Panel Forecasting with Asymmetric Grouping

Didier Nibbering*[a] and Richard Paap[b]

[a]Department of Econometrics and Business Statistics, Monash University
[b]Econometric Institute, Tinbergen Institute, Erasmus University Rotterdam

EI-2019-30

Abstract

This paper proposes an asymmetric grouping estimator for panel data forecasting. The estimator relies on the observation that the bias-variance trade-off in potentially heterogeneous panel data may be different across individuals. Hence, the group of individuals used for parameter estimation that is optimal in terms of forecast accuracy, may be different for each individual. For a specific individual, the estimator uses cross-validation to estimate the bias-variance of all individual groupings, and uses the parameter estimates of the optimal grouping to produce the individual-specific forecast. Integer programming and screening methods deal with the combinatorial problem of a large number of individuals. A simulation study and an application to market leverage forecasts of U.S. firms demonstrate the promising performance of our new estimators.

Keywords: Panel data, forecasting, parameter heterogeneity
JEL Classification: C23, C51, C53

*Correspondence to: Department of Econometrics & Business Statistics, Monash University, Clayton VIC 3800, Australia, e-mail: didier.nibbering@monash.edu

We would like to thank Patrick Groenen for helpful discussions.

1 Introduction

Forecast accuracy suffers in many applications from substantial parameter estimation uncertainty due to a small number of available observations. The availability of potentially relevant additional 'panels' of data may decrease forecast variance and thus increase forecast performance. When the panels are short, one commonly gains in efficiency by estimating the parameter values on the pooled observations from all panels (Baltagi et al., 2008). However, forecasts based on the assumption of complete parameter homogeneity across all panels may suffer from substantial bias. The accuracy of panel-specific forecasts relies on the ability of the researcher to model possible parameter heterogeneity across panels, balancing the efficiency gains from pooling and the bias due to panel heterogeneity.

An alternative to assuming either complete heterogeneity or homogeneity across panels, is to assume that panels can be classified into groups with homogeneous parameters, while allowing for heterogeneity across groups. To estimate the unknown number of groups and the group membership of each panel, researchers use methods from the machine learning literature; Lin and Ng (2012) and Ando and Bai (2016) use the K-means algorithm to cluster panels and Su et al. (2016) and Wang et al. (2018) develop lasso-type estimators to estimate groups of panels. Estimating a symmetric grouping, in which panel A clusters with panel B if and only if panel B clusters with panel A, takes parameter heterogeneity across panels into account. However, it does not account for heterogeneity in panel-specific forecast accuracy.

This paper proposes an asymmetric grouping estimator. The estimator relies on the observation that the bias-variance trade-off may be different in each panel, and therefore the optimal grouping in terms of forecast accuracy may be different for each panel. The asymmetric grouping estimator separately estimates parameter values for each panel of interest, potentially also using observations from other available panels. For a specific panel, the estimator uses leave-one-out cross-validation to estimate the bias-variance trade-off of all groupings that involve the panel, and uses the parameter estimates of the optimal grouping to produce the panel-specific forecast. The estimator is called asymmetric as it does not have to be the case that the optimal grouping for panel A is the same as the optimal grouping for panel B. We refer to a standard segmentation as symmetric grouping.

Since the asymmetric grouping estimator does not assume a latent group structure, it does not require knowledge of the number of groups. Estimating the number of groups in symmetric grouping estimators involves sequential testing (Lin and Ng, 2012), information criteria (Su et al., 2016; Wang et al., 2018), and/or tuning parameters that introduce additional estimation uncertainty. Even when information about the group structure is available, as assumed in Bester and Hansen (2016), a higher level grouping potentially improves forecast accuracy.

We derive an expression for the mean squared forecast error of grouping estimators. This expression identifies settings in which an asymmetric grouping improves upon forecast accuracy relative to symmetric grouping. Since the asymmetric grouping estimator iterates over all possible combinations of panels, we introduce a sequential integer programming approach that does not need to explore every possible combination. The approximation error is small when the sample covariance matrix of the regressors is approximately the same in each panel, while the gains in computation time are substantial. However, there are no guarantees on the computation time of integer programming. We show that asymmetric grouping with a huge number of panels is feasible by using an initial screening step, under the assumption of bounded forecast bias. This screening step reduces an NP-hard problem to computation time that increases linearly in the number of panels.

We study the theoretical results in a finite sample simulation study and find that asymmetric grouping estimators substantially increase forecast performance in both weakly and strongly heterogeneous panels. An empirical application to market leverage forecasts of publicly traded U.S. firms shows that asymmetric grouping improves upon symmetric grouping estimators in terms of mean squared forecast error in panel data with 10, 25 and 172 firms.

Many other methods are proposed to deal with potentially heterogeneous panel data. First, researchers allow for heterogeneous intercept parameters and assume homogeneous slope parameters. These fixed effects estimators are poorly estimated in short panels (Neyman and Scott, 1948). Bonhomme and Manresa (2015) and Bester and Hansen (2016) address this incidental parameter problem by estimating grouped fixed effects to increase the accuracy of the intercept estimate. The random coefficient model of Swamy (1970) provides heterogeneous parameter estimates that rely on distributional assumptions on the parameters. Individual parameters are shrunk towards a common pooled parameter value, where the pooling and amount of shrinkage depends on the amount of information in the individual time series.

Second, under the assumption of an underlying group structure, we can use statistical pretests for parameter heterogeneity across panels (Danilov and Magnus, 2004; Pesaran and Yamagata, 2008; Jin and Su, 2013; Juhl and Lugovskyy, 2014). Although one may decide to forecast with a pooled regression when the hypothesis of homogeneity across panels cannot be rejected, the alternative is not very helpful. Sequentially testing for homogeneity across subgroups of panels potentially leads to a large number of tests and a substantial increase in forecast variance. The decision to reject a model specification relies on an arbitrarily chosen significance level. Moreover, these tests aim to select a 'true' model, which does not have to correspond to a model that performs best in terms of forecast accuracy.

