, forecasts based on the pooled estimator are given by

$$\tilde{y}_{i,T+1} = \tilde{\boldsymbol{\beta}}' \boldsymbol{x}_{i,T+1} \tag{4}$$

where

$$\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}' \boldsymbol{X})^{-1} \boldsymbol{X}' \boldsymbol{y} \tag{5}$$

with $\boldsymbol{X}$ and $\boldsymbol{y}$ as defined above.[2]

Forecast errors from the two approaches are given by

$$\hat{e}_{i,T+1} = y_{iT+1} - \hat{y}_{i,T+1} = \varepsilon_{i,T+1} - (\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)' \boldsymbol{x}_{i,T+1} \tag{6}$$

$$\tilde{e}_{i,T+1} = y_{iT+1} - \tilde{y}_{i,T+1} = \varepsilon_{i,T+1} - (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_i)' \boldsymbol{x}_{i,T+1} \tag{7}$$

[2] One could also consider fixed and random effects estimators by separating the intercept terms from the slope coefficients and assuming that the slopes are homogeneous. This would be an example of sub-set homogeneity. Here, we do not consider such sub-set homogeneity to simplify the derivations but we do consider random effects estimators in our Monte Carlo analysis and applications.

Both forecasts are unbiased under Assumptions 1–6, so the mean squared forecast error (MSFE) equals the variance of each forecast. This allows us to establish the following proposition, which is established in Appendix A.1:

Proposition 1. *(i) Under Assumptions 1–4, the MSFE resulting from individual-specific estimation of the parameters is*

$$\begin{aligned}\text{Var}(\hat{e}_{i,T+1}|\boldsymbol{X}_i,\boldsymbol{x}_{i,T+1}) &= \sigma_i^2 + T^{-1}\sigma_i^2\boldsymbol{x}_{i,T+1}'\boldsymbol{Q}_{iT}^{-1}\boldsymbol{x}_{i,T+1} \\ &= \sigma_i^2 + O_p\left(T^{-1}\right),\end{aligned} \tag{8}$$

where $\boldsymbol{Q}_{iT} = T^{-1}\boldsymbol{X}_i'\boldsymbol{X}_i$.
(ii) Under Assumptions 1–6, the MSFE resulting from pooled estimation of the parameters is

$$\text{Var}(\tilde{e}_{i,T+1}|\boldsymbol{X}_i,\boldsymbol{x}_{i,T+1}) = \sigma_i^2 + \boldsymbol{x}_{i,T+1}'\boldsymbol{\Omega}_\eta\boldsymbol{x}_{i,T+1} + O_p\left(N^{-1}\right) \tag{9}$$

The following list of remarks helps interpret these results:

Remark 1 For typical panel data sets, T is not large. In practice, the parameter estimation uncertainty captured by the $O_p\left(T^{-1}\right)$ term in (8) can therefore be important. Parameter heterogeneity, in contrast, does not affect the accuracy of the forecast in (8).

Remark 2 A comparison of (8) and (9) suggests that for large T, forecasts based on individual estimation will have a lower MSFE than forecasts based on pooled estimation.

Remark 3 Since estimates of $\boldsymbol{\beta}_i$ are not the focus of our analysis, the inverse of $\boldsymbol{X}_i'\boldsymbol{X}_i$ can be replaced by a generalized inverse whenever some of the eigenvalues of $\boldsymbol{X}_i'\boldsymbol{X}_i$ are close to zero.

Remark 4 Forecasts based on individual estimates have optimality properties even if one or more of the predictors in $\boldsymbol{x}_{it}$ are weakly exogenous. In contrast, forecasts based on pooled regressions require the stronger assumption of strict exogeneity.

Remark 5 Individual estimates of $\boldsymbol{\beta}_i$ are not affected by parameter heterogeneity even if such heterogeneity is correlated with the predictors, $\boldsymbol{x}_{it}$, and under Assumptions 1-3 we have $\text{E}\left(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\right) = 0$. However, the same is not true of the pooled estimates. Using (5), we have that

$$\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(N^{-1}\sum_{i=1}^{N}\boldsymbol{X}_i'\boldsymbol{X}_i\right)^{-1} N^{-1}\sum_{i=1}^{N}\boldsymbol{X}_i'\boldsymbol{X}_i\boldsymbol{\eta}_i + (N^{-1}\boldsymbol{X}'\boldsymbol{X})^{-1}N^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon},$$

and under Assumptions 1–4, we have

$$\operatorname*{plim}_{N\to\infty}\left(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)=\boldsymbol{\Psi}^{-1}\boldsymbol{b}, \tag{10}$$

where

$$\boldsymbol{\Psi}=\lim_{N\to\infty}N^{-1}\sum_{i=1}^{N}\mathrm{E}\left(\boldsymbol{X}_i'\boldsymbol{X}_i\right),\quad \boldsymbol{b}=\lim_{N\to\infty}N^{-1}\sum_{i=1}^{N}\mathrm{E}\left(\boldsymbol{X}_i'\boldsymbol{X}_i\boldsymbol{\eta}_i\right).$$

This result holds for any fixed T so long as $\boldsymbol{\Psi}$ is non-singular. For the pooled estimator to be unbiased, Assumptions 4–6 are also needed to ensure that $\mathrm{E}\left(\boldsymbol{X}_i'\boldsymbol{X}_i\boldsymbol{\eta}_i\right)=\mathbf{0}$. Therefore, parameter heterogeneity could be particularly problematic if $\boldsymbol{\eta}_i=\boldsymbol{\beta}_i-\mathrm{E}(\boldsymbol{\beta}_i)$ and $\boldsymbol{X}_i'\boldsymbol{X}_i$ are correlated; see also Pesaran (2015, Section 28.3). Further, (8) holds even if $\boldsymbol{X}_i'\boldsymbol{X}_i$ and $\boldsymbol{\eta}_i$ are correlated, but the result in (9) requires that $\boldsymbol{X}_i$ and $\boldsymbol{\eta}_i$ are distributed independently, thus ruling out correlated heterogeneity as well as weak regressor exogeneity.

Remark 6 From (8) and (9) it is clear that larger values of $\|\boldsymbol{\Omega}_\eta\|$, corresponding to a greater degree of heterogeneity in regression parameters, are associated with a relative deterioration of the expected accuracy of forecasts based on pooled estimation. However, forecast accuracy over the cross-section units depends on $\boldsymbol{x}_{i,T+1}'\boldsymbol{\Omega}_\eta\boldsymbol{x}_{i,T+1}$ and therefore also on the magnitude and the dispersion of the predictors across the units and forecast periods.

2.3 Optimal forecast combinations

Given the MSFE trade-off associated with the forecasts in (2) and (4), a forecast that combines the individual and pooled forecasts, $\hat{y}_{i,T+1}$ and $\tilde{y}_{i,T+1}$, may be desirable. As noted in the forecast combination literature (e.g., Timmermann, 2006), forecast combinations tend to perform particularly well if the correlation between individual forecast errors is weak. Correlations between forecast errors based on the individual and pooled estimation schemes decrease with (i) a lower variance of the common unpredictable component (σ_i^2), (ii) a greater difference in the estimates of $\boldsymbol{\beta}_i$ resulting from larger estimation errors (small T and N) or greater heterogeneity (large $\boldsymbol{\Omega}_\eta$), and (iii) estimation bias of the pooled estimator due to correlated heterogeneity.

In contrast, if the level of heterogeneity is either very large or very small, either the individual or pooled estimation approach will be dominant, thereby reducing the potential gains from forecast combination. Similarly,

if T is very small but N is large and there is little parameter heterogeneity, we would expect pooled estimation to dominate individual estimation by a sufficiently large margin that forecast combination offers small, if any, gains. Conversely, if T is very large and N is relatively small, forecasts using individual estimates will dominate forecasts using pooled estimates by a sufficient margin that renders forecast combination less attractive.

Building on these observations, consider the combined forecast

$$y^*_{i,T+1} = \omega_i \hat{y}_{i,T+1} + (1-\omega_i)\tilde{y}_{i,T+1}, \tag{11}$$

with associated forecast error

$$e^*_{i,T+1} = \omega_i \hat{e}_{i,T+1} + (1-\omega_i)\tilde{e}_{i,T+1}. \tag{12}$$

The error variance of the combined forecast is

$$\begin{aligned}\mathrm{Var}(e^*_{i,T+1}) = \omega_i^2 \mathrm{Var}(\hat{e}_{i,T+1}) + (1-\omega_i)^2 \mathrm{Var}(\tilde{e}_{i,T+1}) \\ + 2\omega_i(1-\omega_i)\mathrm{Cov}(\hat{e}_{i,T+1}, \tilde{e}_{i,T+1}),\end{aligned}$$

and the optimal ω_i^*, chosen to minimize the MSFE, is given by

$$\omega_i^* = \frac{\mathrm{Var}(\tilde{e}_{i,T+1}) - \mathrm{Cov}(\hat{e}_{i,T+1}, \tilde{e}_{i,T+1})}{\mathrm{Var}(\hat{e}_{i,T+1}) + \mathrm{Var}(\tilde{e}_{i,T+1}) - 2\mathrm{Cov}(\hat{e}_{i,T+1}, \tilde{e}_{i,T+1})}. \tag{13}$$

The optimal combination weights therefore depend on the variances of the underlying forecast errors as well as the covariances between forecast errors. Using (13) and our earlier expressions for the forecast errors, we have the following result:

Proposition 2. *For fixed $T > T_0$ such that Assumptions 1–6 hold, the optimal combination weights that minimize the MSFE conditional on $\boldsymbol{X}_i$ and $\boldsymbol{x}_{i,T+1}$, $\boldsymbol{\Omega}_\eta$ and σ_i^2 are given by*

$$\omega_i^* = \frac{\boldsymbol{x}'_{i,T+1}\boldsymbol{\Omega}_\eta \boldsymbol{x}_{i,T+1}}{\boldsymbol{x}'_{i,T+1}\left[T^{-1}\sigma_i^2 \boldsymbol{Q}_{iT}^{-1} + \boldsymbol{\Omega}_\eta\right]\boldsymbol{x}_{i,T+1}} + O_p\left(\frac{1}{N}\right), \tag{14}$$

for $i = 1, 2, \ldots, N$.

The proof is in Appendix A.2.

Note that for $T \to \infty$ the weights tend to 1 as any parameter heterogeneity will outweigh the vanishing uncertainty of the individual estimation. For

a finite T, however, the optimal weights in (14) depend on the unknown population parameters $\boldsymbol{\Omega}_\eta$ and σ_i^2. In practice, we can replace these parameters with estimates, $\hat{\boldsymbol{\Omega}}_\eta$ and $\hat{\sigma}_i^2$, to obtain

$$\hat{\omega}_i^* = \frac{\boldsymbol{x}_{i,T+1}' \hat{\boldsymbol{\Omega}}_\eta \boldsymbol{x}_{i,T+1}}{\boldsymbol{x}_{i,T+1}' \left[\frac{1}{T}\hat{\sigma}_i^2 \boldsymbol{Q}_{iT}^{-1} + \hat{\boldsymbol{\Omega}}_\eta\right] \boldsymbol{x}_{i,T+1}}, \tag{15}$$

where

$$\begin{aligned} \hat{\boldsymbol{\Omega}}_\eta &= \frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{\boldsymbol{\beta}}_i - \bar{\hat{\boldsymbol{\beta}}}\right)\left(\hat{\boldsymbol{\beta}}_i - \bar{\hat{\boldsymbol{\beta}}}\right)', \quad \bar{\hat{\boldsymbol{\beta}}} = N^{-1}\sum_{i=1}^{N}\hat{\boldsymbol{\beta}}_i, \\ \hat{\sigma}_i^2 &= (T-k)^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_i)'(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_i). \end{aligned}$$

Under Assumptions 1–6, $\mathrm{E}(\hat{\boldsymbol{\Omega}}_\eta) = \boldsymbol{\Omega}_\eta + \frac{1}{NT}\sum_{i=1}^{N}\sigma_i^2\boldsymbol{Q}_{iT}^{-1}$, and an unbiased estimator of $\boldsymbol{\Omega}_\eta$ is given by

$$\tilde{\boldsymbol{\Omega}}_\eta = \hat{\boldsymbol{\Omega}}_\eta - \frac{1}{NT}\sum_{i=1}^{N}\hat{\sigma}_i^2\boldsymbol{Q}_{iT}^{-1} \tag{16}$$

which yields the following estimate of ω_i^*

$$\tilde{\omega}_i^* = \frac{\boldsymbol{x}_{i,T+1}'\left[\hat{\boldsymbol{\Omega}}_\eta - \frac{1}{NT}\sum_{i=1}^{N}\hat{\sigma}_i^2\boldsymbol{Q}_{iT}^{-1}\right]\boldsymbol{x}_{i,T+1}}{\boldsymbol{x}_{i,T+1}'\left[\frac{1}{T}\hat{\sigma}_i^2\boldsymbol{Q}_{iT}^{-1} + \hat{\boldsymbol{\Omega}}_\eta - \frac{1}{NT}\sum_{i=1}^{N}\hat{\sigma}_i^2\boldsymbol{Q}_{iT}^{-1}\right]\boldsymbol{x}_{i,T+1}} \tag{17}$$

which we will refer to as bias-corrected weights.

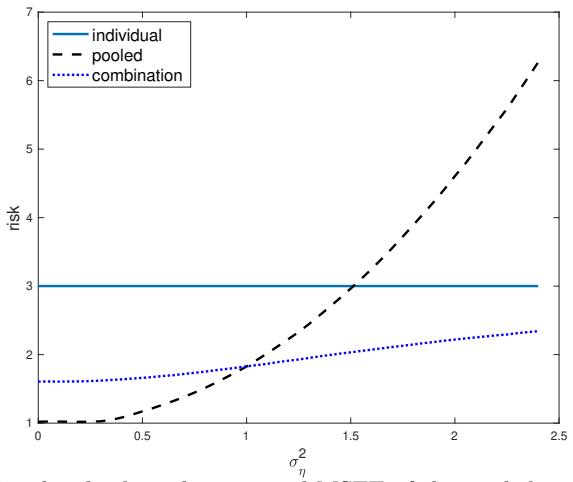
A downside to using the bias-corrected weights in (17) is that estimates of the covariance matrix of the parameters (16) are not guaranteed to be positive definite and the weights can be negative or exceed one. Moreover, if the denominator is close to zero, the weights may take on extreme values which can adversely affect forecasting performance. To avoid such scenarios, we propose to restrict the weights in (17) to values between zero and one and we will do so in the Monte Carlo experiments in Section 3 and in the applications in Section 4.

The second input in (15) are the individual error variances. The standard unbiased estimator of σ_i^2, $\hat{\sigma}_i^2 = (T-K)^{-1}\sum_{t=1}^{T}(y_{it} - \hat{\boldsymbol{\beta}}_i'\boldsymbol{x}_{it})^2$, tends to be poorly estimated for small T. In Section 2.4, we introduce an alternative estimator of σ_i^2 based on a mean group estimator of $\boldsymbol{\beta}$, which is not adversely affected by individual outlier estimates, $\hat{\boldsymbol{\beta}}_i$.

To compare the above forecasting schemes, Figure 1 plots the risk functions, that is, the expected MSFE values of the forecasts based on (i) the individual estimates; (ii) the pooled estimator; and (iii) the optimal combination using the weights in (15). The plot, obtained using simulations for $K = 1$ with details given in Appendix B, maps risk as a function of the degree of heterogeneity measured in terms of σ_η^2, a scalar version of $\mathbf{\Omega}_\eta$.

For low levels of parameter heterogeneity, the pooled forecast has a distinctly lower risk as compared to the individual forecasts, with the forecast combination falling between these two. Differences between the pooled and combination forecasts are due to the estimation error in the covariance matrix and the error variance. The combination forecast is, however, a clear improvement over the individual forecasts. As the degree of heterogeneity increases, the risk of the pooled forecast rises above that of the individual approach (which has a constant risk) with the forecast combination now having a lower risk than either of the other two forecasts. Overall, the combination forecast has an attractive risk profile as it avoids producing high levels of risk regardless of the degree of heterogeneity.

Figure 1: Risk versus parameter heterogeneity



Note: The plot displays the expected MSFE of the pooled, individual, and combination forecasts. The horizontal axis measures the degree of parameter heterogeneity in the simple panel regression model. Details of the construction of the plot is provided in Appendix B.

2.4 Forecast-based tests for pooling

Figure 1 shows that there are regions in the parameter variance space where the pooled forecast is more precise than the individual forecast and *vice versa*. A possible strategy is then to apply a pre-test to determine which of these cases applies to a given data set. One obvious approach would be to directly test the null of parameter homogeneity, namely $\boldsymbol{\beta}_i = \boldsymbol{\beta}$, for all i. There are a number of such tests in the literature that focus on the dispersion of $\hat{\boldsymbol{\beta}}_i$ around $\bar{\hat{\boldsymbol{\beta}}}$, proposed by Swamy (1970) and developed further by Pesaran and Yamagata (2008). For a recent survey see Hsiao (2022, Ch. 13). However, forecast errors depend on $\boldsymbol{x}'_{i,T+1}(\hat{\boldsymbol{\beta}}_i - \bar{\hat{\boldsymbol{\beta}}})$ rather than on $(\hat{\boldsymbol{\beta}}_i - \bar{\hat{\boldsymbol{\beta}}})$ and a more appropriate criterion would be the difference between the MSFE of individual and pooled forecasts, namely $\text{MSFE}(\hat{y}_{i,T+1}) - \text{MSFE}(\tilde{y}_{i,T+1})$.

Under Assumptions 1–6 and given the expressions for the MSFE of the individual and pooled forecasts in (8) and (9), we have

$$\begin{aligned}&\text{MSFE}(\hat{y}_{i,T+1}) - \text{MSFE}(\tilde{y}_{i,T+1}) \\ &\quad = T^{-1}\sigma_i^2 \boldsymbol{x}'_{i,T+1}\boldsymbol{Q}_{iT}^{-1}\boldsymbol{x}_{i,T+1} - \boldsymbol{x}'_{i,T+1}\boldsymbol{\Omega}_\eta \boldsymbol{x}_{i,T+1} + O_p\left(N^{-1}\right). \end{aligned} \tag{18}$$

The first term on the right hand side represents the small T estimation error of the individual approach while the second term captures increased variance of the pooled method due to the parameter heterogeneity. The pooled approach will dominate individual forecasts if, on average, the estimation error is larger than the parameter heterogeneity. Importantly, however, the estimation error and the parameter heterogeneity are weighted by the regressors in the forecast period, and as such it makes more sense to consider the average dispersion of $z_{i,NT} = \boldsymbol{x}'_{i,T+1}(\hat{\boldsymbol{\beta}}_i - \bar{\hat{\boldsymbol{\beta}}})$ rather than $(\hat{\boldsymbol{\beta}}_i - \bar{\hat{\boldsymbol{\beta}}})$. Accordingly, we propose a test based on $z_{i,NT}^2$, as set out in the following proposition.

Proposition 3. *Suppose that Assumptions 1–6 hold, $\boldsymbol{\eta}_i$ and $\boldsymbol{\varepsilon}_i$ are normally distributed, and $\boldsymbol{\eta}_i$ are cross-sectionally independent. Then, under the null of equal forecast accuracy defined by*

$$H_{0,PF} : T^{-1}\sigma_i^2 \boldsymbol{x}'_{i,T+1}\boldsymbol{Q}_{iT}^{-1}\boldsymbol{x}_{i,T+1} = \boldsymbol{x}'_{i,T+1}\boldsymbol{\Omega}_\eta \boldsymbol{x}_{i,T+1}, \quad \forall i, \tag{19}$$

there exists a finite T_0 such that for all $T > T_0$ and as $N \to \infty$

$$PF_{NT} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(\frac{\omega_{i,NT}^2 - 1}{\sqrt{2}} \right) \xrightarrow{d} \text{N}(0,1), \tag{20}$$

where

$$\omega_{i,NT}^2 = \frac{z_{i,NT}^2}{2\boldsymbol{x}_{i,T+1}'\boldsymbol{\Omega}_\eta \boldsymbol{x}_{i,T+1}} = \frac{T\left[\boldsymbol{x}_{i,T+1}'(\hat{\boldsymbol{\beta}}_i - \bar{\hat{\boldsymbol{\beta}}})\right]^2}{2\sigma_i^2 \boldsymbol{x}_{i,T+1}'\boldsymbol{Q}_{iT}^{-1}\boldsymbol{x}_{i,T+1}}. \tag{21}$$

The proof is provided in Appendix A.3.

If the PF_{NT} test statistic exceeds the critical value for a given significance level, the individual forecast should yield the more precise forecast. Therefore, the PF_{NT} test in (20) should be applied as a one-sided test. The assumption of Gaussianity is needed for deriving the variance of $z_{i,NT}^2$. The assumption of error cross-sectional independence can be relaxed but involves technical challenges. When $K = 1$, the PF_{NT} test becomes equivalent to slope homogeneity tests. In this case $\omega_{i,NT}^2 = \frac{1}{2}\left(\frac{\boldsymbol{x}_i'\boldsymbol{x}_i}{\sigma_i^2}\right)(\hat{\boldsymbol{\beta}}_i - \bar{\hat{\boldsymbol{\beta}}})^2$ which no longer depends on $\boldsymbol{x}_{i,T+1}$, and corresponds to s_i in Pesaran and Yamagata (2008, p. 56).

We need suitable estimates of σ_i^2 for the test to be applicable when N is large relative to T. The time series estimator of σ_i^2, namely $\hat{\sigma}_i^2 = (T-k)^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_i)'(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_i)$, is unbiased but requires large T to be consistent. To reduce the dependence of the estimator of σ_i^2 on individual estimates of $\boldsymbol{\beta}_i$, we propose the following alternative estimator based on the mean group estimator, $\bar{\hat{\boldsymbol{\beta}}} = N^{-1}\sum_{i=1}^N \hat{\boldsymbol{\beta}}_i$,

$$\tilde{\sigma}_{i,NT}^2 = \frac{(\boldsymbol{y}_i - \boldsymbol{X}_i\bar{\hat{\boldsymbol{\beta}}})'(\boldsymbol{y}_i - \boldsymbol{X}_i\bar{\hat{\boldsymbol{\beta}}})}{T + \mathrm{E}\left(\boldsymbol{x}_{i,T+1}'\boldsymbol{Q}_{iT}^{-1}\boldsymbol{x}_{i,T+1}\right)}. \tag{22}$$

Under the null of equal forecast accuracy, (19), and assuming that

$$T^{-1}\sum_{t=1}^T \mathrm{E}\left(\boldsymbol{x}_{it}\boldsymbol{x}_{it}'\right) = \mathrm{E}\left(\boldsymbol{x}_{i,T+1}\boldsymbol{x}_{i,T+1}'\right), \tag{23}$$

which holds, for example, if $\boldsymbol{x}_{it}$ is stationary. It then follows that

$$\mathrm{E}\left(\tilde{\sigma}_{i,NT}^2\right) = \sigma_i^2 + O\left(N^{-1}\right). \tag{24}$$

See Appendix A.4 for a proof. Using this result in (20) now yields

$$\widetilde{PF}_{NT} = \frac{1}{\sqrt{N}}\sum_{i=1}^N \left(\frac{\tilde{\omega}_{i,NT}^2 - 1}{\sqrt{2}}\right)$$

where

$$\tilde{\omega}^2_{i,NT} = \frac{T\left[\boldsymbol{x}'_{i,T+1}(\hat{\boldsymbol{\beta}}_i - \bar{\hat{\boldsymbol{\beta}}})\right]^2}{2\tilde{\sigma}^2_{i,NT}\boldsymbol{x}'_{i,T+1}\boldsymbol{Q}^{-1}_{iT}\boldsymbol{x}_{i,T+1}}. \tag{25}$$

Finally, we propose to approximate $\mathrm{E}\left(\boldsymbol{x}'_{i,T+1}\boldsymbol{Q}^{-1}_{iT}\boldsymbol{x}_{i,T+1}\right)$, which appears in the denominator of $\tilde{\sigma}^2_{i,NT}$, by

$$\hat{\mathrm{E}}\left(\boldsymbol{x}'_{i,T+1}\boldsymbol{Q}^{-1}_{iT}\boldsymbol{x}_{i,T+1}\right) \approx N^{-1}\sum_{i=1}^{N}\boldsymbol{x}'_{i,T+1}\boldsymbol{Q}^{-1}_{iT}\boldsymbol{x}_{i,T+1},$$

which holds when $\boldsymbol{x}_{it}$ are random draws from a common distribution. Due to the sampling errors involved in using $\tilde{\sigma}^2_{i,NT}$ for σ^2_i, and $\hat{\mathrm{E}}\left(\boldsymbol{x}'_{i,T+1}\boldsymbol{Q}^{-1}_{iT}\boldsymbol{x}_{i,T+1}\right)$ for $\mathrm{E}\left(\boldsymbol{x}'_{i,T+1}\boldsymbol{Q}^{-1}_{iT}\boldsymbol{x}_{i,T+1}\right)$, the use of $\widetilde{PF}_{NT}$ is asymptotically justified only if N and T are relatively large.

2.5 Shrinkage forecasts

Maddala et al. (1997) consider three iterative shrinkage estimators for improving estimation in panels with heterogeneous parameters, namely the prior likelihood, Bayesian, and empirical Bayes initially introduced by Lee and Griffiths (1979). They do not focus on forecasting, yet it is straightforward to use their estimators for this purpose.[3] The estimators take the same basic form:

$$\hat{\boldsymbol{\beta}}^*_i = \left(\frac{1}{\hat{\sigma}^2_i}\boldsymbol{X}'_i\boldsymbol{X}_i + \hat{\boldsymbol{\Omega}}^{*-1}\right)^{-1}\left(\frac{1}{\hat{\sigma}^2_i}\boldsymbol{X}'_i\boldsymbol{X}_i\hat{\boldsymbol{\beta}}_i + \hat{\boldsymbol{\Omega}}^{*-1}\bar{\boldsymbol{\beta}}^*\right) \tag{26}$$

where $\bar{\boldsymbol{\beta}}^* = N^{-1}\sum_{i=1}^{N}\hat{\boldsymbol{\beta}}^*_i$ is the average of the respective estimators across individual units. The three estimators differ in the choices of $\hat{\sigma}^2_i$ and $\hat{\boldsymbol{\Omega}}^*$.

Specifically, the prior likelihood estimator uses

$$\hat{\sigma}^2_i = \frac{1}{T}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}^*_i\right)\left(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}^*_i\right)' \tag{27}$$

and

$$\hat{\boldsymbol{\Omega}}^* = \frac{1}{N}\sum_{i=1}^{N}(\hat{\boldsymbol{\beta}}^*_i - \bar{\boldsymbol{\beta}}^*)(\hat{\boldsymbol{\beta}}^*_i - \bar{\boldsymbol{\beta}}^*)' \tag{28}$$

[3] Maddala et al. (1997) also discuss a Stein-type estimator. In our applications, this estimator is far less accurate so, for brevity, we omit this estimator.

The Bayesian estimator sets

$$\hat{\sigma}_i^2 = \frac{1}{T+2}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_i^*\right)\left(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_i^*\right)' \tag{29}$$

and

$$\hat{\boldsymbol{\Omega}}^* = \frac{1}{N-k-1}\left[\boldsymbol{R} + \sum_{i=1}^{N}(\hat{\boldsymbol{\beta}}_i^* - \bar{\boldsymbol{\beta}}^*)(\hat{\boldsymbol{\beta}}_i^* - \bar{\boldsymbol{\beta}}^*)'\right] \tag{30}$$

where $\boldsymbol{R}$ is the priors for $\boldsymbol{\Omega}$. Maddala et al. (1997) choose a relatively uninformative prior, setting $\boldsymbol{R}$ to a diagonal matrix with small positive entries on the diagonal as they note that the choice of $\boldsymbol{R}$ can have implications for the convergence of the maximization to obtain the parameter estimates. Finally, the empirical Bayes estimator uses

$$\hat{\sigma}_i^2 = \frac{1}{T-k}\left(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_i^*\right)\left(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_i^*\right)' \tag{31}$$

and

$$\hat{\boldsymbol{\Omega}}^* = \frac{1}{N-1}\left[\boldsymbol{R} + \sum_{i=1}^{N}(\hat{\boldsymbol{\beta}}_i^* - \bar{\boldsymbol{\beta}}^*)(\hat{\boldsymbol{\beta}}_i^* - \bar{\boldsymbol{\beta}}^*)'\right] \tag{32}$$

None of these three estimators have closed form solutions and require an iterative approach for estimation.

3 Monte Carlo experiments

In this section, we use Monte Carlo simulations to explore the finite-sample performance of a number of panel forecasting methods including those listed above.

3.1 Design

We adopt a fairly general design which allows for dynamics, parameter heterogeneity, and correlations between the regressors and coefficients. We also consider the nature of the trade-off between heterogeneity and estimation uncertainty under different degrees of fit of the underlying panel regressions. In particular, we consider the following data generating process (DGP):

$$y_{i,t+1} = \alpha_i + \rho_i y_{it} + \gamma_i x_{it} + \kappa\sigma_i\varepsilon_{i,t+1}, \tag{33}$$

where $\varepsilon_{i,t+1} \sim \text{iidN}(0,1)$, $\sigma_i^2 \sim \text{iid}\left(1+\chi_1^2\right)/2$,

$$x_{it} = \mu_{xi} + \xi_{it}, \quad \mu_{xi} = (z_i^2 - 1)/\sqrt{2}, \quad z_i \sim \text{iidN}\left(0,1\right),$$

$$\xi_{it} = \rho_{xi}\xi_{i,t-1} + \sigma_{xi}\left(1-\rho_{xi}^2\right)^{1/2}\nu_{it}, \quad \nu_{it} \sim \text{iidN}\left(0,1\right),$$

and $\sigma_{xi}^2 \sim \text{iid}\left(1+\chi_1^2\right)/2$, for individual units $i = 1, 2, \ldots, N$ and observation periods $t = -100, -99, \ldots, -1, 0, 1, \ldots, T$, where the draws for $t = -100, -99, \ldots, -1$, are discarded and the following panel regressions are estimated to compute forecasts of $y_{i,T+1}$:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta}_i + \kappa\sigma_i\boldsymbol{\varepsilon}_i,$$

where $\boldsymbol{y}_i = (y_{i1}, y_{i2}, y_{i3}, \ldots, y_{iT})'$, $\boldsymbol{X}_i = (\boldsymbol{\iota}_T, \boldsymbol{y}_{i,-1}, \boldsymbol{x}_{i,-1})$, $\boldsymbol{\iota}_T$ is a $T \times 1$ vector of ones, $\boldsymbol{y}_{i,-1} = (y_{i0}, y_{i1}, \ldots, y_{i,T-1})'$, $\boldsymbol{x}_{i,-1} = (x_{i0}, x_{i1}, \ldots, x_{i,T-1})'$, $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT})'$, and $\boldsymbol{\beta}_i = (\alpha_i, \rho_i, \gamma_i)'$. Given our focus on large N panels, we set $N = 500$ and consider different time dimensions, namely $T = \{20, 30, 50, 100\}$.

The autocorrelations of x_{it} are generated as $\rho_{xi} \sim \text{iid Uniform}(0, 0.95)$, thus allowing for a high degree of dynamic heterogeneity in the regressors. The coefficients of lagged dependent variables, $y_{i,t-1}$, are generated as $\rho_i \sim \text{iid Uniform}(0, \bar{\rho})$, where we vary $\bar{\rho}$ to capture different degrees of dynamic heterogeneity. The static panel arises as a special case with $\bar{\rho} = 0$. The value of $\bar{\rho}$ depends on the chosen R^2, which is discussed below.

Unlike previous research, we also consider cases where the regressors and the coefficients are correlated. Specifically, we set

$$\alpha_i = \phi\mu_{xi} + \sigma_\eta\eta_i, \text{ and } \gamma_i = 1 + \theta\mu_{xi} + \sigma_\zeta\zeta_i,$$

where $\eta_i, \zeta_i \sim \text{iidN}(0,1)$. Importantly, such non-zero correlations need not bias the pooled estimates. As noted in Remark 5, what matters is the correlation between $\boldsymbol{X}_i'\boldsymbol{X}$ and $(\boldsymbol{\beta}_i - \boldsymbol{\beta})$. This has implications for the distribution of μ_{xi}, which drives the correlation. A symmetric distribution, such as the normal distribution, of μ_{xi} would imply that the correlation between regressors and coefficients is non-zero but, due to the symmetry of the normal distribution, we still have that $\text{E}[(\gamma_i - \gamma)x_{it}^2] = 0$. To avoid such an outcome, we draw μ_{xi} from a chi-square distribution.

To determine the magnitude of the correlation, consider taking expectations with respect to both i and t,

$$\begin{aligned}
&\text{E}\left(\gamma_i\right) = 1, \quad \text{Var}(\gamma_i) = \theta^2 + \sigma_\zeta^2, \\
&\text{E}\left(x_{it}\right) = \text{E}\left(\mu_{xi} + \xi_{it}\right) = 0, \quad \text{Var}\left(x_{it}\right) = \text{E}\left(x_{it} - \mu_{xi}\right)^2 = \sigma_{xi}^2,
\end{aligned}$$

and $\mathrm{E}\left[\mathrm{Var}\left(x_{it}\right)\right] = \mathrm{E}\left(1+\chi_1^2\right)/2 = 1$. Also, since ν_{it} is distributed independently of η_j and ζ_j for all t, i and j, $\mathrm{Cov}\left(\gamma_i, x_{it}\right) = \theta$ and $\mathrm{Corr}\left(\gamma_i, x_{it}\right) = \theta\left(\sigma_\zeta^2 + \theta^2\right)^{-1/2}$. To achieve a given level of $\mathrm{Corr}(\gamma_i, x_{it}) = r_{\gamma x}$, we set

$$\theta = \frac{r_{\gamma x}\sigma_\zeta}{\left(1 - r_{\gamma x}^2\right)^{1/2}}. \tag{34}$$

Similarly, to achieve $\mathrm{Corr}(\alpha_i, x_{it}) = r_{\alpha x}$, we set

$$\phi = \frac{r_{\alpha x}\sigma_\eta}{\left(1 - r_{\alpha x}^2\right)^{1/2}}. \tag{35}$$

Defining $\mathrm{Var}(\gamma_i) \equiv \sigma_\gamma^2 = \theta^2 + \sigma_\zeta^2$, we can use (34) to see that $\theta = r_{\gamma x}\sigma_\gamma$. An equivalent result emerges for ϕ where, for $\sigma_\alpha^2 = \mathrm{Var}(\alpha_i)$, we have $\phi = r_{\mu x}\sigma_\alpha$. We thus use the parameters σ_α^2, σ_γ^2, and $\bar{\rho}$ to vary the degree of parameter heterogeneity in α_i, γ_i and ρ_i, respectively.

We show in Appendix D that the value of the pooled R^2 of the individual regressions, PR^2, limits the value that ρ_i can take. We therefore use different values depending on PR^2. For each value of T, we set $\rho_{\beta x} = \rho_{\alpha x} \in \{0, 0.5\}$, and for each of these we run simulations for four combinations of $\{\bar{\rho}, \sigma_\eta^2, \sigma_\zeta^2\}$: For $PR^2 = 0.2$, we use $\{0, 0, 0\}$, $\{0.2, 0.1, 0.1\}$, $\{0.4, 0.25, 0.25\}$, and $\{0.6, 0.5, 0.5\}$. Similarly, for $PR^2 = 0.6$, we use $\{0, 0, 0\}$, $\{0.5, 0.1, 0.1\}$, $\{0.725, 0.25, 0.25\}$, and $\{0.95, 0.5, 0.5\}$.

Finally, we set κ^2 to achieve a desired level of average fit as measured by the pooled R^2 of the individual regressions. Details of how κ^2 is determined can be found in Appendix D. The resulting parameter values are reported in Table 7 in Appendix D.

We hold the parameters constant across replications but redraw the errors, ε_{it} and ν_{it} and generate data for different values of the average PR^2 for the individual estimation.

To provide a more comprehensive comparison of forecasting performance across different models, we add to our list two widely used methods, namely random effects estimation and median group estimation.

In summary, forecasts based on the following estimators are considered:

- Individual estimation;
- Pooled estimation;
- Random effects estimator of Goldberger (1962);
- Median Group estimator;

- Forecast combination with weights (15) and (17) using the variance introduced in Section 2.4 assuming serially independent data;
- Forecasts based on the pre-test discussed in Section 2.4, using the variance introduced in that section assuming serially independent data;
- Forecasts based on the three shrinkage estimators used by Maddala et al. (1997): prior likelihood, Bayesian, and empirical Bayes.