Third, instead of estimating an underlying group structure, one can also average over models with different groupings. Wang et al. (2015) combine forecasts from models with different panel groupings and estimate the weights based on the Mallows criterion. Desbordes et al. (2018) use Bayesian model averaging to combine models with different panel groupings. These methods implicitly assume a symmetric grouping by using the same model weights for each individual panel forecast. Finite mixture models jointly estimate different sets of parameter values and the corresponding probability weights for each panel (McLachlan and Peel, 2000; Frühwirth-Schnatter, 2006; Kasahara and Shimotsu, 2009). Although weighting is panel-specific, each forecast is based on the same fixed number of parameter values with nonzero weights.

Fourth, Maddala et al. (1997) propose shrinkage estimators that shrink the parameter estimates using only the data in each panel to the parameter estimates based on the pooled panel data. This idea is also applied in hierarchical Bayes models for panel data by, for instance, Chib (2008). Shrinking the parameter estimates for each panel to the same mean is only reasonable under the assumption that the panel-specific parameters have similar values. Moreover, the shrinkage estimators require the researcher to set a regularization strength, which is imposed to be the same for each panel.

This paper is structured as follows. Section 2 introduces the asymmetric grouping estimator, derives the settings under which asymmetric grouping improves the forecast accuracy, and discusses estimation when the number of available panels is large. Section 3 derives theoretical results on the conditions where the estimator is useful. The finite sample performance of the asymmetric grouping estimator is compared to other estimators in a Monte Carlo study in Section 4. Section 5 discusses an empirical application. Finally, Section 6 concludes.

2 Methods

This section develops a forecasting method for panel data. We first introduce the general panel data model setup. Second, we show how panel data forecasts can be constructed in this setting. Third, we introduce the idea of forecasting using an asymmetric grouping of panel data, and fourth we introduce algorithms for estimating the grouping that minimizes the mean squared forecast error. This section concludes with a discussion on the interpretation of the asymmetric grouping estimator.

2.1 Setup

Consider the panel regression model

$$y_i = X_i\beta_i + \varepsilon_i, \quad \varepsilon_i \sim \text{i.i.d}(0, \sigma_i^2 I_{T_i}), \quad i = 1, \ldots, N, \tag{1}$$

where $y_i = (y_{i1}, \ldots, y_{iT_i})'$ is a $T_i \times 1$ response vector, $X_i = (X'_{i1}, \ldots, X'_{iT_i})'$ a $T_i \times p$ regressor matrix, $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iT_i})$ a $T_i \times 1$ independent and identically distributed error vector with mean zero and variance σ_i^2, and N the number of available panels. The regressors in X_{it} are assumed to be uncorrelated with the error term ε_{it}. The coefficients $\beta_i = (\beta_{i1}, \ldots, \beta_{ip})'$ are assumed to be fixed but allowed to differ across panels.

We study the mean squared forecast error of the individual panel point forecast $\hat{y}_{i,T_i+1}$ for y_{i,T_i+1}. The mean squared forecast error for any estimator $\hat{\beta}_i$ for β_i is defined as

$$\rho_i = E\left[\left(y_{i,T_i+1} - x'_{i,T_i+1}\hat{\beta}_i\right)^2 - \sigma_i^2\right], \tag{2}$$

where the variance σ_i^2 is subtracted as it arises from the error ε_{i,T_i+1}, which is unpredictable for any method, and we take the expectation over the error terms ε_{it}, for $t = 1, \ldots, T_i$, and $i = 1, \ldots, N$.

The forecast $\hat{y}_{i,T_i+1}$ can be constructed by estimating β_i in (1) by ordinary least squares only using the data in panel i. However, when the sample size or the signal-to-noise ratio in panel i is low, substantial parameter uncertainty can lead to inaccurate forecasts. When data from other panels is available, the mean squared forecast error may benefit from estimating β_i using data from other panels as well. The pooled estimator uses all available panels to produce a forecast. This estimator is more efficient than the individual estimator, but can be biased when coefficients are strongly heterogeneous.

2.2 Panel forecasts from grouping estimators

The bias-variance trade-off between the individual and pooled estimator motivates the use of grouping estimators. Estimating β_i on a subset of multiple panels may introduce less forecast bias than the pooled estimator and less forecast variance than the individual estimator. Grouping estimators are able to exploit this bias-variance trade-off by using the data from the set of panels that minimizes the mean squared forecast error in (2), potentially improving upon both the individual and pooled estimator in terms of forecast accuracy.

Denote the estimator for β_i based on the data in a set of panels s by $\hat{\beta}_i(s)$. The set $s \in S_i$ contains a subset of the numbers $1, \ldots, N$, indicating the panels used to estimate β_i. We impose that s always includes panel

i. Denote the number of elements in s by $|s|$. Let S_i be the superset of s that contains all 2^{N-1} combinations s that include panel i, and S be the superset of S_i including all $2^N - 1$ unique combinations of N panels of length $|s| = 1, \ldots, N$. The estimator $\hat{\beta}_i(s)$ for β_i equals

$$\hat{\beta}_i(s) = \left(\sum_{l \in s} X_l' X_l \right)^{-1} \sum_{l \in s} X_l' y_l, \tag{3}$$

which uses the data in the set of panels $s \in S_i$. The point forecast for y_{i,T_i+1} based on $\hat{\beta}_i(s)$ is denoted by

$$\hat{y}_{i,T_i+1}(s) = x'_{i,T_i+1} \hat{\beta}_i(s). \tag{4}$$

The grouping estimator in (3) includes the case in which panels in s are pooled up to a panel-specific fixed effect or scaling factor. When we exclude an intercept from X_i, (1) can be rewritten to

$$y_i = \alpha_i + X_i c_i \beta_i + u_i, \quad u_i \sim \text{i.i.d}(0, c_i^2 \sigma_i^2 I_{T_i}), \quad i = 1, \ldots, N, \tag{5}$$

where the scalars α_i and c_i represent a panel-specific fixed effect and scaling factor, respectively. Define $\mu_{y_i} = \frac{1}{T_i} \sum_{t=1}^{T_i} y_{it}$, $\mu_{X_i} = \frac{1}{T_i} \sum_{t=1}^{T_i} X_{it}$, and $\sigma_{y_i} = \sqrt{\frac{1}{T_i} \sum_{t=1}^{T_i} (y_{it} - \mu_{y_i})^2}$, then β_i in (5) can be estimated using (3) by replacing X_i and y_i by $\tilde{X}_i = X_i - \mu_{X_i}$ and $\tilde{y}_i = \frac{y_i - \mu_{y_i}}{\sigma_{y_i}}$, respectively. The estimate for α_i equals $\hat{\alpha}_i = \frac{\mu_{y_i} - \mu_{X_i} \hat{\beta}_i}{\sigma_{y_i}}$ and for c_i we get $\hat{c}_i = \frac{1}{\sigma_{y_i}}$. Setting $\sigma_{y_i} = 1$ only allows for panel-specific fixed effects within s, and setting $\mu_{y_i} = 0$ and $\mu_{X_i} = 0$ only for panel-specific scaling within s.

The individual estimator and the pooled estimator are two widely used special cases of (3). When $s = \{i\}$, (3) only uses the data in panel i to estimate β_i, which is equivalent to the individual estimator. The pooled estimator uses all available panels to produce a forecasts. This boils down to (4) with $s = \{1, \ldots, N\}$.

The grouping estimator in (3) allows s to contain more than one but less than N elements, which results in 2^{N-1} unique combinations s for forecasting

$\hat{y}_{i,T_i+1}(s)$. Among others, Lin and Ng (2012), Ando and Bai (2016), Su et al. (2016), and Wang et al. (2018) estimate s by restricting the total number of possible combinations in S. They propose estimators that account for a group structure by assuming a 'true' grouping of panels. The coefficients within a group of panels are homogeneous, but coefficients are heterogeneous across groups of panels. Applying these 'symmetric grouping' methods to forecasting with panel data, restricts each forecast $\hat{y}_{i,T_i+1}$ to be constructed from the same underlying group structure. In other words, y_{i,T_i+1} is forecast by $\hat{y}_{i,T_i+1}(s)$ with $j \in s$ if and only if y_{j,T_j+1} is forecast by $\hat{y}_{j,T_j+1}(s)$ with $i \in s$ and vice versa. Hence, symmetry in grouping is assumed. In the next section we will relax this assumption.

2.3 Asymmetric grouping

Grouping estimators potentially increase forecast accuracy by trading gains in efficiency against increase in bias due to heterogeneity across panels. This trade-off may be different in each panel, and therefore the optimal grouping in terms of forecast accuracy may be different for each panel. An asymmetric grouping estimator allows for different panel groupings for each individual panel forecast, and hence allow for asymmetry.

To illustrate the potential gains of asymmetric grouping estimators versus symmetric grouping estimators, we consider the data generating process in (1) with $x_{it} = 1$ for $t = 1, \dots, T_i$, and $i = 1, \dots, N = 2$,

$$y_1 = \beta_1 + \varepsilon_1, \quad \varepsilon_1 \sim \text{i.i.d}(0, \sigma_1^2 I_{T_1}), \tag{6}$$

$$y_2 = \beta_2 + \varepsilon_2, \quad \varepsilon_2 \sim \text{i.i.d}(0, \sigma_2^2 I_{T_2}), \tag{7}$$

where $S = \{\{1\}, \{2\}, \{1, 2\}\}$ and the mean squared forecast errors equal

$$\rho_1(\{1\}) = \sigma_1^2/T_1, \quad \rho_1(\{1,2\}) = \frac{T_2^2}{(T_1+T_2)^2}(\beta_1 - \beta_2)^2 + \frac{\sigma_1^2 T_1 + \sigma_2^2 T_2}{(T_1+T_2)^2}, \tag{8}$$

$$\rho_2(\{2\}) = \sigma_2^2/T_2, \quad \rho_2(\{1,2\}) = \frac{T_1^2}{(T_1+T_2)^2}(\beta_1 - \beta_2)^2 + \frac{\sigma_1^2 T_1 + \sigma_2^2 T_2}{(T_1+T_2)^2}. \tag{9}$$

There are two cases in which a symmetric grouping achieves the lowest mean squared forecast error for both panels. First, it is optimal to forecast both y_{1,T_1+1} and y_{2,T_2+1} by pooling the two panels together if

$$(\beta_1 - \beta_2)^2 < \min\left(\left(\frac{2}{T_2} + \frac{1}{T_1}\right)\sigma_1^2 - \frac{1}{T_2}\sigma_2^2, \left(\frac{2}{T_1} + \frac{1}{T_2}\right)\sigma_2^2 - \frac{1}{T_1}\sigma_1^2\right), \quad (10)$$

which implies that it is optimal to group the two panels when the bias that arises from grouping is small relative to the error variances. Second, both panels benefit more the individual estimator if

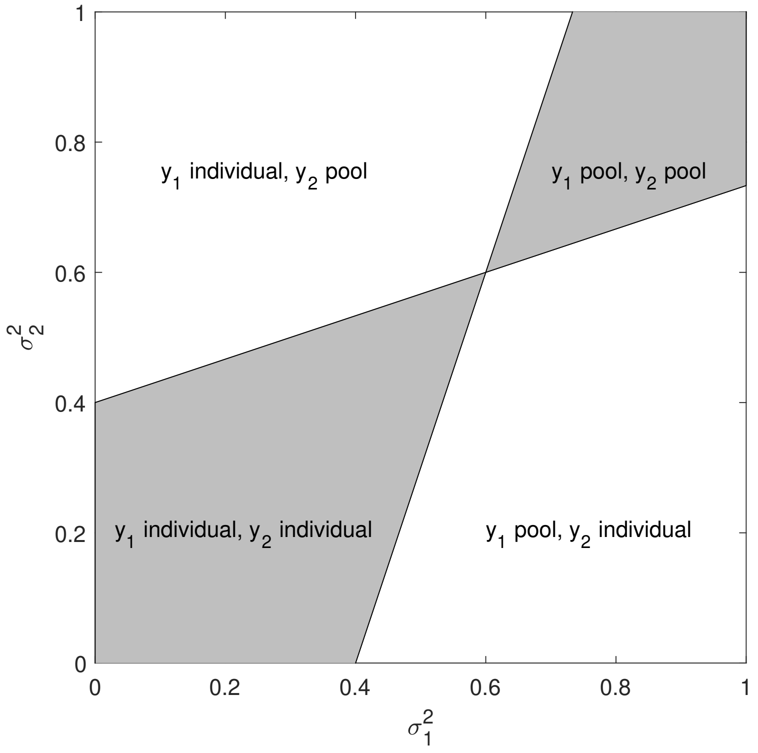
$$(\beta_1 - \beta_2)^2 > \max\left(\left(\frac{2}{T_2} + \frac{1}{T_1}\right)\sigma_1^2 - \frac{1}{T_2}\sigma_2^2, \left(\frac{2}{T_1} + \frac{1}{T_2}\right)\sigma_2^2 - \frac{1}{T_1}\sigma_1^2\right). \quad (11)$$

Since the optimal group structure does not change by forecasting in another panel, we consider the individual estimator as a special case of symmetric grouping. When the error variances are small relative to the bias, grouping does not result in more accurate forecasts.