Further details of these estimators are provided in Appendix C.

Heterogeneity across units can lead to misleading results when calculating MSFE ratios that are averaged across all units. To obtain a more robust measure, for each method we report the median MSFE ratio computed across all units:

$$\text{median MSFE}(j) = \text{median}_{i\in(1,N)}\left[\frac{\mathcal{R}^{-1}\sum_{r=1}^{\mathcal{R}}\left(y_{i,T+1}-\hat{y}_{i,T+1}^{(j)}\right)^2}{\mathcal{R}^{-1}\sum_{r=1}^{\mathcal{R}}\left(y_{i,T+1}-\hat{y}_{i,T+1}^{(\text{bench})}\right)^2}\right], \quad (36)$$

where $\hat{y}_{i,j,T+1}^{(r)}$ denotes the forecast of method j and $\hat{y}_{i,bench,T+1}^{(r)}$ is the benchmark forecast, both for unit i in the rth simulation for $r = 1, 2, \ldots, \mathcal{R}$, where the number of replications is set to $\mathcal{R} = 10,000$.

3.2 Simulation results

Tables 1 and 2 report results for $R^2 = 0.2$ and $R^2 = 0.6$, respectively. Across rows we vary the correlation between the regressors and the coefficients ($r_{\gamma x}$) and the variance of the coefficients (σ_γ^2) which controls the degree of heterogeneity. The time-series dimension, T, varies across the columns. We keep N fixed at 500 as initial results suggest that variation in N has a much smaller influence on the results compared to variations in T. The MSFE ratios are computed using the forecasts from the individual forecasts as the benchmark with values below unity suggesting that a particular method produces more accurate forecasts than this benchmark, while values above unity suggest the reverse.

Panels in the top row display the results for the pooled estimation, the BLUP random effects estimation, and the median group estimator. Panels in the middle row show results for the combination forecasts and the pre-test forecasts. The bottom row contains the forecast results for the three shrinkage estimators of Maddala et al. (1997).

We summarize our observations from these simulations as follows:

Table 1: Monte Carlo results: relative MSFE, average $R^2 = 0.2$

$r_{\gamma,x}$	$\sigma_\gamma^2 \backslash T$	20	50	100	20	50	100	20	50	100
		Pooled			Random effects			Median Group		
0	0	0.841	0.938	0.969	0.884	0.957	0.979	0.844	0.939	0.970
0	0.1	0.899	0.993	1.025	0.907	0.978	1.003	0.913	0.999	1.029
0	0.25	0.939	1.035	1.066	0.929	1.002	1.026	0.971	1.051	1.077
0	0.5	0.952	1.040	1.064	0.943	1.017	1.037	1.002	1.061	1.076
0.5	0	0.841	0.938	0.970	0.884	0.957	0.979	0.844	0.939	0.970
0.5	0.1	0.924	1.014	1.047	0.908	0.980	1.003	0.928	1.006	1.037
0.5	0.25	0.977	1.070	1.095	0.929	1.002	1.023	0.992	1.061	1.082
0.5	0.5	0.987	1.072	1.099	0.942	1.018	1.038	1.018	1.073	1.089
		Combination, $\hat{\mathbf{\Omega}}_\eta$			Combination, $\tilde{\mathbf{\Omega}}_\eta$			PF test		
0	0	0.908	0.963	0.982	0.852	0.940	0.971	0.841	0.938	0.969
0	0.1	0.914	0.971	0.989	0.892	0.969	0.988	0.899	0.996	1.000
0	0.25	0.919	0.976	0.992	0.922	0.981	0.993	0.941	1.000	1.000
0	0.5	0.921	0.976	0.992	0.935	0.982	0.993	0.955	1.000	1.000
0.5	0	0.908	0.963	0.982	0.852	0.940	0.971	0.841	0.938	0.970
0.5	0.1	0.918	0.973	0.991	0.969	0.993	1.003	0.923	1.000	1.000
0.5	0.25	0.925	0.979	0.993	0.989	1.005	1.009	0.975	1.000	1.000
0.5	0.5	0.927	0.979	0.993	0.994	1.007	1.010	0.985	1.000	1.000
		Prior likelihood			Bayes			Empirical Bayes		
0	0	0.844	0.939	0.970	0.842	0.939	0.969	0.843	0.939	0.969
0	0.1	0.914	0.997	0.995	0.907	0.992	0.994	0.908	0.995	0.995
0	0.25	0.973	0.992	1.010	0.953	0.988	1.010	0.956	0.995	1.013
0	0.5	1.005	1.000	1.000	0.965	1.001	1.001	0.972	1.020	1.005
0.5	0	0.844	0.939	0.970	0.842	0.939	0.969	0.842	0.939	0.969
0.5	0.1	0.923	0.972	0.994	0.914	0.972	0.993	0.914	0.976	0.994
0.5	0.25	0.927	0.988	1.010	0.929	0.985	1.009	0.962	0.986	1.009
0.5	0.5	0.945	0.994	1.003	0.955	0.993	1.002	0.972	1.000	1.008

Note: The table reports the median of the individual MSFE of the respective method as a ratio of the MSFE of the individual forecasts. The DGP is given in Section 3.1, $N = 500$, and the parameterization to achieve an average R^2 of 0.2 is in Table 7. Details on the estimators are in Appendix C. Results are from 10,000 draws of the DGP.

Table 2: Monte Carlo results: relative MSFE, average $R^2 = 0.6$

$r_{\gamma,x}$	$\sigma^2_\gamma \backslash T$	20	50	100	20	50	100	20	50	100
		Pooled			Random effects			Median Group		
0	0	0.841	0.938	0.969	0.883	0.957	0.979	0.843	0.939	0.970
0	0.1	1.144	1.228	1.249	1.024	1.095	1.114	1.245	1.292	1.304
0	0.25	1.240	1.382	1.420	1.085	1.193	1.223	1.533	1.634	1.649
0	0.5	1.248	1.262	1.332	1.103	1.155	1.205	2.287	1.651	1.729
0.5	0	0.840	0.938	0.969	0.883	0.957	0.979	0.843	0.939	0.970
0.5	0.1	1.181	1.289	1.345	1.011	1.091	1.121	1.245	1.310	1.370
0.5	0.25	1.330	1.439	1.451	1.107	1.189	1.200	1.772	1.706	1.678
0.5	0.5	1.279	1.327	1.377	1.100	1.168	1.199	2.822	2.012	1.913
		Combination, $\hat{\mathbf{\Omega}}_\eta$			Combination, $\tilde{\mathbf{\Omega}}_\eta$			PF test		
0	0	0.909	0.964	0.982	0.851	0.940	0.971	0.841	0.938	0.969
0	0.1	0.941	0.987	0.997	1.023	1.018	1.003	1.000	1.000	1.000
0	0.25	0.951	0.993	0.999	1.035	1.050	1.017	1.000	1.000	1.000
0	0.5	0.954	0.987	0.997	1.000	1.027	1.025	1.000	1.000	1.000
0.5	0	0.909	0.964	0.982	0.851	0.940	0.971	0.841	0.938	0.969
0.5	0.1	0.946	0.990	0.999	1.003	1.016	1.021	1.000	1.000	1.000
0.5	0.25	0.960	0.995	0.999	1.001	1.008	1.014	1.000	1.000	1.000
0.5	0.5	0.958	0.990	0.998	1.000	1.000	1.002	1.000	1.000	1.000
		Prior likelihood			Bayes			Empirical Bayes		
0	0	0.844	0.940	0.970	0.841	0.938	0.969	0.842	0.938	0.969
0	0.1	0.992	1.000	0.995	0.976	0.996	0.994	0.983	1.006	0.995
0	0.25	1.005	0.985	0.996	1.010	0.985	0.996	1.028	0.986	0.996
0	0.5	1.006	0.994	0.999	0.994	0.994	0.998	1.017	1.015	1.005
0.5	0	0.844	0.940	0.970	0.841	0.938	0.969	0.841	0.938	0.970
0.5	0.1	0.968	1.002	0.995	0.957	1.005	0.995	0.962	1.028	0.995
0.5	0.25	1.027	0.985	0.995	1.015	0.986	0.996	1.021	1.006	0.996
0.5	0.5	1.001	0.993	1.008	0.984	0.992	1.007	0.998	1.004	1.008

Note: See footnote of Table 1.

- The pooled forecast is most precise under parameter homogeneity. Additionally, in the low R^2 environment and with $T = 20$, pooling provides fairly competitive forecasts even when parameters are heterogeneous. For larger T, the individual forecasts are more precise. In contrast, in the high R^2 environment, forecasts from individual estimation tend to be relatively more precise for all T when parameters are heterogeneous. In all settings, the pooled forecast is also relatively more adversely affected by a positive correlation between the slope parameter and the regressor than the individual forecast.

- The precision of forecasts based on random effects estimation often falls between those of the individual and pooled estimations when parameters are heterogeneous. In the low R^2 environment, forecasts based on random coefficients estimation can be more precise than either. The relative performance of the random effects forecast is not adversely affected by a positive correlation between the slope parameter and the regressor.

- The median group forecast improves over the individual forecasts when T is small but is generally less precise for larger T values. It is more precise than pooling only for high levels of heterogeneity and in large samples but can be fairly imprecise in other settings. This method performs particularly poorly in the high R^2 setting under high levels of parameter heterogeneity and when regressors and slope coefficients are correlated.

- The combination method that uses $\hat{\mathbf{\Omega}}_\eta$ provides the most precise forecast more often than any other methods and, in the remaining cases, is generally close to the best forecast. This method is uniformly more precise than the individual forecast. When $R^2 = 0.2$ and parameters are heterogeneous it is the most precise or close to being the most precise forecast with the only exception of $T = 20$ and low parameter heterogeneity. Also, the forecast precision of this method, measured relative to the individual forecasts, is robust with regards to any correlation between parameters and regressors. Under $R^2 = 0.6$ it is the most precise or very close to being the most precise method whenever parameters are heterogeneous for all T.

- Using the unbiased estimator of $\tilde{\mathbf{\Omega}}_\eta$ in (16) improves over the $\hat{\mathbf{\Omega}}_\eta$-based forecasts under parameter homogeneity, especially if T is small. In the low R^2 environment and if parameters and regressors are uncorrelated,

this method provides the best or very close to the best forecast. This forecast does, however, perform notably worse when parameters and regressors are correlated. In the high R^2 environment, the correction to the estimate $\tilde{\mathbf{\Omega}}_\eta$ is not as important as the numerical instability that it introduces and the forecast is less precise than that based on $\hat{\mathbf{\Omega}}_\eta$ under parameter heterogeneity.

- The accuracy of the pre-test forecast that tests between the individual and pooled forecasts is very close to that of the pooled forecast when this is better than the individual forecasts but equals that of the individual forecast when this is more precise than the pooled forecast. This holds for both values of R^2 and suggests that the PF test has a high success rate in identifying whether the pooled or individual forecasts will be most accurate.

- The methods considered by Maddala et al. (1997) generally perform well under parameter homogeneity, with a forecasting performance that is close to the pooled forecast. However, under parameter heterogeneity, their performance is mixed. They deliver the most precise forecast at times but at other times fail to beat the individual forecasts, though any under-performance relative to this benchmark tends to be modest, i.e., almost always less than 3% in our simulations. Although no uniform ordering emerges among the three forecasts, the Bayesian forecast tends to perform best overall.

Overall, these simulations demonstrate that the combination forecast using $\hat{\mathbf{\Omega}}_\eta$ is the most precise or close to being the most precise method whenever parameters are heterogeneous. The pre-test generally tends to choose the more precise of the pooled or individual forecasts but the resulting forecasts can be beaten by using forecast combination, unless the degree of heterogeneity is very small. The remaining methods show mixed results and their forecasting performance tends to depend more on the specific setting.

Our simulations also highlight the importance of correlated heterogeneity, that is, the nature of the correlation between γ_i and x_{it}. The pooled forecasts tend to perform poorly when $E\left[(\gamma_i - \gamma)\, x_{it}^2\right] \neq 0$.

The fit of the model which, in our simulations, is captured through the R^2 value also matters to the performance of the prediction methods. The better the fit (higher R^2), the more accurately the individual coefficients are estimated, and the more costly it becomes to ignore parameter heterogeneity.

Adding dynamics to the data generating process favors pooling over other methods when T is small because it tends to exacerbate estimation error,

which matters particularly when T is small and parameters are imprecisely estimated.

Individual empirical applications differ in how closely they resemble the spectrum of DGPs assumed in Tables 1 and 2. To illustrate this point, we consider three very different applications in the next section.

4 Empirical Applications

We apply our list of panel forecasting methods to the following three empirical applications: House price inflation in U.S. metropolitan areas, inflation of CPI sub-indices, and stock returns on U.S. firms. These three applications have very contrasting in-sample fit: the returns of individual stocks have an average R^2 below 0.01 when estimating parameters individually; for the CPI application the in-sample R^2 is around 0.2, and for the house price models it exceeds 0.8.

4.1 Measures of forecasting performance

Forecasts are evaluated using the out-of-sample MSFE of the individual units (i)

$$\text{MSFE}_{ij} = \frac{1}{T-T_1}\sum_{t=T_1}^{T-1}(y_{i,t+1} - \hat{y}_{i,j,t+1|t})^2, \tag{37}$$

where $\hat{y}_{i,j,t+1|t}$ is the forecast of $y_{i,t+1}$ using method j, conditional on the information at time t. We then calculate the ratio of the MSFE of method j relative to that of the benchmark method (individual estimation), and report the median ratio. We split the full sample of T observations into an estimation sample containing observations $t = 1, 2, \ldots, T_1$, and a test sample covering observations $T_1 + 1, T_1 + 2, \ldots, T$. Details of the size of the estimation sample are reported with each application.

We also report the proportion of units in the cross-section for which a given method produces a smaller MSFE than the benchmark

$$\frac{1}{N}\sum_{i=1}^{N}\text{I}\left[\frac{1}{T-T_1}\sum_{t=T_1}^{T-1}(y_{i,t+1} - \hat{y}_{i,j,t+1|t})^2 < \frac{1}{T-T_1}\sum_{t=T_1}^{T-1}(y_{it+1} - \hat{y}_{i,B,t+1|t})^2\right], \tag{38}$$

where $\text{I}(\cdot)$ is the indicator function that equals unity if the expression inside the operator is true and zero otherwise, and $\hat{y}_{i,B,t+1|t}$ denotes the benchmark

forecast. Additionally, we report the proportion of units in the cross-section for which a given method has the smallest or largest MSFE value computed as

$$\frac{1}{N}\sum_{i=1}^{N} \mathrm{I}\left[\mathrm{MSFE}_{ij} = \min_{l} \mathrm{MSFE}_{il}\right],$$
$$\frac{1}{N}\sum_{i=1}^{N} \mathrm{I}\left[\mathrm{MSFE}_{ij} = \max_{l} \mathrm{MSFE}_{il}\right]. \tag{39}$$

These measures help us better understand the risk of under-performance and, on the upside, the possibility of superior predictive accuracy. Note that these proportions can add up to more than one due to ties between forecasting methods, for example when the pre-test forecasts selects the forecasts from pooling for the entire forecast window for a particular unit.

Finally, we provide density plots of the individual ratios of MSFEs for selected methods. These plots give more detailed insights into the distribution of the forecast performance across different units.

4.2 U.S. house prices

Our first application uses quarterly data on real house price inflation in 377 U.S. Metropolitan Statistical Areas (MSAs) from the first quarter of 1975 to the fourth quarter of 2014—a data set that has previously been analyzed by Yang (2021). From the Freddie Mac house price index, we calculate annual house price inflation rates for each MSA which we then deflate by the CPI. Our forecasts focus on the one-quarter-ahead MSA-level rate of house price changes. The sample covers a total of 160 quarterly periods which we split into two halves using the first as the training sample and the second to evaluate the forecasts. Forecasts start in the first quarter of 1995 and end in the fourth quarter of 2014, a total of 80 quarterly periods. We use a rolling window of 60 quarters to estimate the parameters.

Our prediction model for the house price inflation rate in quarter t for MSA i, y_{it}, takes the form

$$y_{it} = \alpha_i + \rho_i y_{i,t-1} + \rho_i^* y_{i,t-1}^* + \gamma_{Ri}\bar{y}_{i,t-1}^{(R)} + \gamma_{Ci}\bar{y}_{t-1}^{(C)} + \varepsilon_{it}, \tag{40}$$

where $i = 1, 2, \ldots, N$ denotes individual MSAs and $t = 1, 2, \ldots, T$ refers to the time period, $y_{it}^* = \sum_{k=1, k\neq i}^{N} \omega_{ik} y_{jt}$ is the spatial effect for a set of spatial weights ω_{ik}, $\bar{y}_{it}^{(R)}$ is the average house price inflation in the region of unit i, and $\bar{y}_t^{(C)}$ is the country-wide average house price inflation. The spatial

weights, ω_{ik}, measure the spatial effect of house prices in MSA k on house prices in MSA i. We construct spatial weights from geographic distance, that is $\omega_{ik} = v_{ik}/\sum_{k=1}^{N} v_{ik}$ and $v_{ik} = 1$ if the MSAs are at most 100 miles apart and zero otherwise. For details of these weights see Yang (2021). Our analysis excludes MSAs that do not have any neighbors within 100 miles, which leaves 362 MSAs in our sample.

We consider two forecasting models: the first, denoted SAR, is a spatial autoregressive model that excludes the regional and country-wide averages, i.e., (40) with $\gamma_{Ri} = \gamma_{Ci} = 0$. The second, denoted SARX, is the model in (40) with all coefficients left unrestricted.

Table 3 reports the results. In the columns headed Median MSFE, the first row shows the median MSFE value for the different MSAs for the individual forecasts. Subsequent rows report the median of the MSFE ratios for the respective methods computed relative to the individual forecasts. Values of unity mean that for half of the MSAs, forecasts produced by the method listed in the row are at least as accurate as those from the benchmark with the remaining half being equally good or worse. Values below unity imply that the row method is more accurate than the benchmark for more than half of the MSAs, while values above unity suggest the opposite. The next two columns headed 'freq. beating benchmark' report the proportion of MSAs for which the respective methods have a smaller MSFE than the benchmark, while the columns headed 'freq. smallest MSFE' and 'freq. largest MSFE' show the proportion of MSAs for which the respective methods have the smallest or largest mean square forecast error among all forecasting methods.