Figure 1 shows the optimal group structure for different values of the error variances σ_1^2 and σ_2^2 when the bias term $(\beta_1 - \beta_2)^2 = 1$ and $T_1 = T_2 = 10$. The gray parameter space represents the error variances for which symmetric grouping is optimal. In the left lower corner both error variances are small relative to the bias term and each panel has its own group. In the right upper corner the error variances dominate the bias term and the pooled estimator is for both panels optimal. The point of intersection is determined by the sample sizes T_1 and T_2 and the magnitude of the bias. When the bias increases, the parameter space corresponding to pooling decreases and for a zero bias it is never optimal to forecast both panels by the individual estimator.

When there is enough variation in the error variances across the panels, Figure 1 shows that it is suboptimal to forecast each panel from the same group structure. Depending on the bias and the sample size, the asymmetric grouping can be optimal in a large part of the parameter space and seems by no means restricted to extreme cases. The idea of this paper is to develop an estimator that exploits this uncolored parameter space to gain in terms of forecast accuracy.

Figure 1: Optimal grouping over parameter space



This figure shows the optimal group structure for different values of the error variances σ_1^2 and σ_2^2 when the bias term $(\beta_1 - \beta_2)^2 = 1$ and $T_1 = T_2 = 10$.

2.4 Group selection

The asymmetric grouping estimator allows each panel forecast to be constructed from a different subset of panels. The subset s is treated as a hyperparameter in the forecast $\hat{y}_{i,T_i+1}(s)$, which we select by cross-validation.

The infeasible best estimator of s for forecasting y_{i,T_i+1} minimizes the mean squared forecast error and selects the optimal subset of panels $\hat{s}$ by

$$\hat{s} = \underset{s \in S_i}{\operatorname{argmin}}\, \rho_i(s), \tag{12}$$

where $\rho_i(s)$ is defined as the mean squared forecast error in (2) based on $\hat{\beta}_i = \hat{\beta}_i(s)$.

To obtain a feasible estimate of the optimal grouping strategy for each panel forecast we use cross-validation. A sample estimate $\hat{\rho}_i(s)$ of $\rho_i(s)$ is

$$\hat{\rho}_i(s) = \frac{1}{T_i} \sum_{t=1}^{T_i} e_{it}(s)^2, \tag{13}$$

where $e_{it}(s)$ denote the leave-one-out prediction residuals. These residuals are based on the leave-one-out estimator

$$\hat{\beta}_i^{-t}(s) = \left(\sum_{l \in s/i} X_l' X_l + \sum_{j \neq t} x_{ij} x_{ij}' \right)^{-1} \left(\sum_{l \in s/i} X_l' y_l + \sum_{j \neq t} x_{ij} y_{ij} \right), \tag{14}$$

where s/i denotes the set s without element i, by the formula

$$e_{it}(s) = y_{it} - x_{it} \hat{\beta}_i^{-t}(s) = \frac{\hat{\varepsilon}_{it}(s)}{1 - x_{it}' \left(\sum_{l \in s} X_l' X_l \right)^{-1} x_{it}}, \tag{15}$$

where $\hat{\varepsilon}_{it}(s) = y_{it} - x_{it} \hat{\beta}_i(s)$. The forecast for y_{i,T_i+1} equals

$$\hat{y}_{i,T_i+1} = x_{i,T_i+1}' \hat{\beta}_i(\hat{s}), \tag{16}$$

with $\hat{\beta}_i(s)$ as in (3), and $\hat{s} = \operatorname{argmin}_{s \in S_i} \hat{\rho}_i(s)$ with $\hat{\rho}_i(s)$ defined in (13). Algorithm 1 shows the three simple steps of forecasting with this asymmetric grouping estimator.

Algorithm 1 Asymmetric grouping estimator

```
for all $s \in S_i$ do
    $\hat{\beta}_i(s) = \left( \sum_{l \in s} X_l' X_l \right)^{-1} \sum_{l \in s} X_l' y_l$
end for
$\hat{s} \quad = \operatorname{argmin}_{s \in S_i} \sum_{t=1}^{T_i} \left( \frac{y_{it} - x_{it} \hat{\beta}_i(s)}{1 - x_{it}' \left( \sum_{l \in s} X_l' X_l \right)^{-1} x_{it}} \right)^2$
$\hat{y}_{i,T_i+1} = x_{i,T_i+1}' \hat{\beta}_i(\hat{s})$
```

2.4.1 Large number of panels

Since Algorithm 1 performs only one linear operation for each combination of panels, the estimator is reasonably fast when the number of available panels is small. However, the estimator iterates over all available 2^{N-1} combinations to forecast in one panel, which means that Algorithm 1 works when $N = 10$, but is computationally infeasible when $N = 50$.

To solve this problem, we propose an asymmetric grouping estimator that is feasible with a large number of panels by approximating the sample estimate of the mean squared forecast errors of each panel combination. This approximation allows for a sequential integer programming approach that does not need to explore every possible combination.