A first observation is that the individual forecasts from the SAR model that excludes the regional and national averages tend to be slightly more accurate, on average, than their counterparts based on the SARX model but also have a wider spread in forecasting performance. The individual models produce the most accurate forecasts from the SAR model for only 7.7% of the MSAs but the worst forecasts for a staggering 45.0% of MSAs. Similarly, the individual SARX forecasts are most accurate for only 3.6% of the MSAs, but least accurate for 40.1% of MSAs. Some form of pooling is clearly advantageous for this data set.

The pooled method improves over the benchmark by reducing the median MSFE by between 1% and 3%, and beating the benchmark by up to two-thirds of the MSAs. For the SAR model, the pooled forecasts generate the most accurate forecasts for 3.0% of the MSAs while they are least accurate for 11.9% of the MSAs. These numbers are somewhat worse for the SARX model, amounting to 1.1% and 20.2%. The pooled forecasts, there-

Table 3: House price inflation forecasts

Forecast methods	Median MSFE		freq. beating benchmark		freq. smallest MSFE		freq. largest MSFE	
	SAR	SARX	SAR	SARX	SAR	SARX	SAR	SARX
Individual	2.536	2.569	–	–	0.077	0.036	0.450	0.401
Pooled	0.971	0.989	0.660	0.539	0.030	0.011	0.119	0.202
RE	0.952	0.952	0.754	0.682	0.221	0.091	0.041	0.044
Median group	0.952	0.941	0.727	0.688	0.312	0.318	0.050	0.069
Optimal combination								
Naive	0.980	0.975	0.876	0.934	0.019	0.047	0.000	0.000
Bias adj.	0.974	0.966	0.859	0.914	0.069	0.119	0.006	0.006
Pre-test								
PF	0.984	0.974	0.608	0.691	0.102	0.185	0.213	0.091
Shrinkage								
Prior lik.	0.970	0.963	0.715	0.622	0.047	0.088	0.105	0.149
Bayes.	0.960	0.948	0.749	0.699	0.058	0.047	0.006	0.003
Emp. Bayes.	0.956	0.954	0.754	0.652	0.064	0.058	0.011	0.036

Note: SAR denotes the spatial autoregressive model and SARX the same model with regional and nationwide house prices average added. The results in this table use geographic spatial weighting: being in a 100km neighborhood, and an estimation window of 60 observations. In the column headed Median MSFE, the first line gives the median level of the MSFE of the individual forecasts and the following lines the median ratios of MSFEs of the respective method relative to that of the individual forecast. The column headed 'freq. beating benchmark' gives the proportion of MSAs where the respective method has a smaller MSFE than the benchmark, the column headed 'freq. smallest MSFE' gives the proportion of MSAs where the respective method has the smallest MSFE, and the column headed 'freq. largest MSFE' gives the proportion of MSAs where the respective method has the largest MSFE. See Appendix C for details of the estimators.

fore, produce the worst forecasts for far more MSAs than they produce the best ones, which suggests that pooling is not the optimal method for this data set either.

Random effects forecasts of house price inflation are more precise than both the pooled and individual forecasts. It performs best for the SAR model, where it beats the benchmark for 75.4% of the MSAs, is the most precise method for 22.1% of the MSAs, and only produces the least accurate forecasts for 4.1% of the MSAs. For the SARX model, it produces the most precise forecasts less often (9.1%) and produces the worst forecasts at about half of this rate (4.4%). Estimating heterogeneous intercepts seems to work well in this application. This is consistent with substantial heterogeneity in average house price appreciation across the MSAs in our data.

The median group estimator produces the most accurate forecasts in

this application both on average, across all MSAs, and also in terms of the frequency at which it is the best forecast at the unit level among all forecasting methods. It beats the benchmark for 72.7% in the case of the SAR models and 68.8% in the case of the SARX model and provides the most precise forecasts for 31.2% and 31.8% of MSAs. In contrast, it provides the least precise forecast only for 5.0% and 6.9% of MSAs. This suggests that the median group estimator successfully adapts to the moderate level of parameter heterogeneity that is present in this application.

Combination forecasts also improve over both the individual forecasts and the pooled forecasts. The optimal combination forecasts using bias-corrected weights are more accurate than those from the combination scheme that uses naive weights and beat the benchmark for 85.9% and 91.4% of MSAs. The combination forecasts with bias-corrected weights are the most accurate ones for 6.9% and 11.9% of MSAs but only for 1.9% and 4.7% when using the naive weights. Remarkably, however, it is extremely rare for them to produce the worst performance. The combination forecast using the naive weights never produces the worst forecast and the bias-corrected weights only does so for 0.6% of the MSAs. This is a feature we will find again in our other applications.

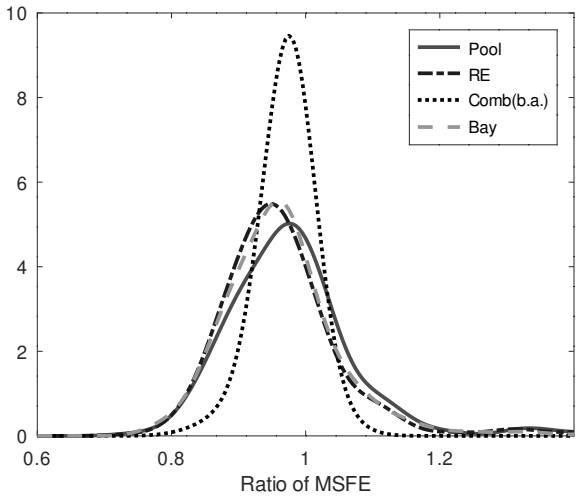
Pre-tests improve on the individual forecasts and in the case of the SARX model also on the pooled forecast. They beat the benchmark in about two thirds of the MSAs. Interestingly, while the pre-tests produce the most accurate forecasts at only half the rate at which they produce the least accurate forecasts (10.2% versus 21.3%) for the SAR model, this relation flips for the SARX model for which the pre-tests produce the smallest MSFE (18.5%) at twice the rate at which these forecasts are worst (9.1%).[4]

The shrinkage forecasts are also more precise than both the individual and pooled forecasts. The Bayesian and empirical Bayes forecasts tend to be slightly more accurate than the prior likelihood forecasts. The shrinkage forecasts beat the benchmark in between 62.2% and 75.4% of the MSAs. They are the most precise forecasts for between 4.7% and 8.8% of the MSAs and produce the worst forecast for between 0.3% and 14.9% of the MSAs with the Bayesian forecasts being notably less likely than the other shrinkage methods to yield the worst forecast for individual MSAs.

Our findings can be summarized by considering Figure 2, which plots densities fitted to the cross-sectional distribution of MSFE ratios for four of the forecasting methods. The pooled forecasts are centered around 0.97 and

[4]The median value of the PF test statistic is 1.70 and 2.82 for the SAR and SARX models, respectively.

Figure 2: Density of ratios of MSFE: House prices



Note: The graph shows density plots of the ratios of the MSFEs for selected forecasts (pooled, random effects, bias-adjusted combination, and Bayesian) for the SAR model. The density estimates use a normal kernel with bandwidth 0.03.

have a high dispersion relative to the other methods. The MSFE distribution of the random effect forecasts is shifted to the left and has a slightly smaller dispersion than the pooled forecasts. The forecast combination and Bayesian approaches have a much lower dispersion than the other methods and rarely underperform for individual MSAs.

4.3 CPI inflation of sub-indices

Our second application covers inflation rates for 202 sub-indices of the US consumer price index (CPI). The data is measured at the monthly frequency and spans the 60-year period from November 1958 to December 2018. We use rolling estimation windows with 60 observations and require each estimation sample to be balanced, excluding individual units without a complete set of observations in a given window. After accounting for the necessary pre-samples, we generate 590 forecasts for each series, with the first forecast computed for November 1969.

We consider three forecasting models: (i) a purely autoregressive specification with lags 1, 2, and 12, denoted AR; (ii) the same AR specification augmented with the lagged value of the first principal component of the data,

Table 4: CPI inflation forecasting results

Forecast method	Median MSFE			freq. beating benchmark			freq. smallest MSFE			freq. largest MSFE		
	AR	AR-PC	AR-X	AR	AR-PC	AR-X	AR	AR-PC	AR-all	AR	AR-PC	AR-X
Individual	1.568	1.573	1.641	–	–	–	0.059	0.054	0.035	0.064	0.054	0.183
Pooled	1.076	1.077	1.114	0.351	0.347	0.361	0.144	0.153	0.149	0.119	0.124	0.114
RE	1.153	1.155	1.232	0.213	0.218	0.233	0.015	0.015	0.000	0.579	0.564	0.554
Median group	1.038	1.038	1.016	0.337	0.342	0.450	0.030	0.030	0.059	0.124	0.119	0.099
Optimal combination												
Naive	0.975	0.974	0.951	0.936	0.936	0.950	0.317	0.297	0.243	0.000	0.000	0.000
Bias adj.	0.973	0.971	0.964	0.678	0.673	0.688	0.074	0.074	0.045	0.000	0.000	0.005
Pre-test												
PF	1.000	1.000	1.003	0.356	0.485	0.465	0.030	0.030	0.000	0.0.03	0.025	0.045
Shrinkage												
Prior lik.	0.991	0.989	0.951	0.574	0.554	0.723	0.069	0.059	0.069	0.054	0.020	0.000
Bayes.	0.982	0.980	0.931	0.644	0.683	0.807	0.114	0.104	0.144	0.000	0.000	0.000
Emp. Bayes.	0.994	0.996	0.927	0.584	0.550	0.782	0.173	0.188	0.257	0.035	0.094	0.000

Note: The column denoted 'AR' reports the results for a purely autoregressive specification, 'AR-PC' adds the first principle component of the panel of sub-indices, and 'AR-X' additionally the default yield and the term spread to the model. The absolute MSFEs of the individual forecasts are multiplied by 10^5. For further details see the footnote of Table 3.

denoted AR-PC; and (iii) the AR-PC model augmented with the lagged default yield and lagged term spread, denoted AR-X.

The first three columns of Table 4 show that, in contrast to the previous application, forecasts from individual estimation are now more accurate, on average, than the pooled, random effects, and, to a lesser extent, median group forecasts. These methods have a higher median MSFE and beat the benchmark only for about one-third of the indices. Pooled forecasts yield the smallest MSFE (14%) only slightly more often than they yield the largest ones (12%). Random effects forecasts, in contrast, rarely have the smallest MSFE and now generate the largest MSFE values for a majority of indices (55-57%). Median group forecasts are most accurate for 3-6% of the series but least accurate for 10-12% of the series, so this approach is not adapting particularly well to the high degree of heterogeneity observed across the models estimated on the CPI data–a finding that is consistent with the Monte Carlo simulations.

Forecast combinations produce a more precise median MSFE value than both the individual and pooled forecasts. They beat the benchmark for up to 95% of the indices and are the most precise forecast in up to 31.7% of the cases. Remarkably, forecast combinations based on the naive weights do not generate the largest MSFE value for any of the price indices. Overall, the

naive weights produce more precise forecasts than the bias-corrected weights, particularly in terms of their performance at the level of the individual CPI sub-indices.

For this application, the pre-test forecasts frequently select the individual-specific forecasts, consistent with there being a high level of heterogeneity in the forecasting models fitted on CPI sub-indices.[5] These forecasts improve on the predictive accuracy of the benchmark for between 35.6% and 48.5% of the price indices and produce the smallest MSFE values about as often as they generate the largest MSFE (3%) for the AR and AR-PC models. For the AR-X model, the pre-test approach performs a little worse, however, failing to generate the most accurate forecast for any of the sub-indices and generating the worst forecast for 4.5% of the series.

The prior likelihood, Bayesian, and empirical Bayes estimators again deliver forecasts with a lower median MSFE than both the pooled and individual forecasts. They beat the benchmark for between 57.4% and 80.7% of the price indices, have the lowest MSFE value for between 5.9% and 25.7% of indices and the highest MSFE for between 0 and 9.4% of indices.

Overall, the Bayesian forecasts perform best among these shrinkage approaches, both in terms of median MSFE, the frequency at which the forecasts beat the benchmark, and the ability to produce forecasts with the lowest MSFE without also increasing the risk of generating the least accurate forecasts. The ordering of the forecasts is not completely uniform, however. For example, the empirical Bayes approach works very well for the AR-X model.

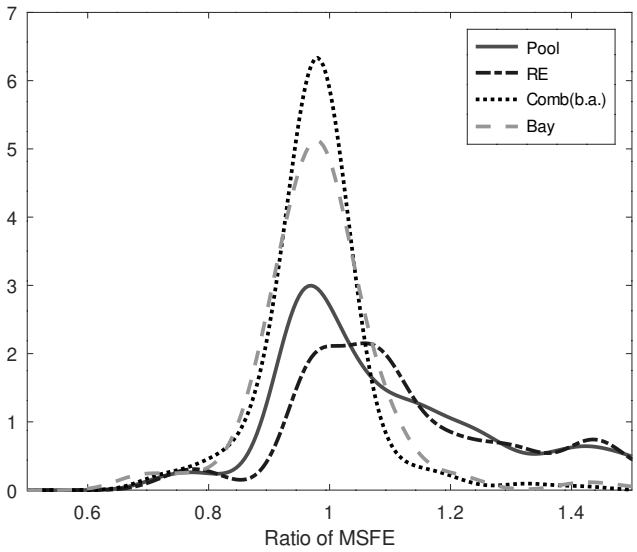
Our evidence is summarized in Figure 3 which shows densities of the MSFE ratios across the different sub-indices as produced by the AR model. As in the previous application, the forecast combination and Bayesian approaches have by far the smallest dispersion, consistent with these methods producing relatively few cases with the best or worst MSFE performance. The pooled forecasts have a thick right tail, indicating severe underperformance (median MSFE ratios above one) for a large proportion of price indices. The random effects approach performs even worse in this application with few MSFE ratios below one and a very thick right tail.

4.4 Individual stock returns

Our final application considers a panel of 23,121 individual US firm-level stock returns recorded at the monthly frequency. Our data sample runs from

[5]The median values of the PF test statistic for the AR, AR-PC and AR-X models (8.29, 7.91, and 3.13) support using the individual forecasts for the CPI sub-indices.

Figure 3: Density of ratios of MSFE: CPI inflation



Note: The graph shows density plots of the ratios of the MSFE over the CPI sub-indices for selected forecasts (pooled, random effects, bias-adjusted combination, and Bayesian) for the AR model. The density estimates use a normal kernel with bandwidth 0.04.

January 1977 through December 2017. We use a rolling estimation window of 120 observations, so our out-of-sample forecasts span the 31-year period from January 1987 through December 2017, corresponding to 372 monthly forecasts per stock. We require the panel for each estimation window to be balanced. In practice, this means that, in a given period, between 1,116 and 2,726 stocks are included in the panels we use for estimation, the average being 2,151 stocks.

Our prediction model is based on univariate regressions of the form

$$y_{i,t+1} = \alpha_i + \beta_i x_{it} + \varepsilon_{it+1}, \tag{41}$$

where x_{it} is the 6-month momentum of the particular stock, measured using cumulative returns up to the previous month. Return momentum is often identified as being among the best predictors of stock returns. While results will of course vary across predictors, similar findings hold for other predictors that we examined.

Table 5 reports our results. As in the other applications, the top row shows results for the model estimated on individual stocks. In the second

Table 5: Stock market results

Forecast method	Median R-squared	freq. beating prevail.mean	freq. beating individual	freq. smallest MSFE	freq. largest MSFE
Individual	-1.399	0.352	–	0.088	0.583
Prevailing mean	–	–	0.648	0.164	0.089
Pooled	0.232	0.609	0.648	0.106	0.018
RE	-0.220	0.439	0.647	0.200	0.206
Median Group	0.302	0.623	0.648	0.151	0.011
Optimal combination					
Naive	−0.244	0.414	0.647	0.049	0.002
Bias adj.	0.184	0.577	0.648	0.117	0.026
Pre-test					
PF	0.232	0.609	0.648	0.106	0.018
Shrinkage					
Prior lik.	0.274	0.602	0.648	0.136	0.074
Bayes.	0.295	0.622	0.648	0.033	0.002
Emp. Bayes.	0.295	0.623	0.648	0.003	0.000

Note: The table reports the median out-of-sample R^2, defined as 1 minus the ratio of MSFE of the respective method over that of the prevailing mean forecast, such that positive values correspond to MSFE ratios less than one. For further details see the footnote of Table 3.

row we now present results for the so-called prevailing mean model that uses the historical average computed on data up to the forecast date, i.e., model (41) constrained to set $\beta_i = 0$. Welch and Goyal (2008) find that this benchmark is very difficult to outperform in out-of-sample forecasts for the aggregate stock market, while Gu, Kelly and Xiu (2020) find that a simple forecast of zero is difficult to improve upon for individual stocks.

In line with the empirical finance literature, the first column reports out-of-sample R^2 values rather than MSFE ratios. The out-of-sample R^2 measure proposed by Campbell and Thompson (2008) is given by

$$R^2_{ij} = 1 - \frac{(T-T_1)^{-1}\sum_{t=T_1}^{T-1}(y_{i,t+1} - \hat{y}_{i,j,t+1|t})^2}{(T-T_1)^{-1}\sum_{t=T_1}^{T-1}(y_{i,t+1} - \bar{y}_{i,t+1|t})^2} = 1 - \frac{MSFE_{ij}}{MSFE_{iB}}.$$

where $\bar{y}_{it+1|t}$ is the prevailing mean (historical average) forecast, $i = 1, 2, \ldots, N$ denotes the stock and j the forecasting method with corresponding forecast $\hat{y}_{i,j,t+1|t}$. As shown after the second equation sign, the out-of-sample R^2 measure is a simple transformation of the MSFE ratio for method j ($MSFE_{ij}$) measured relative to that of the benchmark ($MSFE_{iB}$) with positive values corresponding to MSFE ratios less than one. As R^2 values are very small

in this application, we report them in percentage terms so that, e.g., 0.5% is equivalent to 0.005.

Columns two and three report the frequency with which a given forecasting method produces lower MSFE values than the prevailing mean (column 2) or the individual forecasts (column 3), while columns four and five report the proportion of stocks for which a method produces the lowest or highest MSFE values, respectively.