We estimate the mean squared forecast error $\rho_i(s)$ using the leave-one-out prediction residuals

$$v_{it}(s) = \frac{1}{k}(y_{it} - x_{it}'\hat{\beta}_i^{-t}(\{i\})) + \frac{1}{k}\sum_{j \neq i}^{N} w_{ij}(s)(y_{it} - x_{it}'\hat{\beta}_j(\{j\})), \tag{17}$$

where $k = |s|$, and the weights $w_{ij}(s)$, $j = 1, \ldots, i-1, i+1, \ldots, N$, equal 1 when $j \in s$, and zero otherwise. The residuals $v_{it}(s)$ in (17) approximate the residuals $e_{it}(s)$ in (13) by a linear approximation[1]

$$v_{it}(s) = \frac{1}{k-1}\sum_{j \neq i}^{N} w_{ij}(s) v_{it}(\{j\}), \tag{18}$$

where the panel-specific prediction residuals $v_{it}(\{j\})$ are specified as

$$v_{it}(\{j\}) = \frac{1}{k}(y_{it} - x_{it}'\hat{\beta}_i^{-t}(i)) + \frac{k-1}{k}(y_{it} - x_{it}'\hat{\beta}_j(\{j\})). \tag{19}$$

The panel-specific prediction residuals in (19) are constructed only from data in panel i and panel j. Therefore, this linear approximation enables the sample estimate of the mean squared forecast error to be written as

$$\hat{\rho}_i(s) = \frac{1}{T_i}\sum_{t=1}^{T_i} v_{it}(s)^2 = \frac{1}{T_i}(k-1)^{-2} w_i(s)' V_i w_i(s), \tag{20}$$

where $V_i = v_i v_i'$, with $v_i = (v_i(\{1\}), \ldots, v_i(\{i-1\}), v_i(\{i+1\}), \ldots, v_i(\{N\}))'$,

[1] Section 3.2 discusses the assumptions under which these approximations are valid.

$v_i(\{j\}) = (v_{i1}(\{j\}), \ldots, v_{iT}(\{j\}))'$, and $w_i(s) = (w_{i1}(s), \ldots, w_{iN}(s))'$.

To estimate the optimal subset of panels $\hat{s}$ we minimize the sample estimate of the mean squared forecast error. We iteratively minimize (20) for each value of k. The $s \in S_i$ with $k = 1$ corresponds to the sample estimate of the mean squared forecast error $\hat{\rho}_i(i)$ as defined in (13). For $k > 1$, we estimate the optimal combination by solving the optimization problem

$$\begin{aligned} \min \quad & (k-1)^{-2} w' V_i w, && (21) \\ & \sum_j w_j = k - 1, && (22) \\ & w \in \{0,1\}^{N-1}, && (23) \end{aligned}$$

which can be solved by integer programming as in Matsypura et al. (2018). We select the combination $\hat{s}$ with the lowest $\hat{\rho}_i(s)$ from the set of optimal combinations s of length $k = 1, \ldots, N$. Algorithm 2 outlines the forecasting steps with the asymmetric grouping estimator and a large number of panels.

Algorithm 2 Asymmetric grouping with large number of panels

```
for all $k = 2, \ldots, N$ do
    for all $j \neq i$ do
        $v_i(\{j\}) = \frac{1}{k} e_i(i) + \frac{k-1}{k}(y_i - x_i' \hat{\beta}_j(\{j\}))$
    end for
    $v_i = (v_i(\{1\}), \ldots, v_i(\{i-1\}), v_i(\{i+1\}), \ldots, v_i(\{N\}))'$
    $\min_w \rho_i^k(w) = (k-1)^{-2} w' v_i v_i' w$ s.t. $\sum_j w_j = k - 1$ and $w_j \in \{0,1\}$
    Set $\hat{s}$ according to $\hat{w}^k$ if $\rho_i^k(\hat{w}^k) < \rho_i^{k-1}(\hat{w}^{k-1})$
end for
$\hat{y}_{i,T_i+1} = x_{i,T_i+1}' \hat{\beta}_i(\hat{s})$
```

2.4.2 Huge number of panels

Algorithm 2 runs a series of integer programming problems to forecast with a large number of panels. Matsypura et al. (2018) show that these problems have NP-hard complexity, which means that there is no guarantee on a feasible computation time for data sets with a very large number of panels. This section proposes an alternative forecasting method, for which the computation time increases linearly in the number of panels, and therefore suits a huge number of panels.

Instead of estimating the mean squared forecast error for all possible combinations of panels, we first select a small set of panel combinations which contains a panel combination with a mean squared forecast error that is close or equal to the optimal one. This initial screening step is based on the conjecture that a panel combination that contains a panel that induces a forecast bias and variance in the mean squared forecast error, is only likely to be optimal when it also includes the panels that induce smaller bias and variance terms in the mean squared forecast error. This bias-variance trade-off for forecasting in panel i with panel l is captured by $\rho_i(\{j\})$.

When the number of panels N is large we propose an algorithm that only estimates the mean squared forecast errors of $2N$ panel combinations instead of 2^{N-1}. The initial screening step estimates the mean squared forecast errors $\rho_i(\{l\})$ for $l = 1, \ldots, N$. Subsequently, we run Algorithm 1 only for the N panel combinations consisting of panels corresponding to the k smallest $\rho_i(\{l\})$, with $k = 1, \ldots, N$. Algorithm 3 shows the computation steps of this asymmetric grouping estimator with large N.

Algorithm 3 Asymmetric grouping with huge number of panels

```
for $j = 1, \ldots, N$ do
    $\hat{\beta}_i(\{j\}) = \left(X_j' X_j\right)^{-1} X_j' y_j$
    $\hat{\rho}_i(\{j\}) = \sum_{t=1}^{T_i} \left(y_{it} - x_{it}\hat{\beta}_i(\{j\})\right)^2$
end for
for $k = 1, \ldots, N$ do
    $h$ = set of $k$ panels selected by smallest $\hat{\rho}_i(\{j\})$
    $\hat{\beta}_i(h) = \left(\sum_{l \in h} X_l' X_l\right)^{-1} \sum_{l \in h} X_l' y_l$
end for
$\hat{h} = \mathrm{argmin}_h \sum_{t=1}^{T_i} \left(\frac{y_{it} - x_{it}\hat{\beta}_i(h)}{1 - x_{it}'\left(\sum_{l \in h} X_l' X_l\right)^{-1} x_{it}}\right)^2$
$\hat{y}_{i,T_i+1} = x_{i,T_i+1}'\hat{\beta}_i(\hat{h})$
```

2.5 A shrinkage interpretation of grouping estimators

The grouping estimator defined in (3) borrows information from other panels when the set s contains more than one panel. This idea is directly related to shrinkage estimators in panel data. To illustrate this, we rewrite the grouping estimator to

$$\hat{\beta}_i(s) = \sum_{l \in s} (U_s^{-1} Q_l) \hat{\beta}_l(\{l\}) = (U_s^{-1} Q_i) \hat{\beta}_i(\{i\}) + \sum_{l \in s/i} (U_s^{-1} Q_l) \hat{\beta}_l(\{l\}), \quad (24)$$

where $Q_l = X_l' X_l$ and $U_s = \sum_{l \in s} Q_l$. The grouping estimator in (24) is a weighted average of the individual panel estimator in panel i and a combination of the individual estimators in the other panels included in s. The grouping estimator shrinks the panel estimates to the weighted average of other panels in s.

From a Bayesian perspective, this shrinkage is defined in terms of prior distributions. The posterior mean of a parameter is shrinked from the sample mean to the prior mean. Consider the panel regression model in (1) and assume normally distributed error terms. The natural conjugate prior distribution for β_i and diffuse prior for σ_i^2 are specified as

$$p(\beta_i | \sigma_i^2) \sim \mathrm{N}(b_i, \sigma_i^2 B_i), \quad p(\sigma_i^2) \propto \sigma_i^{-2}, \quad (25)$$

where b_i defines the prior mean and $\sigma_i^2 B_i$ the covariance matrix of the prior distribution for β_i. This prior assumes no shrinkage when b_i and B_i do not use cross-sectional information. The marginal posterior distribution of β_i is

$$\beta_i \sim \mathrm{t}(\tilde{\beta}_i, \tilde{\sigma}_i^2 (X_i' X_i + B_i^{-1})^{-1}, T_i), \quad (26)$$

$$\tilde{\beta}_i = (X_i' X_i + B_i^{-1})^{-1} (X_i' y_i + B_i^{-1} b_i), \quad (27)$$

$$\tilde{\sigma}_i^2 = \frac{1}{N} ((y_i - X_i \tilde{\beta}_i)'(y_i - X_i \tilde{\beta}_i) + (b_i - \tilde{\beta}_i)' B_i^{-1} (b_i - \tilde{\beta}_i), \quad (28)$$

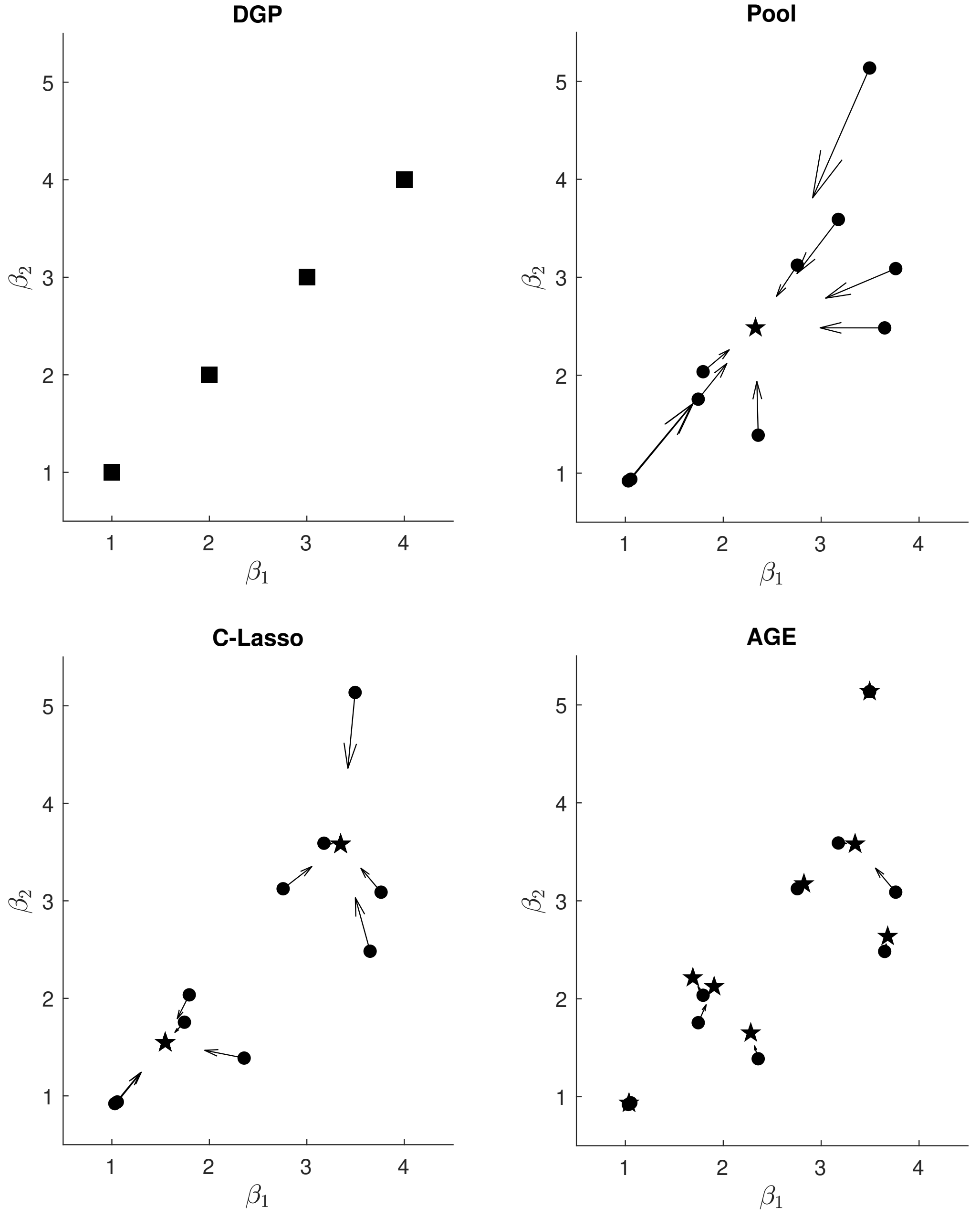
see e.g. Chapter 3 of Zellner (1971), from which follows that the posterior mean of β_i equals $\hat{\beta}_i(s)$ if

$$b_i = \left(\sum_{l \in s/i} X_l' X_l \right)^{-1} \sum_{l \in s/i} X_l' y_l, \qquad B_i = \left(\sum_{l \in s/i} X_l' X_l \right)^{-1}. \tag{29}$$

The prior parameters in (29) shrink the individual panel estimator in panel i to the ordinary least squares estimator using the other panels included in s. A symmetric grouping only allows for groups of panels for which the parameter estimates are shrunken to a common group-specific parameter value. The asymmetric grouping estimator also allows for different prior parameters for each panel, which implies that the shrinkage direction can be different for all panel-specific parameter estimates.