Table 5 shows that the pooled, median group, bias-corrected optimal combination, pre-testing, and three shrinkage forecasts generate median R^2 values between 0.184% and 0.302%. These forecasts also beat the prevailing mean forecasts for between 57.7% and 62.3% of the stocks and further beat the individual forecasts for nearly 65% of the stocks. Conversely, the individual, random effects and naive combination weights produce negative R^2 values.

This application effectively has a very low level of parameter heterogeneity due to the extremely low predictive power of the regressor which means that even large variation in slope coefficients do not translate into very different forecasts. Consistent with this, the pre-test always selects the pooled over the individual specific method.[6]

To get a sense of the significance of these results, note that a large literature in finance finds that univariate regression models such as that in (41) often produce higher out-of-sample MSFE-values than the prevailing mean model (Goyal and Welch, 2008, Rapach et al., 2010). This happens because stock returns are very difficult to predict due to the low signal-to-noise ratio in predictive return regressions and high levels of parameter estimation error which tends to dominate any predictive signal and so produces negative R^2 values. Campbell and Thompson (2008) indicate that a monthly out-of-sample R^2 value of 0.5% or higher can be exploited for sizeable economic gains for mean-variance investors with moderate levels of risk aversion. Their analysis is for aggregate stock market returns, which tend to be less volatile than individual stock-level returns. Viewed in this context, many of the panel forecasting approaches perform quite well.

Individual return forecasts produce the lowest MSFE for only 8.8% of the stocks, and produce the highest MSFEs for 58.3% of the stocks, suggesting that this approach has a very unattractive risk profile across the population of stocks. The traditional benchmark forecasts from the prevailing mean model perform much better here, producing the most accurate forecasts for

[6]The median value of the PF test statistic is -19.85, strongly favoring pooling for the stock market application.

16.4% of the stocks and the least accurate forecasts for only 8.9% of the stocks. The pooled forecasts are more conservative, producing the smallest MSFE for 10.6% of stocks and the largest value for only 1.8% of cases. The random effect forecasts produce both the most accurate and the least accurate forecasts for 20% of the stocks. Similar to the pooled forecasts, the median group, combination, and pre-test forecasts generate a substantially higher proportion of stocks with the smallest MSFE than the proportion of stocks with the largest MSFE value. Specifically, the median group, bias-corrected optimal combination and pre-test forecasts produce the best forecasts for 10-15% of the stocks and only generate the worst forecasts for less than 3% of the stocks.

The shrinkage forecasts produce more accurate forecasts than both the prevailing mean and individual forecasts for between 60 and 65% of the stocks. While the prior likelihood approach produces the most accurate forecasts for 13.6% of the stocks and the worst forecasts for 7.4% of the stocks, these rates are much lower for the Bayes and empirical Bayes methods.

Figure 4 shows density plots for the cross-section of R^2 values measured relative to the prevailing mean model. The individual forecasts are centered furthest to the left–recall that this is an undesirable feature for the R^2 measure–and have the highest dispersion followed by the random effects forecasts. The distributions of the pooled, combination, and Bayesian forecasts have densities with similar dispersion that are centered further to the right, indicating larger average R^2 values and, thus, better out-of-sample forecasting performance.

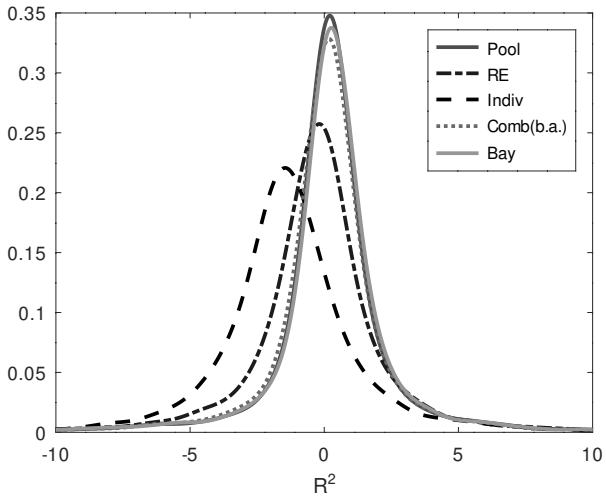
4.5 A decomposition of MSFE values

We, finally, propose a simple decomposition of the MSFE values for a better understanding of the trade-off offered by the different forecasting methods in different empirical applications. To isolate the effects of estimation uncertainty in the proposed decomposition we use individual-based forecasts fitted on the out-of-sample test period as the benchmark. However, the decomposition is completely general and can readily be used with alternative benchmarks.

Let $\overline{\boldsymbol{\beta}}_i$ be the estimate of $\boldsymbol{\beta}_i$ from the individual regressions fitted on the out-of-sample period, $t = T_1 + 1, T_1 + 2, \ldots, T$, and let $\overline{\varepsilon}_{it}$ be the associated (in-sample) residual at time t

$$y_{it} = \overline{\boldsymbol{\beta}}_i' \boldsymbol{x}_{it} + \overline{\varepsilon}_{it} \equiv \overline{y}_{it} + \overline{\varepsilon}_{it}, \tag{42}$$

Figure 4: Density of ratios of R^2: Stock market data



Note: The graph shows density plots of the out-of-sample R^2 for selected forecasts (pooled, random effects, bias-adjusted combination, and Bayesian). The density estimates use a normal kernel with bandwidth 0.5.

such that $\sum_{t=T_1+1}^{T} \overline{\varepsilon}_{it} = 0$, assuming that an intercept is always included in the forecasting model. Fitting the individual model only on the data from the out-of-sample period accomplishes two things. First, it ensures that the forecasts are unbiased and the residuals have zero mean. Second, the MSE computed from the residuals in (42) is an infeasible lower bound against which we can measure the MSFE performance of other forecasting methods; it is infeasible because the regression uses information from future observations to estimate $\boldsymbol{\beta}_i$.[7]

To compare the infeasible forecast in (42) to forecasts from the other methods, let $\widehat{\boldsymbol{\beta}}_{it}$ be the period-t (real-time) estimate of $\boldsymbol{\beta}_i$ from one of the methods under consideration, based on information up to time t, while $\widehat{\varepsilon}_{it}$ is the associated forecast error

$$y_{it} = \widehat{\boldsymbol{\beta}}_{it}' \boldsymbol{x}_{it} + \widehat{\varepsilon}_{it} \equiv \widehat{y}_{it} + \widehat{\varepsilon}_{it}, \quad t = T_1 + 1, T_1 + 2, \ldots, T. \tag{43}$$

Taking the difference between (42) and (43), we have

$$\widehat{\varepsilon}_{it} = \overline{y}_{it} - \widehat{y}_{it} + \overline{\varepsilon}_{it} \equiv \Delta \widehat{y}_{it} + \overline{\varepsilon}_{it},$$

[7]Note also that the sample estimate of the individual model's MSFE over the period $t = T_1 + 1, T_1 + 2, \ldots, T$ is biased downwards relative to the population MSFE value.

where we defined $\Delta\widehat{y}_{it} = \overline{y}_{it} - \widehat{y}_{it}$. This yields the following decomposition:

$$\begin{aligned}\frac{1}{T-T_1}\sum_{t=T_1+1}^{T}\widehat{\varepsilon}_{it}^2 &= \frac{1}{T-T_1}\sum_{t=T_1+1}^{T}\Delta\widehat{y}_{it}^2 + \frac{1}{T-T_1}\sum_{t=T_1+1}^{T}\overline{\varepsilon}_{it}^2 \\ &\quad + \frac{2}{T-T_1}\sum_{t=T_1+1}^{T}\Delta\widehat{y}_{it}\overline{\varepsilon}_{it}.\end{aligned} \tag{44}$$

Subtracting the MSFE from the benchmark model from both sides of (44), we have

$$\begin{aligned}&\frac{1}{T-T_1}\sum_{t=T_1+1}^{T}\widehat{\varepsilon}_{it}^2 - \frac{1}{T-T_1}\sum_{t=T_1+1}^{T}\overline{\varepsilon}_{it}^2 \\ =\ &\frac{1}{T-T_1}\sum_{t=T_1+1}^{T}\Delta\widehat{y}_{it}^2 + \frac{2}{T-T_1}\sum_{t=T_1+1}^{T}\Delta\widehat{y}_{it}\overline{\varepsilon}_{it} \\ =\ &\frac{1}{T-T_1}\sum_{t=T_1+1}^{T}\left(\Delta\widehat{y}_{it} - \hat{\mu}_{\widehat{y}_i}\right)^2 + \hat{\mu}_{\widehat{y}_i}^2 + \frac{2}{T-T_1}\sum_{t=T_1+1}^{T}\Delta\widehat{y}_{it}\overline{\varepsilon}_{it},\end{aligned} \tag{45}$$

where $\hat{\mu}_{\widehat{y}_i} = (T-T_1)^{-1}\sum_{t=T_1+1}^{T}\Delta\widehat{y}_{it}$, is the estimated bias of $\widehat{y}_{it}$ relative to that of $\overline{y}_{it}$ in the test sample. By construction, $\overline{y}_{it}$ has zero bias in the test sample so $\hat{\mu}_{\widehat{y}_i}$ is also equal to the test-sample bias of $\widehat{y}_{it}$. Moreover, $(T-T_1)^{-1}\sum_{t=T_1+1}^{T}\left(\Delta\widehat{y}_{it} - \hat{\mu}_{\widehat{y}_i}\right)^2$ can loosely be interpreted as variance originating from estimation error.

Estimates of the out-of-sample MSFE measured relative to the MSFE of the individual forecasts (with parameters estimated on the out-of-sample period) can therefore be decomposed into the variance of the forecast differential (relative to the benchmark) plus the squared sample bias of the forecasting method plus two times the covariance between the forecast differential and the baseline model's forecast error.

Note that (45) will be positive because the $\overline{y}_{it}$ forecasts are computed using a model whose parameters are optimized on the test sample $t = T_1 + 1, T_1+2, \ldots, T$. Across different forecasting methods, the relative size of the variance component depends on whether estimation uncertainty (which, for example, is $O_p(N^{-1})$ under pooling) outweighs parameter heterogeneity or if the opposite holds.

Table 6 shows the outcome of this decomposition for our three empirical applications in the form of median (across units) values of the three com-

Table 6: Decompositions of the forecasts

Application	CPI data								
	AR			AR-PC			AR-X		
	Var	2Cov	E^2	Var	2Cov	E^2	Var	2Cov	E^2
Individual	0.283	−0.137	0.000	0.285	−0.083	0.001	0.417	−0.080	0.002
Pooled	0.306	−0.122	0.016	0.280	−0.077	0.015	0.392	−0.057	0.019
RE	0.268	−0.068	0.202	0.274	−0.053	0.193	0.500	−0.036	0.175
Median group	0.235	−0.148	0.048	0.220	−0.106	0.046	0.252	−0.096	0.044
Optimal combination									
Naive	0.208	−0.158	0.001	0.210	−0.105	0.001	0.283	−0.097	0.002
Bias adj.	0.220	−0.143	0.002	0.218	−0.109	0.002	0.319	−0.106	0.003
Pre-test									
PF	0.282	−0.122	0.001	0.279	−0.080	0.001	0.386	−0.079	0.002
Shrinkage									
Prior lik	0.203	−0.103	0.002	0.194	−0.065	0.003	0.267	−0.074	0.005
Bayes	0.186	−0.121	0.003	0.182	−0.072	0.003	0.241	−0.080	0.005
Emp Bayes	0.184	−0.094	0.003	0.192	−0.058	0.004	0.243	−0.080	0.005
Application	House price data						Stock prices		
	SAR			SARX					
	Var	2Cov	E^2	Var	2Cov	E^2	Var	2Cov	E^2
Individual	0.274	0.083	0.008	0.579	0.074	0.018	0.933	2.018	0.130
Pooled	0.241	0.003	0.002	0.577	-0.022	0.011	0.429	0.171	0.332
RE	0.201	0.009	0.001	0.506	-0.001	0.002	0.450	0.246	0.947
Median group	0.182	0.007	0.006	0.437	-0.032	0.006	0.425	0.142	0.356
Optimal combination									
Naive	0.239	0.057	0.006	0.522	0.047	0.013	0.449	1.005	0.157
Bias adj.	0.236	0.042	0.005	0.514	0.034	0.011	0.467	0.260	0.279
Pre-test									
PF	0.292	0.027	0.005	0.630	-0.059	0.013	0.429	0.171	0.332
Shrinkage									
Prior lik	0.185	0.045	0.006	0.513	0.001	0.015	0.437	0.159	0.348
Bayes	0.195	0.019	0.004	0.475	0.004	0.008	0.426	0.151	0.340
Emp Bayes	0.191	0.019	0.005	0.462	0.014	0.007	0.426	0.151	0.340

Note: The table reports the variance, twice the covariance, and squared expectated bias given in (45) estimated in the respective forecast samples. The results for the CPI and stock market data are scaled up by 10^5 and 10^4 for readability.

ponents of (45).[8] The layout of this table is similar to that of the earlier tables, except that the top row (individual) now decomposes the MSFE per-

[8]Because medians are computed separately for each of the three components, the reported values cannot be added up to get the total MSFE ratio of the median unit.

formance of the individual forecasts with recursively estimated parameters relative to the individual forecasts whose parameters are estimated on the test sample and so, by construction, produce smaller MSFE values.

Starting with the house price application (lower left quadrant), the individual forecasts and pre-tests generate by far the highest median variance component, while the median group estimator, optimal combination and shrinkage forecasts are at the opposite end of the spectrum, producing far lower variance terms.

The individual method also produces a far higher positive covariance term than the other approaches. Among the remaining methods, the forecast combinations generate quite high covariances, while the pooled, random effect, median group and Bayesian shrinkage approaches produce very low covariance values. Finally, the squared bias term is small for all approaches in this application.

These results show that the individual forecasting method performs relatively poorly in the housing application due to the twin effects of estimation errors that have high variance and are strongly correlated with the forecast errors. Conversely, the median group estimator, combinations, and shrinkage approaches produce relatively accurate forecasts as a result of their small variance and covariance terms.

Turning to the CPI inflation forecasts (top panel), the variance component is notably smaller for the median group, combination and shrinkage methods compared to especially the individual, pooled and pre-test methods. The optimal combinations, along with the median group estimator, also tend to have slightly larger negative covariance terms which helps to reduce their MSFE values.

The random effects approach achieves a substantially smaller reduction in MSFE performance from the covariance term than the other methods. Moreover, the random effects forecasts stand out as being highly biased, which helps explain their very poor performance for the CPI application.

Conversely, the good CPI forecasting performance of the median group, combination and shrinkage approaches reflects both their small variance and larger negative covariance components.

Finally, in the stock price application, the variance and covariance components for the individual forecasting method are both very large – twice as large as those of the second-highest method. While the individual method generates a much smaller squared bias than the other approaches, this term is generally quite small and so gets dominated by the larger variance and covariance terms. The exception to this pattern is the random effects forecasts whose very large bias explains their poor performance for stock returns.

The median group, pooling, bias-corrected combination and shrinkage approaches all perform quite well in this application because they manage to keep all three MSFE components relatively small.

These results suggest that the accuracy of different panel forecasting methods generally hinges on how high the variance of their estimation errors is and how strongly they correlate with forecast errors. The only method for which the squared bias term seems to matter a great deal is the random effects estimator which in some settings produces strongly biased forecasts.

5 Conclusion

We provide a comprehensive examination of the out-of-sample predictive accuracy of a large set of existing and novel panel forecasting methods, including individual estimation, pooling, random effects, median group, optimal combination, pre-test, and (Bayesian) shrinkage. Our analysis characterizes analytically the determinants of squared error performance as it relates to (squared) bias and estimation error variance components. We show how parameter heterogeneity, predictive power, and sample sizes regulate the bias-variance trade-off that determines predictive accuracy. To illustrate these theoretical points, we consider three empirical applications for house prices, CPI inflation, and stock market returns.

Our main findings can be summarized in three points. First, we generally find that a number of panel approaches perform systematically better than individual forecasts computed equation by equation. Our empirical applications to three very different data sets demonstrate sizeable gains from exploiting panel information to obtain forecasts that are more accurate both on average, across units, and also for the majority of individual units.

Second, our analytical results and Monte Carlo simulations show that one should not expect a single forecasting approach to be uniformly dominant across applications that differ in terms of the cross-sectional and time-series dimensions, strength of predictive power, and, most importantly, degree of heterogeneity in intercept and slope coefficients.

This point is validated in our empirical analysis. For example, pooling produces notably better forecasts than individual estimation in the house price application, while conversely individual forecasts perform much better in the CPI inflation application. For both panels, we find that the combined forecasts and (Bayesian) shrinkage forecasts are more precise than the pooled and individual forecasts, sometimes by a substantial margin. A novel pre-test method that chooses between pooled and individual estimation gen-

erally produces forecasts that are more precise than either of the underlying forecasting methods.

For the application to stock market returns where pooling is very hard to beat, the pre-test nearly always correctly chooses pooling. The accuracy of the pre-test method that includes the combined forecast is similar to that of a bias-corrected optimal combined forecast. Moreover, shrinkage forecasts also deliver relatively accurate forecasts. These methods all beat prevailing mean and individual forecasts for the majority of stocks.

Third, while it is not possible to pinpoint a single universally dominant approach to panel forecasting, the methods clearly differ in terms of their risk profiles, particularly their ability to reduce the probability of generating very poor forecasts for individual units in a cross-section. While the individual, pooled, random effect, and median group methods perform very poorly in at least one of our empirical applications, forecast combination and Bayesian shrinkage methods typically have a very small chance of being the worst method for individual units, while at the same time retaining some probability of being the best method.

In a nutshell, our simulations and empirical applications suggest that forecast combinations and shrinkage methods offer insurance against poor performance. Forecast combinations, in particular, perform well across the board, while the performance of shrinkage methods tends to vary a bit more across applications. Compared to the alternative forecasting methods we consider, this better "risk-return" trade-off makes the combination and shrinkage methods attractive in forecast applications with panel data.

References

Baltagi, B.H. (2008) "Forecasting with panel data" *Journal of Forecasting* 27, 153–173.

Baltagi, B.H. (2013) "Panel data forecasting" Ch. 18 in Elliott, G. and A. Timmermann, *Handbook of Economic Forecasting*, volume 2B. North Holland: Elsevier.

Boot, T., and A. Pick (2020) "Does modeling a structural break improve forecast accuracy?" *Journal of Econometrics* 215, 35–59.

Brückner, H., and B. Siliverstovs (2006) "On the estimation and forecasting of international migration: how relevant is heterogeneity across countries" *Empirical Economics* 31, 735–754.