Figure 2 shows the shrinkage behaviour of different grouping estimators. The coefficients are estimated on linear panel data simulated from DGP 3 with $N = 10$ and $R^2 = 0.9$, discussed in Section 4. The squares in the first panel of Figure 2 display the true values of the 4 different regression parameter settings used in the DGP. The circles in the other panels display the individual estimates, while the stars denote one of the grouping estimates. The pooled estimator with $s = \{1, \ldots, N\}$ for all panels shrinks all panel-specific coefficients to a common estimate. This approach does not allow for parameter heterogeneity across panels. The second panel of Figure 2 shows the individual panel estimates together with their weighted average estimated by the pooled estimator. A symmetric grouping estimator does allow for heterogeneity by shrinking the individual panel estimates to common cluster estimates. The third panel in Figure 2 shows that the coefficients shrink to two different clusters. The asymmetric grouping estimator does not necessarily shrink panel coefficients to identical values. The final panel in Figure 2 shows that the two coefficient vectors in the lower left corner shrink together, the coefficient vector in the upper right corner do not move at all, and other coefficients shrink only in each others direction.

Figure 2: Parameter estimates grouping estimators



The figure in the upper left corner shows the parameter values of β_1 and β_2 for data simulated from DGP 3 in Section 4. The upper right figure shows the individual panel estimates (circles) with the pooled estimate (stars), the lower left figure with the C-lasso estimates (stars), and the lower right with the asymmetric grouping estimates (stars).

3 Theoretical results

So far, we have introduced our new asymmetric grouping estimator. In this section we derive under which conditions our estimator is useful and under which conditions our approximations can be applied. The results are based on the panel regression model in (1). This model assumes strictly exogenous explanatory variables. Under this assumption, the individual estimator is unbiased and we can derive explicit expressions for the bias-variance trade-off of the proposed estimators.

3.1 Mean squared forecast error grouping estimators

Lemma 1 provides an expression for the mean squared forecast error $\rho_i(s)$ for the forecast $\hat{y}_{i,T_i+1}(s)$ in (4).

Lemma 1 *The mean squared forecast error $\rho_i(s)$ for $\hat{y}_{i,T_i+1}(s)$ is*

$$\begin{aligned}\rho_i(s) &= B_i(s) + V_i(s) &(30)\\ &= \left(x'_{i,T_i+1} U_s^{-1} \sum_{l\in s} Q_l(\beta_l - \beta_i) \right)^2 + \sum_{l\in s} \sigma_l^2 x'_{i,T_i+1} U_s^{-1} Q_l U_s^{-1} x_{i,T_i+1},\end{aligned}$$

with $s \in S_i$ and where we denote $Q_l = X_l' X_l$ and $U_s = \sum_{l\in s} Q_l$.

The proof is given in Appendix A.

Lemma 1 shows that the mean squared forecast error is a composition of a bias and a variance term. The first term in (30) represents the squared forecast bias that may arise from using multiple panels to forecast panel i. For the individual estimator $s = \{i\}$ the bias term is zero. The bias term also equals zero when the coefficients are homogeneous across panels, that is $\beta = \beta_i$ for all i. In this case both the individual and the pooled estimator with $s = \{1, \dots, N\}$ results in zero forecast bias. The bias is large when the coefficients are strongly heterogeneous and $|s| > 1$.

The second term in (30) represents the forecast variance. The variance for $s = \{i\}$ equals $\sigma_i^2 x'_{i,T_i+1} Q_i^{-1} x_{i,T_i+1}$, which increases in the error variance

σ_i^2. To show that pooling may decrease the forecast variance, we rewrite the variance term as

$$V_i(s) = \sigma_i^2 x'_{i,T_i+1} U_s^{-1} x_{i,T_i+1} + \sum_{l \in s/i} (\sigma_l^2 - \sigma_i^2) x'_{i,T_i+1} U_s^{-1} Q_l U_s^{-1} x_{i,T_i+1}. \quad (31)$$

For homogeneous error variances, $\sigma^2 = \sigma_i^2$, the forecast variance is minimized by the pooled estimator. In case of heterogeneous error variances, the forecast variance of $\hat{y}_{i,T_i+1}(s)$ benefits from pooling with panels l for which $\sigma_l^2 < \sigma_i^2$. The forecast variance also decreases in the number of total observations in s via U_s^{-1}.

The bias-variance trade-off in the mean squared forecast error determines whether heterogeneous grouping results in optimal forecasts. Theorem 1 formalizes this intuition to a general setting.

Theorem 1 *For two panels i and j with data generating process* (1)*, the mean squared forecast errors as defined in* (2) *satisfy*

$$\rho_j(\{ij\}) < \rho_j(\{j\}) \text{ and } \rho_i(\{ij\}) > \rho_i(\{i\}), \quad (32)$$

when the following condition holds,

$$\begin{aligned} \sigma_i^2 U_s Q_i^{-1} U_s - Q_j(\beta_i - \beta_j)(\beta_i - \beta_j)' Q_j \prec \sigma_i^2 Q_i + \sigma_j^2 Q_j \\ \prec \sigma_j^2 U_s Q_j^{-1} U_s - Q_i(\beta_i - \beta_j)(\beta_i - \beta_j)' Q_i, \end{aligned} \quad (33)$$

where $A \prec B$ means that $B - A$ is a positive definite matrix, $Q_i = X_i' X_i$, $Q_j = X_j' X_j$ and $U_s = Q_i + Q_j$.

The proof is given in Appendix B.