Campbell, J.Y. and S.B. Thompson (2008) "Predicting Excess Stock Returns out of Sample: Can Anything Beat the Historical Average?" *Review of Financial Studies* 21, 1509–1531.

Goldberger, A.S. (1962) "Best linear unbiased prediction in the generalized linear regression model" *Journal of the American Statistical Association* 57, 369–375.

Goyal, A., and I. Welch (2008) "A comprehensive look at the empirical performance of equity premium prediction" *Review of Financial Studies* 21, 1455–1508.

Gu, S., B. Kelly, and D. Xiu (2020) "Empirical Asset Pricing via Machine Learning" *Review of Financial Studies* 33, 2223–2273.

Hsiao, C. (2022) *Analysis of Panel Data*, 4th ed., Cambridge University Press (forthcoming).

Lee, L.-F., and W.E. Griffiths (1979) "The prior likelihood and best linear unbiased prediction in stochastic coefficient linear models" Center for Economic Research, University of Minnesota, Discussion paper 79–107.

Liu, L., H.R. Moon, and F. Schorfheide (2020) "Forecasting with dynamic panel data models" *Econometrica* 88, 171–201.

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

Pesaran, H.M. (2015) *Time Series and Panel Data Econometrics*, Oxford University Press.

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

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

Rapach, D., J.K. Strauss, and G. Zhou (2010) "Out-of-Sample Equity Premium Prediction: Combination Forecasts and Links to the Real Economy" *Review of Financial Studies* 23, 821–862.

Timmermann, A. (2006) "Forecast combinations" Ch. 4 in G. Elliott, C. W. J. Granger and A. Timmermann (Eds.) *Handbook of Economic Forecasting*, volume. 1, North Holland: Elsevier.

Trapani, L. and G. Urga (2009) "Optimal forecasting with heterogeneous panels: A Monte Carlo Study" *International Journal of Forecasting* 25, 567–586.

Wang, W., X. Zhang, and R. Paap (2019) "To pool of not to pool: What is a good strategy for parameter estimation and forecasting in panels" *Journal of Applied Econometrics* 34, 724–745.

Welch, I. and A. Goyal (2008) "A Comprehensive Look at The Empirical Performance of Equity Premium Prediction" *Review of Financial Studies* 21, 1455–1508.

Yang, Cynthia F. (2021) "Common Factors and Spatial Dependence: An Application to US House Prices" *Econometric Reviews* 40, 14–50.

Online Appendix

Appendix A Mathematical derivations

Note that for simplicity we drop the fact that we condition on $\boldsymbol{X}$ and $\boldsymbol{x}_{i,T+1}$ from the notation. However, all mathematical expectations derived are conditional, which is justified given Assumptions 1–6.

A.1 MSFE values for pooled and individual estimation

Using (3), under (1) we have

$$\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i = \left(\boldsymbol{X}_i'\boldsymbol{X}_i\right)^{-1}\boldsymbol{X}_i'\boldsymbol{\varepsilon}_i,$$

and the first result in Proposition 1 follows since under Assumptions 1-3 $\mathrm{E}\left(\boldsymbol{\varepsilon}_i\right) = \mathbf{0}$, and $\mathrm{Var}\left(\boldsymbol{X}_i'\boldsymbol{\varepsilon}_i\right) = \sigma_i^2\left(\boldsymbol{X}_i'\boldsymbol{X}_i\right)$.

To prove the second part of Proposition 1, using (5) we note that

$$\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_i = \boldsymbol{b}_{i,NT} + \boldsymbol{\xi}_{NT}, \tag{46}$$

where

$$\boldsymbol{b}_{i,NT} = N^{-1}\sum_{j=1}^{N}\boldsymbol{S}_{j,NT}\,\boldsymbol{\eta}_j - \boldsymbol{\eta}_i, \text{ and } \boldsymbol{\xi}_{NT} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\sum_{j=1}^{N}\boldsymbol{X}_j'\boldsymbol{\varepsilon}_j, \tag{47}$$

and

$$\boldsymbol{S}_{j,NT} = \left(\frac{\boldsymbol{X}'\boldsymbol{X}}{NT}\right)^{-1}\left(\frac{\boldsymbol{X}_j'\boldsymbol{X}_j}{T}\right).$$

By Assumption 5, $\boldsymbol{\varepsilon}_i$ and $\boldsymbol{\eta}_j$ are distributed independently for all i and j, so

$$\mathrm{Var}(\tilde{y}_{i,T+1}) = \sigma_i^2 + \boldsymbol{x}_{i,T+1}'\left[\mathrm{Var}(\boldsymbol{b}_{i,NT}) + \mathrm{Var}(\boldsymbol{\xi}_{NT})\right]\boldsymbol{x}_{i,T+1}. \tag{48}$$

Also, under Assumption 4 and 5, $\mathrm{E}\left(\boldsymbol{b}_{i,NT}\right) = \mathbf{0}$, and $\mathrm{Var}(\boldsymbol{b}_{i,NT}) = \mathrm{E}\left(\boldsymbol{b}_{i,NT}\boldsymbol{b}_{i,NT}'\right)$ and we have

$$\mathrm{Var}(\boldsymbol{b}_{i,NT}) = \mathrm{E}\left[\left(N^{-1}\sum_{j=1}^{N}\boldsymbol{S}_{j,NT}\boldsymbol{\eta}_j - \boldsymbol{\eta}_i\right)\left(N^{-1}\sum_{l=1}^{N}\boldsymbol{S}_{l,NT}\boldsymbol{\eta}_l - \boldsymbol{\eta}_i\right)'\right].$$

Since, conditional on $\boldsymbol{X}$, $\boldsymbol{S}_{j,NT}$ is given,

$$\begin{aligned}\mathrm{Var}(\boldsymbol{b}_{i,NT}) &= \mathrm{E}\left(\boldsymbol{\eta}_i\boldsymbol{\eta}_i'\right) + N^{-2}\sum_{j=1}^{N}\sum_{j'=1}^{N}\boldsymbol{S}_{j,NT}\mathrm{E}\left(\boldsymbol{\eta}_j\boldsymbol{\eta}_{j'}'\right)\boldsymbol{S}_{j',NT}' \\ &\quad -N^{-1}\sum_{j=1}^{N}\boldsymbol{S}_{j,NT}\mathrm{E}\left(\boldsymbol{\eta}_j'\boldsymbol{\eta}_i\right) - N^{-1}\sum_{j=1}^{N}\mathrm{E}\left(\boldsymbol{\eta}_i\boldsymbol{\eta}_j'\right)\boldsymbol{S}_{j,NT}.\end{aligned}$$

Let $\mathrm{E}\left(\boldsymbol{\eta}_i\boldsymbol{\eta}_j'\right) = \boldsymbol{\Omega}_{ij}$, and note that $\mathrm{E}\left(\boldsymbol{\eta}_i\boldsymbol{\eta}_i'\right) = \boldsymbol{\Omega}_{\eta}$, $\boldsymbol{\Omega}_{ij}' = \boldsymbol{\Omega}_{ji}$, and $\|\boldsymbol{\Omega}_{ij}\| = \|\boldsymbol{\Omega}_{ji}\|$. Then

$$\begin{aligned}\mathrm{Var}(\boldsymbol{b}_{i,NT}) &= \boldsymbol{\Omega}_{\eta} + N^{-2}\sum_{j=1}^{N}\sum_{j'=1}^{N}\boldsymbol{S}_{j,NT}\,\boldsymbol{\Omega}_{jj'}\boldsymbol{S}_{j',NT}' \\ &\quad -N^{-1}\sum_{j=1}^{N}\boldsymbol{S}_{j,NT}\boldsymbol{\Omega}_{ji} - N^{-1}\sum_{j=1}^{N}\boldsymbol{\Omega}_{ij}\boldsymbol{S}_{j,NT},\end{aligned}$$

which can be written more compactly as

$$\mathrm{Var}(\boldsymbol{b}_{i,NT}) = \boldsymbol{\Omega}_{\eta} + N^{-1}\left(\boldsymbol{A}_{NT} + \boldsymbol{B}_{NT} + \boldsymbol{B}_{NT}'\right), \tag{49}$$

where

$$\begin{aligned}\|\boldsymbol{A}_{NT}\| &= \left\|N^{-1}\sum_{j=1}^{N}\sum_{j'=1}^{N}\boldsymbol{S}_{j,NT}\,\boldsymbol{\Omega}_{jj'}\boldsymbol{S}_{j',NT}'\right\| \\ &\leq N^{-1}\sum_{j=1}^{N}\sum_{j'=1}^{N}\|\boldsymbol{S}_{j,NT}\|\left\|\boldsymbol{S}_{j',NT}\right\|\left\|\boldsymbol{\Omega}_{jj'}\right\| \\ &\leq \left(\sup_j\|\boldsymbol{S}_{j,NT}\|\right)\left(\sup_{j'}\left\|\boldsymbol{S}_{j',NT}\right\|\right)\left(N^{-1}\sum_{j=1}^{N}\sum_{j'=1}^{N}\left\|\boldsymbol{\Omega}_{jj'}\right\|\right) \\ &\leq \left(\sup_j\|\boldsymbol{S}_{j,NT}\|\right)^2\sup_j\left(\sum_{j'=1}^{N}\left\|\boldsymbol{\Omega}_{jj'}\right\|\right).\end{aligned}$$

Further,

$$\sup_j\|\boldsymbol{S}_{j,NT}\| \leq \left\|\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{NT}\right)^{-1}\right\|\sup_j\left\|\frac{\boldsymbol{X}_j'\boldsymbol{X}_j}{T}\right\|,$$

and, under Assumptions 3 and 6,

$$\left\|\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{NT}\right)^{-1}\right\| = O_p(1), \quad \sup_j \left\|\frac{\boldsymbol{X}_j'\boldsymbol{X}_j}{T}\right\| = O_p(1), \quad \sup_j \left(\sum_{j'=1}^{N} \left\|\boldsymbol{\Omega}_{jj'}\right\|\right) < C.$$

Hence, $\|\boldsymbol{A}_{NT}\| = O_p(1)$. Similarly,

$$\begin{aligned} \left\|\boldsymbol{B}_{NT} + \boldsymbol{B}_{NT}'\right\| &< 2\left\|\boldsymbol{B}_{NT}\right\| \leq 2\sum_{j=1}^{N} \left\|\boldsymbol{S}_{j,NT}\boldsymbol{\Omega}_{ji}\right\| \\ &\leq 2\sup_j \left\|\boldsymbol{S}_{j,NT}\right\| \sup_i \left(\sum_{j=1}^{N} \left\|\boldsymbol{\Omega}_{ji}\right\|\right), \end{aligned}$$

and we also have $\|\boldsymbol{B}_{NT} + \boldsymbol{B}_{NT}'\| = O_p(1)$. Using these results in (49) yields

$$\operatorname{Var}(\boldsymbol{b}_{i,NT}) = \boldsymbol{\Omega}_\eta + O_p\left(\frac{1}{N}\right). \tag{50}$$

Finally, under Assumptions 1–3, $\mathrm{E}\left(\boldsymbol{\xi}_{NT}\right) = \boldsymbol{0}$, and $\mathrm{E}\left(\boldsymbol{\varepsilon}_j\boldsymbol{\varepsilon}_{j'}'\right) = \boldsymbol{0}$ for $j \neq j'$ and $\mathrm{E}\left(\boldsymbol{\varepsilon}_j\ \boldsymbol{\varepsilon}_j'\right) = \sigma_j^2 < \infty$, and we have

$$\begin{aligned} \operatorname{Var}(\boldsymbol{\xi}_{NT}) = &\left(\tfrac{\boldsymbol{X}'\boldsymbol{X}}{NT}\right)^{-1} \tfrac{1}{N^2}\textstyle\sum_{j=1}^{N}\sum_{j'=1}^{N} \frac{\boldsymbol{X}_j'\mathrm{E}\left(\boldsymbol{\varepsilon}_j\boldsymbol{\varepsilon}_{j'}'\right)\boldsymbol{X}_{j'}}{T^2} \left(\tfrac{\boldsymbol{X}'\boldsymbol{X}}{NT}\right)^{-1} \\ = &\ \tfrac{1}{N}\left(\tfrac{\boldsymbol{X}'\boldsymbol{X}}{NT}\right)^{-1} \left(\tfrac{1}{N}\textstyle\sum_{j=1}^{N}\sigma_j^2\frac{\boldsymbol{X}_j'\boldsymbol{X}_j}{T}\right) \left(\tfrac{\boldsymbol{X}'\boldsymbol{X}}{NT}\right)^{-1} = O_p\left(\tfrac{1}{N}\right). \end{aligned} \tag{51}$$

The result in the second part of Proposition 1 now follows by using (50) and (51) in (48).

A.2 Combination weights

Consider ω_i^* given by (13) which we reproduce here for convenience:

$$\omega_i^* = \frac{\operatorname{Var}(\tilde{e}_{i,T+1}) - \operatorname{Cov}(\hat{e}_{i,T+1}, \tilde{e}_{i,T+1})}{\operatorname{Var}(\hat{e}_{i,T+1}) + \operatorname{Var}(\tilde{e}_{i,T+1}) - 2\operatorname{Cov}(\hat{e}_{i,T+1}, \tilde{e}_{i,T+1})}. \tag{52}$$

Using (8) and (9), we have

$$\begin{aligned} \operatorname{Var}(\hat{e}_{i,T+1}) + \operatorname{Var}(\tilde{e}_{i,T+1}) = &\ 2\sigma_i^2 + T^{-1}\sigma_i^2\boldsymbol{x}_{i,T+1}'\boldsymbol{Q}_{iT}^{-1}\boldsymbol{x}_{i,T+1} \\ &+\boldsymbol{x}_{i,T+1}'\ \boldsymbol{\Omega}_\eta\boldsymbol{x}_{i,T+1} + O_p\left(N^{-1}\right), \end{aligned} \tag{53}$$

where $\boldsymbol{Q}_{iT} = T^{-1}\boldsymbol{X}_i'\boldsymbol{X}_i$. Also, using (6) and (7) we have

$$\begin{aligned}\text{Cov}(\hat{e}_{i,T+1}, \tilde{e}_{i,T+1}) &= \text{Cov}\left[\varepsilon_{i,T+1} - (\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)'\boldsymbol{x}_{i,T+1}, \varepsilon_{i,T+1} - (\tilde{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)'\boldsymbol{x}_{i,T+1}\right] \\ &= \sigma_i^2 + \boldsymbol{x}_{i,T+1}'\text{E}\left[(\,\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_i)'\right]\boldsymbol{x}_{i,T+1}\end{aligned}$$

and, under Assumptions 1–3,

$$\begin{aligned}&\text{E}\left[(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_i)'\right] = \text{E}\left[\left(\boldsymbol{X}_i'\boldsymbol{X}_i\right)^{-1}\boldsymbol{X}_i'\boldsymbol{\varepsilon}_i\left(\boldsymbol{b}_{i,NT} + \boldsymbol{\xi}_{NT}\right)'\right] \\ &= \left(\,\boldsymbol{X}_i'\boldsymbol{X}_i\right)^{-1}\boldsymbol{X}_i'E\left(\boldsymbol{\varepsilon}_i\boldsymbol{\xi}_{NT}'\right) = \left(\boldsymbol{X}_i'\boldsymbol{X}_i\right)^{-1}\boldsymbol{X}_i'\text{E}\left(\boldsymbol{\varepsilon}_i\sum_{j=1}^{N}\boldsymbol{\varepsilon}_j'\boldsymbol{X}_j(\boldsymbol{X}'\boldsymbol{X})^{-1}\right) \\ &= \sigma_i^2\left(\boldsymbol{X}_i'\boldsymbol{X}_i\right)^{-1}\boldsymbol{X}_i'\boldsymbol{X}_i(\boldsymbol{X}'\boldsymbol{X})^{-1} = N^{-1}T^{-1}\sigma_i^2\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{TN}\right)^{-1} = O_p(N^{-1}).\end{aligned}$$

Hence

$$\text{Cov}(\hat{e}_{i,T+1}, \tilde{e}_{i,T+1}) = \sigma_i^2 + O_p(N^{-1}). \tag{54}$$

Using this result together with (53) we now have the following expression for the denominator of (52)

$$\begin{aligned}&\text{Var}(\hat{e}_{i,T+1}) + \text{Var}(\tilde{e}_{i,T+1}) - 2\text{Cov}(\hat{e}_{i,T+1}, \tilde{e}_{i,T+1}) \\ &\quad = T^{-1}\sigma_i^2\boldsymbol{x}_{i,T+1}'\boldsymbol{Q}_{iT}^{-1}\boldsymbol{x}_{i,T+1} + \boldsymbol{x}_{i,T+1}'\,\boldsymbol{\Omega}_\eta\boldsymbol{x}_{i,T+1} + O_p(N^{-1}).\end{aligned} \tag{55}$$

Similarly, using (48) and (54) the numerator of (52) is given by

$$\text{Var}(\tilde{e}_{i,T+1}) - \text{Cov}(\hat{e}_{i,T+1}, \tilde{e}_{i,T+1}) = \boldsymbol{x}_{i,T+1}'\boldsymbol{\Omega}_\eta\boldsymbol{x}_{i,T+1} + O_p(N^{-1}). \tag{56}$$

The result in Proposition 2 now follows by using (55) and (56) in (52).

A.3 Proof of Proposition 3

We first note that

$$\hat{\boldsymbol{\beta}}_i - \bar{\hat{\boldsymbol{\beta}}} = \boldsymbol{\eta}_i - \bar{\boldsymbol{\eta}}_N + \boldsymbol{\xi}_{iT} - \bar{\boldsymbol{\xi}}_{NT}, \tag{57}$$

where

$$\bar{\boldsymbol{\eta}}_N = N^{-1}\sum_{i=1}^{N}\boldsymbol{\eta}_i, \quad \boldsymbol{\xi}_{iT} = (\boldsymbol{X}_i'\boldsymbol{X}_i)\boldsymbol{X}_i'\boldsymbol{\varepsilon}_i, \quad \bar{\boldsymbol{\xi}}_{NT} = N^{-1}\sum_{i=1}^{N}\boldsymbol{\xi}_{iT}, \tag{58}$$

and

$$z_{i,NT} = \boldsymbol{x}'_{i,T+1}\left(\boldsymbol{\eta}_i - \bar{\boldsymbol{\eta}}_N\right) + \boldsymbol{x}'_{i,T+1}\left(\boldsymbol{\xi}_{iT} - \bar{\boldsymbol{\xi}}_{NT}\right).$$

Under the assumptions of the proposition and conditional on $\boldsymbol{x}_{i,T+1}$ and $\boldsymbol{X}_i$, $\boldsymbol{x}'_{i,T+1}\left(\boldsymbol{\eta}_i - \bar{\boldsymbol{\eta}}_N\right)$ and $\boldsymbol{x}'_{i,T+1}\left(\boldsymbol{\xi}_{iT} - \bar{\boldsymbol{\xi}}_{NT}\right)$ are distributed independently with zero means and variances $\frac{N-1}{N}\left(\boldsymbol{x}'_{i,T+1}\boldsymbol{\Omega}_\eta \boldsymbol{x}_{i,T+1}\right)$ and $\frac{N-1}{NT}\sigma_i^2\left(\boldsymbol{x}'_{i,T+1}\boldsymbol{Q}_{i,T}^{-1}\boldsymbol{x}_{i,T+1}\right)$, respectively. From the assumption that $\boldsymbol{\eta}_i$ and $\boldsymbol{\xi}_{iT}$ are normally distributed, we have that

$$\left[\mathrm{E}\left(z_{i,NT}^2\right)\right]^{-1/2} z_{i,NT} \sim \mathrm{N}\left(0,1\right), \text{ for } i = 1,2,\ldots,N, \tag{59}$$

where

$$\mathrm{E}\left(z_{i,NT}^2\right) = \left(1 - N^{-1}\right)\left[\boldsymbol{x}'_{i,T+1}\boldsymbol{\Omega}_\eta \boldsymbol{x}_{i,T+1} + T^{-1}\sigma_i^2\left(\boldsymbol{x}'_{i,T+1}\boldsymbol{Q}_{i,T}^{-1}\boldsymbol{x}_{i,T+1}\right)\right].$$

Then under the null hypothesis (19), we have

$$\mathrm{E}\left(z_{i,NT}^2\right) = 2(1 - N^{-1})T^{-1}\sigma_i^2\left(\boldsymbol{x}'_{i,T+1}\boldsymbol{Q}_{i,T}^{-1}\boldsymbol{x}_{i,T+1}\right).$$

Using (59), it now follows that $\left[\mathrm{E}\left(z_{i,NT}^2\right)\right]^{-1} z_{i,NT}^2$ is distributed as χ_1^2 and, hence, $\mathrm{Var}\left(\left[\mathrm{E}\left(z_{i,NT}^2\right)\right]^{-1} z_{i,NT}^2\right) = 2$, and under the null hypothesis

$$\mathrm{Var}\left(z_{i,NT}^2\right) = 2\left[\mathrm{E}\left(z_{i,NT}^2\right)\right]^2 = 2(1-N^{-1})^2\left[2T^{-1}\sigma_i^2\left(\boldsymbol{x}'_{i,T+1}\boldsymbol{Q}_{i,T}^{-1}\boldsymbol{x}_{i,T+1}\right)\right]^2,$$

as required. The result in (20) now follows from application of standard central limit theorems, since $\left[\mathrm{Var}(z_{i,NT}^2)\right]^{-1/2}\left(z_{i,NT}^2 - \mathrm{E}(z_{i,NT}^2)\right)$, $i = 1,2,\ldots,N$ are standardized random variables distributed independently over i.

Note also that for large N

$$\begin{aligned}
\frac{z_{i,NT}^2 - \mathrm{E}(z_{i,NT}^2)}{\left[\mathrm{Var}(z_{i,NT}^2)\right]^{1/2}} &= \frac{z_{i,NT}^2 - 2T^{-1}\sigma_i^2\left(\boldsymbol{x}'_{i,T+1}\boldsymbol{Q}_{i,T}^{-1}\boldsymbol{x}_{i,T+1}\right)}{\sqrt{2}\left[2T^{-1}\sigma_i^2\left(\boldsymbol{x}'_{i,T+1}\boldsymbol{Q}_{i,T}^{-1}\boldsymbol{x}_{i,T+1}\right)\right]} \\
&= \frac{\omega_{i,NT}^2 - 1}{\sqrt{2}},
\end{aligned}$$

where $\omega_{i,NT}^2$ is given by (21), as required.

A.4 Large N probability order of $\tilde{\sigma}^2_{i,NT}$

Consider the numerator of (22) and note that, again conditional on $\boldsymbol{X}_i$,

$$\mathrm{E}\left(\tilde{\sigma}^2_{i,NT}\right) = \frac{\mathrm{E}\left[(\boldsymbol{y}_i - \boldsymbol{X}_i\bar{\hat{\boldsymbol{\beta}}})'(\boldsymbol{y}_i - \boldsymbol{X}_i\bar{\hat{\boldsymbol{\beta}}})\right]}{T + \mathrm{E}\left(\boldsymbol{x}'_{i,T+1}\ \boldsymbol{Q}_{iT}^{-1}\boldsymbol{x}_{i,T+1}\right)}. \tag{60}$$

But

$$\begin{aligned}
\widetilde{\mathrm{SSR}}_i &= (\boldsymbol{y}_i - \boldsymbol{X}_i\bar{\hat{\boldsymbol{\beta}}})'(\boldsymbol{y}_i - \boldsymbol{X}_i\bar{\hat{\boldsymbol{\beta}}}) \\
&= [\boldsymbol{\varepsilon}_i + \boldsymbol{X}_i(\boldsymbol{\beta}_i - \bar{\hat{\boldsymbol{\beta}}})]'[\boldsymbol{\varepsilon}_i + \boldsymbol{X}_i(\boldsymbol{\beta}_i - \bar{\hat{\boldsymbol{\beta}}})] \\
&= \boldsymbol{\varepsilon}'_i\boldsymbol{\varepsilon}_i + 2\boldsymbol{\varepsilon}'_i\boldsymbol{X}_i(\boldsymbol{\beta}_i - \bar{\hat{\boldsymbol{\beta}}}) + (\boldsymbol{\beta}_i - \bar{\hat{\boldsymbol{\beta}}})'\boldsymbol{X}'_i\boldsymbol{X}_i(\boldsymbol{\beta}_i - \bar{\hat{\boldsymbol{\beta}}}).
\end{aligned}$$

Using (57) and $\hat{\boldsymbol{\beta}}_i = \boldsymbol{\beta}_i + \bar{\boldsymbol{\xi}}_{iT}$, then $\boldsymbol{\beta}_i - \bar{\hat{\boldsymbol{\beta}}} = \boldsymbol{\eta}_i - \bar{\boldsymbol{\eta}}_N - \ \bar{\boldsymbol{\xi}}_{NT}$ and we have

$$\begin{aligned}
\widetilde{\mathrm{SSR}}_i &= \boldsymbol{\varepsilon}'_i\boldsymbol{\varepsilon}_i + 2\boldsymbol{\varepsilon}'_i\boldsymbol{X}_i\left(\boldsymbol{\eta}_i - \bar{\boldsymbol{\eta}}_N - \ \bar{\boldsymbol{\xi}}_{NT}\right) \\
&\quad + \left(\boldsymbol{\eta}_i - \bar{\boldsymbol{\eta}}_N - \bar{\boldsymbol{\xi}}_{NT}\right)'\left(\boldsymbol{X}'_i\boldsymbol{X}_i\right)\left(\boldsymbol{\eta}_i - \bar{\boldsymbol{\eta}}_N - \bar{\boldsymbol{\xi}}_{NT}\right)',
\end{aligned}$$

and, under Assumptions 1–6, it follows that

$$\begin{aligned}
\mathrm{E}\left(\widetilde{\mathrm{SSR}}_i\right) &= T\sigma_i^2 + \mathrm{E}(\boldsymbol{\eta}'_i\boldsymbol{X}'_i\boldsymbol{X}_i\boldsymbol{\eta}_i) + O\left(\frac{1}{N}\right) \\
&= T\sigma_i^2 + \mathrm{E}\left(\sum_{t=1}^{T}\boldsymbol{x}'_{it}\boldsymbol{\Omega}_\eta\boldsymbol{x}_{it}\right) + O\left(\frac{1}{N}\right).
\end{aligned}$$

Using this result in (60), we have

$$\begin{aligned}
\mathrm{E}\left(\tilde{\sigma}^2_{i,NT}\right) - \sigma_i^2 &= \frac{T\sigma_i^2 + \mathrm{E}\left(\sum_{t=1}^{T}\boldsymbol{x}'_{it}\ \boldsymbol{\Omega}_\eta\boldsymbol{x}_{it}\right) + O\left(\frac{1}{N}\right)}{T + \mathrm{E}\left(\boldsymbol{x}'_{i,T+1}\boldsymbol{Q}_{iT}^{-1}\boldsymbol{x}_{i,T+1}\right)} - \sigma_i^2 \\
&= \frac{\mathrm{E}\left(T^{-1}\sum_{t=1}^{T}\boldsymbol{x}'_{it}\boldsymbol{\Omega}_\eta\boldsymbol{x}_{it}\right) - T^{-1}\sigma_i^2\mathrm{E}\left(\boldsymbol{x}'_{i,T+1}\boldsymbol{Q}_{iT}^{-1}\boldsymbol{x}_{i,T+1}\right)}{1 + T^{-1}\mathrm{E}\left(\boldsymbol{x}'_{i,T+1}\boldsymbol{Q}_{iT}^{-1}\boldsymbol{x}_{i,T+1}\right)} + O\left(\frac{1}{N}\right).
\end{aligned}$$

Under $H_{0,PF}$ defined by (19), $T^{-1}\sigma_i^2\mathrm{E}\left(\boldsymbol{x}'_{i,T+1}\boldsymbol{Q}_{iT}^{-1}\boldsymbol{x}_{i,T+1}\right) = \mathrm{E}\left(\boldsymbol{x}'_{i,T+1}\boldsymbol{\Omega}_\eta\boldsymbol{x}_{i,T+1}\right)$, so

$$\mathrm{E}\left(\tilde{\sigma}^2_{i,NT}\right) = \sigma_i^2 + \frac{\mathrm{E}\left(T^{-1}\sum_{t=1}^{T}\boldsymbol{x}'_{it}\boldsymbol{\Omega}_\eta\boldsymbol{x}_{it}\right) - \mathrm{E}\left(\boldsymbol{x}'_{i,T+1}\boldsymbol{\Omega}_\eta\boldsymbol{x}_{i,T+1}\right)}{1 + T^{-1}\mathrm{E}\left(\boldsymbol{x}'_{i,T+1}\boldsymbol{Q}_{iT}^{-1}\boldsymbol{x}_{i,T+1}\right)} + O\left(\frac{1}{N}\right),$$

and the desired result (24) follows if

$$\mathrm{E}\left(T^{-1}\sum_{t=1}^{T}\boldsymbol{x}_{it}'\,\boldsymbol{\Omega}_{\eta}\boldsymbol{x}_{it}\right)=\mathrm{E}\left(\boldsymbol{x}_{i,T+1}'\boldsymbol{\Omega}_{\eta}\boldsymbol{x}_{i,T+1}\right),$$

This condition can be written equivalently as

$$\mathrm{tr}\left\{\boldsymbol{\Omega}_{\eta}\left[T^{-1}\sum_{t=1}^{T}E\left(\boldsymbol{x}_{it}\boldsymbol{x}_{it}'\right)-E\left(\boldsymbol{x}_{i,T+1}\boldsymbol{x}_{i,T+1}'\right)\right]\right\}=0.$$

which is satisfied under (23).

Appendix B Details of the risk plots

The risk for each forecast is the average MSFE for the respective forecasting procedures. The data are generated as

$$y_{it}=\beta_i x_{it}+\sigma_i\varepsilon_{it},$$

where $x_{it}=\mu_{xi}+\sigma_{xi}v_{it}$, $\sigma_i^2\sim\mathrm{iid}\left(1+\chi_1^2\right)/2$, $\sigma_{xi}^2\sim\mathrm{iid}\left(1+\chi_1^2\right)/2$, $\beta_i=1+\sigma_\eta\eta_i$, $\varepsilon_{it}\sim\mathrm{iidN}\left(0,1\right)$, $v_{it}\sim\mathrm{iidN}\left(0,1\right)$, $\eta_i\sim\mathrm{iidN}\left(0,1\right)$, and $\mu_{xi}\sim\mathrm{iidN}\left(0,1\right)$. We repeat the forecasts 10,000 times to obtain the average MSFE for each forecasting method and value of σ_η^2.

In order to make the risk plots readable while keeping the computational cost feasible, we run the resulting risks functions through a filter for smoothing. In particular, the average MSFE for each forecasting method is smoothed using the Kalman filter and smoother for a local linear model

$$\begin{aligned}z_t&=\mu_t+\nu_t\\\mu_t&=\mu_{t-1}+\xi_t\end{aligned}$$

where the variances are set to $\sigma_\nu^2=0.3$ and $\sigma_\xi^2=0.2$. For the individual forecasts, we simply use the mean.

Appendix C Details of the estimators in the Monte Carlo experiments and applications

Individual estimation This is the forecast in (2). This forecast is the reference forecast and the MSFE of all other methods are reported as ratios relative to the MSFE of this forecast.

Pooled estimation This is the forecast in (4).

Goldberger's random effects BLUP This forecast uses the best linear unbiased predictor (BLUP) of Goldberger (1962) as reviewed by Baltagi (2013). For the random effects model

$$y_{i,T+1} = \alpha + \boldsymbol{\beta}'\boldsymbol{x}_{i,T+1} + u_{i,T+1}$$

where $u_{i,T+1} = \eta_i + \varepsilon_{i,T+1}$, the BLUP forecast is

$$\hat{y}_{i,T+1} = \hat{\alpha}_{\text{GLS}} + \hat{\boldsymbol{\beta}}'_{\text{GLS}}\boldsymbol{x}_{i,T+1} + \frac{\hat{\sigma}^2_\eta}{T\hat{\sigma}^2_\eta + \hat{\sigma}^2_\varepsilon}\ (\boldsymbol{l}'_i \otimes \boldsymbol{\iota}_T)\hat{\boldsymbol{u}}_{\text{GLS}}$$

and $\boldsymbol{l}_i$ is the ith column of $\boldsymbol{I}_N$, $\hat{\alpha}_{\text{GLS}}$ and $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ are estimated by GLS with covariance matrix

$$\boldsymbol{\Sigma} = T\sigma^2_\eta \boldsymbol{P} + \sigma^2_\varepsilon \boldsymbol{I}$$

$\boldsymbol{P} = \boldsymbol{X}_\mu(\boldsymbol{X}'_\mu\boldsymbol{X}_\mu)^{-1}\boldsymbol{X}'_\mu$, $\boldsymbol{X}_\mu = \boldsymbol{I}_N \otimes \boldsymbol{\iota}_T$, $\boldsymbol{\iota}_T$ is a $T \times 1$ vector of ones, $\hat{\boldsymbol{u}}_{\text{GLS}} = \boldsymbol{y} - \hat{\alpha}_{\text{GLS}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{GLS}}$,

$$\hat{\sigma}^2_\varepsilon = \frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=1}^{T}(u_{it} - \bar{u}_i)^2$$

and

$$\hat{\sigma}^2_\eta = \frac{1}{N}\sum_{i=1}^{N}\hat{\eta}^2_i$$

with $\hat{\eta}_i$ obtained from the fixed effects estimation.

Median Group The forecast uses the median group estimator of $\boldsymbol{\beta}$,

$$\hat{\beta}^{(MG)}_k = \text{Median}\left(\{\hat{\beta}_{ik}\}_{i=1,2,\ldots,N}\right),\ \text{for } k = 1, 2, \ldots, K$$

where $\hat{\beta}_{ik}$ is the k-th element of the individual estimator in (2).

Combination: Naive weights This is the forecast in (11) with weights (15) where

$$\hat{\boldsymbol{\Omega}}_\eta = \frac{1}{N}\sum_{i=1}^{N}(\hat{\boldsymbol{\beta}}_i - \bar{\hat{\boldsymbol{\beta}}})(\ \hat{\boldsymbol{\beta}}_i - \bar{\hat{\boldsymbol{\beta}}})' \quad \text{and} \quad \bar{\hat{\boldsymbol{\beta}}} = \frac{1}{N}\sum_{i=1}^{N}\ \hat{\boldsymbol{\beta}}_i$$

and $\tilde{\sigma}^2_i$ is given in (22).

Combination: First-order bias-corrected weights This is the forecast in (11) with weights (15) with $\tilde{\boldsymbol{\Omega}}_\eta$ replacing $\hat{\boldsymbol{\Omega}}_\eta$ where

$$\tilde{\boldsymbol{\Omega}}_\eta = \hat{\boldsymbol{\Omega}}_\eta - \frac{1}{N}\sum_{i=1}^{N} \tilde{\sigma}_i^2 (\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}$$

with $\tilde{\sigma}_i^2$ given in (22).

Pre-test Parameter homogeneity is tested using the test statistic (20) using the $\tilde{\sigma}_i^2$ in (22). If parameter heterogeneity is rejected the individual forecast is chosen, otherwise the pooled forecast is taken.

Prior likelihood This forecast uses the parameter estimator (26) together with (27) and (28).

Bayesian This forecast uses the parameter estimator (26) together with (29) and (30).

Empirical Bayes This forecast uses the parameter estimator (26) together with (31) and (32).

Appendix D Derivation of PR_N^2

Consider the panel data model

$$y_{it} = \alpha_i + \rho_i y_{i,t-1} + \gamma_i x_{it} + \kappa\sigma_i\varepsilon_{it}, \tag{61}$$

$$x_{it} = \mu_{xi} + \xi_{it}, \quad \xi_{it} = \rho_{xi}\xi_{i,t-1} + \sigma_{xi}\sqrt{1-\rho_{xi}^2}\nu_{it}.$$

$\text{Var}(\varepsilon_{it}) = 1$, and $\text{Var}(\nu_{it}) = 1$ as set out in further detail in Section 3. In order to simplify the derivations, we treat x_{it} as strictly exogenous (no feedback from $y_{i,t-1}$), and assume that y_{it} is stationary and has started a long time in the past. To deal with the heterogeneity across the different equations in the panel, we use the following average measure of fit, for a given N,

$$PR_N^2 = 1 - \frac{N^{-1}\sum_{i=1}^{N}\text{Var}\left(\kappa\sigma_i\varepsilon_{it}\,|\boldsymbol{\theta}_i,\, x_{it}\right)}{N^{-1}\sum_{i=1}^{N}\text{Var}(y_{it}\,|\boldsymbol{\theta}_i,\, x_{it},\,)}, \tag{62}$$

where $\boldsymbol{\theta}_i = (\alpha_i, \rho_i, \gamma_i, \sigma_i)'$. For the numerator we have

$$\text{Var}\left(\kappa\sigma_i\varepsilon_{it} \mid \boldsymbol{\theta}_i,\, x_{it}\;\right) = \kappa^2\sigma_i^2. \tag{63}$$

To derive $\mathrm{Var}(y_{it}\,|\boldsymbol{\theta}_i,\, x_{it}\,)$, we note that

$$\begin{aligned}
\mathrm{Var}(y_{it}\,|\boldsymbol{\theta}_i, x_{it}\,) &= \mathrm{E}\left[\mathrm{Var}(y_{it}\,|\boldsymbol{\theta}_i, y_{i,t-1}, x_{it}\,)\right] + \mathrm{Var}\left[\mathrm{E}(y_{it}\,|\boldsymbol{\theta}_i, y_{i,t-1}, x_{it}\,)\right], \\
\mathrm{Var}(y_{it}\,|\boldsymbol{\theta}_i, y_{i,t-1}, x_{it}\,) &= \kappa^2\sigma_i^2, \\
\mathrm{E}(y_{it}\,|\boldsymbol{\theta}_i, y_{i,t-1}, x_{it}\,) &= \alpha_i + \rho_i y_{i,t-1} + \gamma_i x_{it}, \\
\mathrm{Var}\left[\mathrm{E}(y_{it}\,|\boldsymbol{\theta}_i, y_{i,t-1}, x_{it}\,)\right] &= \rho_i^2\mathrm{Var}(y_{it}\,|\boldsymbol{\theta}_i, x_{it}\,) + \gamma_i^2\mathrm{Var}\left(x_{it}\right).
\end{aligned}$$

Hence

$$\mathrm{Var}(y_{it}\,|\boldsymbol{\theta}_i, x_{it}\,) = \frac{\gamma_i^2\mathrm{Var}(\xi_{it}) + \kappa^2\sigma_i^2}{1-\rho_i^2} \tag{64}$$

Now using (63) and (64) in (62), we obtain

$$PR_N^2 = 1 - \kappa^2\left(\frac{N^{-1}\sum_{i=1}^N \sigma_i^2}{N^{-1}\sum_{i=1}^N \frac{\gamma_i^2\sigma_{xi}^2+\kappa^2\sigma_i^2}{1-\rho_i^2}}\right),$$

where $\sigma_{xi}^2 = \mathrm{Var}(\xi_{it})$ and after some simplifications we have

$$PR_N^2 = \frac{b_N + \kappa^2\left(c_N - a_N\right)}{b_N + \kappa^2 c_N}, \tag{65}$$

where

$$\begin{aligned}
a_N &= N^{-1}\sum_{i=1}^N \sigma_i^2, \quad b_N = N^{-1}\sum_{i=1}^N \frac{\gamma_i^2\sigma_{xi}^2}{1-\rho_i^2}, \\
c_N &= N^{-1}\sum_{i=1}^N \frac{\sigma_i^2}{1-\rho_i^2}.
\end{aligned}$$

Then

$$\kappa^2 = \frac{b_N(1-PR_N^2)}{a_N - c_N(1-PR_N^2)}$$

A number of observations follow from this. For $\kappa^2 > 0$, we must have

$$a_N - c_N(1-PR_N^2) > 0,$$

or $PR_N^2 > 1 - a_N/c_N$. So we can not set PR_N too low relative to the distribution of ρ_i over i. It is clear that if we fix PR_N^2 for a given N, then κ will vary across N, as well. For a given value of PR_N^2 and for a finite N,

one can simulate the values of a_N, b_N and c_N from the N random draws of $\sigma_i^2, \sigma_{xi}^2, \rho_i$, and γ_i.

When these parameters are distributed independently, as $N \to \infty$

$$\begin{aligned} a_N &\xrightarrow{p} \mathrm{E}(\sigma_i^2) \\ b_N &\xrightarrow{p} \mathrm{E}(\gamma_i^2)\mathrm{E}(\sigma_{xi}^2)\mathrm{E}\left(\frac{1}{1-\rho_i^2}\right) \\ c_N &\xrightarrow{p} \mathrm{E}(\sigma_i^2)\mathrm{E}\left(\frac{1}{1-\rho_i^2}\right) \end{aligned}$$

Hence, using (65), we note that (as $N \to \infty$)

$$PR_N^2 \to PR^2 = \frac{\mathrm{E}(\gamma_i^2)\mathrm{E}(\sigma_{xi}^2)\mathrm{E}\left(\frac{1}{1-\rho_i^2}\right) + \kappa^2\left(\mathrm{E}(\sigma_i^2)\mathrm{E}\left(\frac{1}{1-\rho_i^2}\right) - \mathrm{E}(\sigma_i^2)\right)}{\mathrm{E}(\gamma_i^2)\mathrm{E}(\sigma_{xi}^2)\mathrm{E}\left(\frac{1}{1-\rho_i^2}\right) + \kappa^2\mathrm{E}(\sigma_i^2)\mathrm{E}\left(\frac{1}{1-\rho_i^2}\right)}.$$

Under our design $\mathrm{E}(\sigma_i^2) = 1$, $\mathrm{E}(\sigma_{xi}^2) = 1$, and the above expression simplifies to

$$PR^2 = \frac{\mathrm{E}(\gamma_i^2)\mathrm{E}\left(\frac{1}{1-\rho_i^2}\right) + \kappa^2\left[\mathrm{E}\left(\frac{1}{1-\rho_i^2}\right) - 1\right]}{\mathrm{E}(\gamma_i^2)\mathrm{E}\left(\frac{1}{1-\rho_i^2}\right) + \kappa^2\mathrm{E}\left(\frac{1}{1-\rho_i^2}\right)}.$$

Hence, we have

$$\kappa^2 = \frac{\mathrm{E}(\gamma_i^2)(1-PR^2)}{\frac{1}{\mathrm{E}\left(\frac{1}{1-\rho_i^2}\right)} - (1-PR^2)}.$$

Since $\kappa^2 > 0$, we must also have

$$\frac{1}{\mathrm{E}\left(\frac{1}{1-\rho_i^2}\right)} - (1-PR^2) > 0 \quad \text{or} \quad PR^2 > 1 - \frac{1}{\mathrm{E}\left(\frac{1}{1-\rho_i^2}\right)}.$$

Note that $\mathrm{E}\left(\frac{1}{1-\rho_i^2}\right) > 0$. If ρ_i is homogeneous, such that $\rho_i = \rho$, $\mathrm{E}\left(\frac{1}{1-\rho_i^2}\right) = 1/(1-\rho^2)$, and the above condition simplifies to the familiar condition $PR^2 > \rho^2$.

In the general case where σ_i^2 is not distributed independently of ρ_i and N is finite we have

$$\begin{aligned} PR_N^2 &> 1 - a_N/c_N \\ &= 1 - \frac{N^{-1}\sum_{i=1}^N \sigma_i^2}{N^{-1}\sum_{i=1}^N \frac{\sigma_i^2}{1-\rho_i^2}}. \end{aligned}$$

In the case where $\rho_i \sim$ iidUniform$(0, \bar{\rho})$, with $\bar{\rho} < 1$, we have that

$$\begin{aligned} \mathrm{E}\left(\frac{1}{1-\rho_i^2}\right) &= \frac{1}{\bar{\rho}} \int_0^{\bar{\rho}} \frac{1}{1-x^2} dx \\ &= \frac{1}{2\bar{\rho}} \int_0^{\bar{\rho}} \left[\frac{1}{1+x} + \frac{1}{1-x}\right] dx \\ &= \frac{1}{2\bar{\rho}} \left[ln(1+x) - ln(1-x)\right]_0^{\bar{\rho}} \\ &= \frac{1}{2\bar{\rho}} ln\left(\frac{1+\bar{\rho}}{1-\bar{\rho}}\right). \end{aligned} \tag{66}$$

and therefore

$$PR^2 > 1 - \frac{2\bar{\rho}}{ln\left(\frac{1+\bar{\rho}}{1-\bar{\rho}}\right)}.$$

In this case, we obtain $PR^2 > 0.481$ if $\bar{\rho} = 0.9$ while for $\bar{\rho} = 0.8$, we obtain the condition $PR^2 > 0.272$. In the homogeneous case the equivalent measure is $PR^2 > 0.4^2 = 0.16$.

Large N and finite T population pooled R^2

In place of using the population moments analyzed above, consider now the finite sample case, where T is finite but the ARDL processes in (61) have started from some distance in the past, such that

$$y_{it} = \frac{\alpha_i}{1-\rho_i} + \gamma_i z_{it}(\rho_i) + \kappa\sigma_i u_{it}, \quad t = 1, 2, \ldots, T$$

where

$$\begin{aligned} z_{it}(\rho_i) &= \rho_i z_{i,t-1}(\rho_i) + x_{it} = \sum_{s=0}^{\infty} \rho_i^s x_{i,t-s}, \\ u_{it} &= \rho_i u_{i,t-1} + \varepsilon_{it} = \sum_{s=0}^{\infty} \rho_i^s \varepsilon_{i,t-s}. \end{aligned}$$

Since the fit is conditional on past observed data, we need to write y_{it} in terms of ε_{it}, namely

$$y_{it} = \frac{\alpha_i}{1-\rho_i} + \gamma_i z_{it}(\rho_i) + \kappa\sigma_i\rho_i u_{i,t-1} + \kappa\sigma_i\varepsilon_{it}.$$

Also

$$\begin{aligned} y_{it} - \bar{y}_{iT} &= \gamma_i \left[z_{it}(\rho_i) - \bar{z}_{iT}(\rho_i)\right] + \kappa\sigma_i\rho_i \left(u_{i,t-1} - \bar{u}_{i,-1T}\right) + \kappa\sigma_i \left(\varepsilon_{it} - \bar{\varepsilon}_{iT}\right) \\ &= \gamma_i \left[z_{it}(\rho_i) - \bar{z}_{iT}(\rho_i)\right] + \kappa\sigma_i \left(u_{it} - \bar{u}_{iT}\right), \end{aligned}$$

where

$$\begin{aligned}\bar{y}_{iT} &= T^{-1}\sum_{t=1}^{T} y_{it}, \quad \bar{z}_{iT} = T^{-1}\sum_{t=1}^{T} z_{it}(\rho_i), \\ \bar{u}_{i,-1,T} &= T^{-1}\sum_{t=1}^{T} u_{i,t-1}, \quad \bar{\varepsilon}_{iT} = T^{-1}\sum_{t=1}^{T} \varepsilon_{it}\,,\ \bar{u}_{iT} = T^{-1}\sum_{t=1}^{T} u_{it}.\end{aligned}$$

Then

$$PR^2_{N,T} = 1 - \frac{\kappa^2_{N,T} T^{-1}\sum_{t=1}^{T} N^{-1}\sum_{i=1}^{N} \sigma_i^2 \left(\varepsilon_{it} - \bar{\varepsilon}_{iT}\right)^2}{N^{-1}T^{-1}\sum_{t=1}^{T}\sum_{i=1}^{N}\left\{\gamma_i\left[z_{it}(\rho_i) - \bar{z}_{iT}(\rho_i)\right] + \kappa_{N,T}\sigma_i\left(u_{it} - \bar{u}_{iT}\right)\right\}^2}.$$

Since σ_i^2 are drawn independently of ε_{it}, and for each i, and ε_{it} is distributed as *iid* over i, then for each t

$$N^{-1}\sum_{i=1}^{N} \sigma_i^2 \left(\varepsilon_{it} - \bar{\varepsilon}_{iT}\right)^2 \xrightarrow{p} \lim_{N\to\infty} N^{-1}\sum_{i=1}^{N} \mathrm{E}\left(\sigma_i^2\right)\mathrm{E}\left(\varepsilon_{it} - \bar{\varepsilon}_{iT}\right)^2 = \frac{T-1}{T}.$$

As in our design $\mathrm{E}\left(\sigma_i^2\right) = 1$, and $\mathrm{E}\left(\varepsilon_{it} - \bar{\varepsilon}_{iT}\right)^2 = 1 - T^{-1}$. For the denominator of $PR^2_{N,T}$ note that

$$\begin{aligned}& N^{-1}\sum_{i=1}^{N}\left\{\gamma_i\left[z_{it}(\rho_i) - \bar{z}_{iT}(\rho_i)\right] + \kappa_{N,T}\sigma_i\left(u_{it} - \bar{u}_{iT}\right)\right\}^2 \\ = \;& N^{-1}\sum_{i=1}^{N}\gamma_i^2\left[z_{it}(\rho_i) - \bar{z}_{iT}(\rho_i)\right]^2 + N^{-1}\sum_{i=1}^{N}\sigma_i^2\left(u_{it} - \bar{u}_{iT}\right)^2 \\ & + 2N^{-1}\sum_{i=1}^{N}\sigma_i\gamma_i\left(u_{it} - \bar{u}_{iT}\right)\left[z_{it}(\rho_i) - \bar{z}_{iT}(\rho_i)\right].\end{aligned}$$

Since σ_i, γ_i, ε_{it} and $\xi_{it'}$ are distributed independently (when x_{it} is strictly exogenous and γ_i and x_{it} are independently distributed), then

$$N^{-1}\sum_{i=1}^{N}\gamma_i^2\left[z_{it}(\rho_i) - \bar{z}_{iT}(\rho_i)\right]^2 \xrightarrow{p} \lim_{N\to\infty} N^{-1}\sum_{i=1}^{N}\mathrm{E}\left\{\gamma_i^2\left[z_{it}(\rho_i) - \bar{z}_{iT}(\rho_i)\right]^2\right\}$$

$$N^{-1}\sum_{i=1}^{N}\sigma_i^2\left(u_{it} - \bar{u}_{iT}\right)^2 \xrightarrow{p} \lim_{N\to\infty} N^{-1}\sum_{i=1}^{N}\mathrm{E}\left(\sigma_i^2\right)\mathrm{E}\left(u_{it} - \bar{u}_{iT}\right)^2$$

$$\mathrm{E}\left(u_{it} - \bar{u}_{iT}\right)^2 = \frac{1}{1-\rho_i^2} + \frac{1}{T}\left[1 + 2\sum_{h=1}^{T-1}\left(1 - \frac{h}{T}\right)\rho_i^h\right] - \frac{2}{T}\sum_{h=1}^{T}\rho_i^{|t-h|}$$

Hence

$$N^{-1}\sum_{i=1}^{N}\sigma_i^2\left(u_{it}-\bar{u}_{iT}\right)^2 \xrightarrow{p} \mathrm{E}\left(\frac{1}{1-\rho_i^2}\right)+\frac{1}{T}\left[1+2\sum_{h=1}^{T-1}\left(1-\frac{h}{T}\right)\mathrm{E}\left(\rho_i^h\right)\right] -\frac{2}{T}\sum_{h=1}^{T}\mathrm{E}\left(\rho_i^{|t-h|}\right),$$

and

$$N^{-1}\sum_{i=1}^{N}\sigma_i\gamma_i\left(u_{it}-\bar{u}_{iT}\right)\left[z_{it}(\rho_i)-\bar{z}_{iT}(\rho_i)\right]\xrightarrow{p}0.$$

Hence (as $N\to\infty$, $\kappa_{N,T}\to\kappa_T$)

$$PR^2_{N,T}\to_p PR^2_T=1-\frac{\left(1-\frac{1}{T}\right)\kappa_T^2}{A_T+\kappa_T^2B_T}, \tag{67}$$

where

$$A_T=T^{-1}\sum_{t=1}^{T}\left[\lim_{N\to\infty}N^{-1}\sum_{i=1}^{N}\mathrm{E}\left(\gamma_i^2\right)\mathrm{E}\left[z_{it}(\rho_i)-\bar{z}_{iT}(\rho_i)\right]^2\right],$$

and

$$B_T = E\left(\frac{1}{1-\rho_i^2}\right)+\frac{1}{T}\left[1+2\sum_{h=1}^{T-1}\left(1-\frac{h}{T}\right)\mathrm{E}\left(\rho_i^h\right)\right] -\frac{2}{T}\sum_{h=1}^{T}T^{-1}\sum_{t=1}^{T}\mathrm{E}\left(\rho_i^{|t-h|}\right).$$

The expression for $\mathrm{E}\left(\frac{1}{1-\rho_i^2}\right)$ is given by (66) and $\mathrm{E}(\rho_i^c)=\bar{\rho}^c/(c+1)$, and A_T can be computed by stochastic simulation as:

$$A_T=\frac{1}{RNT}\sum_{r=1}^{R}\sum_{t=1}^{T}\sum_{i=1}^{N}\gamma_i^2\left[z_{it}^{(r)}(\rho_i)-\bar{z}_{iT}^{(r)}(\rho_i)\right]^2,$$

where $z_{it}^{(r)}(\rho_i)$ is the r^{th} draw from the distribution of $z_{it}^{(r)}(\rho_i)$.

Table 7 reports the numerical values of κ for different values of PR^2, T, σ_β^2, and $\rho_{\beta x}$. The values are obtained from simulations using 10,000 repetitions.

Table 7: Monte Carlo parameterization

PR^2	T	σ^2_β	κ	κ
			$r_{\beta x} = 0$	$r_{\beta x} = 0.5$
0.2	20	0	0.8285	0.8261
		0.1	2.0211	1.9900
		0.25	2.5168	2.4831
		0.5	3.6878	3.7017
	50	0	0.5932	0.5887
		0.1	0.9098	0.9641
		0.25	1.2735	1.2025
		0.5	2.1259	2.1208
	100	0	0.9018	0.8814
		0.1	2.1918	2.1950
		0.25	2.7212	2.7094
		0.5	4.4307	4.3744
0.6	20	0	0.6084	0.5941
		0.1	0.9979	1.0197
		0.25	1.2847	1.2998
		0.5	3.2462	2.9754
	50	0	0.8592	0.8845
		0.1	2.2343	2.2013
		0.25	2.7972	2.8835
		0.5	4.8454	4.6805
	100	0	0.6189	0.6204
		0.1	1.0352	0.9966
		0.25	1.2949	1.3723
		0.5	3.0190	3.0304

The table contains the estimated κ using (67) with 10,000 simulations for a range of parameter values.