Theorem 1 shows that there exist general conditions under which a heterogeneous grouping is optimal. This optimality depends on the bias from grouping, the error variances and the regressor matrices. Along the same lines as in the proof for Theorem 1, we can simplify the expression in (33) by assuming that $Q_l \to T_l\Sigma$,

$$\left(\frac{2\sigma_1^2 - \sigma_2^2}{T_2} + \frac{\sigma_1^2}{T_1}\right)\Sigma \prec \Sigma(\beta_i - \beta_j)(\beta_i - \beta_j)'\Sigma \prec \left(\frac{2\sigma_2^2 - \sigma_1^2}{T_1} + \frac{\sigma_2^2}{T_2}\right)\Sigma, \quad (34)$$

or by using the stricter assumption that $X_i = X_j$,

$$(3\sigma_i^2 - \sigma_j^2)Q_i \prec Q_i(\beta_i - \beta_j)(\beta_i - \beta_j)'Q_i \prec (3\sigma_j^2 - \sigma_i^2)Q_i. \quad (35)$$

When the variation in the error variances across the panels is large enough, the panel with the smaller error variance achieves the highest forecast accuracy by not grouping together, whereas the panel with the larger error variance improves in forecast accuracy by grouping. Applying a symmetric group estimator in this setting inevitably leads to a loss in accuracy.

3.2 Forecast algorithms

Section 2.4 estimates the optimal grouping by cross-validation on the mean squared forecast errors. However, for a large number of panels the computation of the forecast errors of all possible panel combinations becomes infeasible and we rely on approximations. This section shows under which assumptions these approximations are valid.

The heterogeneous grouping estimator in Algorithm 2 approximates the leave-one-out-residuals used in Algorithm 1 to allow for a large number of panels. Theorem 2 shows that the approximation error of Algorithm 2 is small for balanced panels in which the sample covariance matrix of the regressors is approximately the same in each panel.

Theorem 2 *Assume that $\frac{1}{T_i}X_i'X_i \approx \frac{1}{T_j}X_j'X_j$ and $T_i = T_j$ for all $i = 1, \ldots, N$ and $j = 1, \ldots, N$. We have that*

$$v_{it}(s) \approx e_{it}(s). \quad (36)$$

The proof is given in Appendix C.

To deal with the large number of panels, Theorem 2 assumes balanced panel data and the same covariance matrix for the regressors in each panel.

Both assumptions can be checked before the analysis.

Algorithm 3 is proposed for panel data sets with a huge number of panels. The approximation error of this algorithm is small if the difference between the mean squared forecast error between the selected panel combination by Algorithm 1 and the selected panel combination by Algorithm 3 is small. Theorem 3 shows that this difference is due to ignoring the cross bias terms in the mean squared forecast error.

Theorem 3 *Assume that* $\frac{1}{T_i}X_i'X_i \approx \frac{1}{T_j}X_j'X_j$ *for all* $i = 1, \ldots, N$ *and* $j = 1, \ldots, N$. *We have that*

$$\rho_i(s) \approx \frac{\sum_{l \in s} T_l^2 \rho_i(\{l\}) + C_i(s)}{\sum_{l \in s} T_l^2}, \tag{37}$$

where

$$C_i(s) = \sum_{l \in s} \sum_{k \in s/l} T_l T_k x_{i,T_i+1}' (\beta_l - \beta_i)(\beta_k - \beta_i)' x_{i,T_i+1}. \tag{38}$$

The proof is given in Appendix D.

Theorem 3 shows that the mean squared forecast error of a combination of panels indeed consists of a weighted average of the individual bias-variance trade-offs of each panel. However, the mean squared forecasts error consists of an additional term, $C_i(s)$, that represents the cross bias terms of the panels. Algorithm 3 gives identical results to Algorithm 1 when the sample covariance matrix of the regressors in each panel are the same and the cross bias terms are sufficiently small. In the next section we will illustrate the properties of the asymmetric grouping estimator and its approximations in a Monte Carlo study.

4 Simulation Study

To demonstrate the importance of asymmetric grouping estimators in panel data forecasting, we consider varying degrees of panel heterogeneity in a simulation study. The performance of the proposed forecasting algorithms with asymmetric grouping are evaluated by the mean squared forecast error and compared to symmetric grouping estimators.

4.1 Design Monte Carlo experiments

The Monte Carlo experiments have the following data generating process

$$y_{it} = \sum_{l=1}^{3} x_{itl}\beta_{il} + \varepsilon_{it}, \quad \varepsilon_{it} \sim N(0, \sigma_i^2), \quad i = 1, \dots, N, \quad t = 1, \dots, T+1,$$

where $x_{it1} = 1$ and x_{it2} and x_{it3} are independently generated from a standard normal distribution. The size of the variance of the errors σ_i^2 sets the R^2 in each panel equal to a pre-specified value. The parameters are estimated on $t = 1, \dots, T$, and used to forecast $t = T + 1$ for all $i = 1, \dots, N$. The sample size in each panel is $T = 20$, the number of panels $N = 10$, and the error term $\varepsilon_{i,T+1}$ is set to zero.

We vary the degree of heterogeneity in the coefficients with four different specifications.

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

DGP 2 (Weakly heterogeneous)

$$\beta_{i1} = \beta_{i2} = \begin{cases} 1, & i = 1, \dots, [N/2], \\ 3, & i = [N/2] + 1, \dots, N, \end{cases} \tag{39}$$

$$\beta_{i3} = \begin{cases} 1, & i = 1, \dots, [N/3], \\ 3, & i = [N/3] + 1, \dots, N. \end{cases} \tag{40}$$

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

DGP 3 (Strongly heterogeneous)

$$\beta_{i1} = \beta_{i2} = \begin{cases} 1, & i = 1, \dots, [N/4], \\ 2, & i = [N/4] + 1, \dots, [N/2], \\ 3, & i = [N/2] + 1, \dots, [3N/4], \\ 4, & i = [3N/4] + 1, \dots, N, \end{cases} \tag{41}$$

$$\beta_{i3} = \begin{cases} 1, & i = 1, \dots, [N/5], \\ 2, & i = [N/5] + 1, \dots, [2N/5], \\ 3, & i = [2N/5] + 1, \dots, [3N/5], \\ 4, & i = [3N/5] + 1, \dots, N. \end{cases} \tag{42}$$

DGP 4 (Completely heterogeneous) $\beta_{il} = N^{-1} \times i \times l$ for all i and l.

We generate one-step ahead forecasts with the three asymmetric grouping algorithms discussed in Section 2.4. These methods are compared to symmetric grouping estimators: pooled estimator, individual estimator, C-Lasso, and an oracle estimator. The pooled estimator produces a forecast $\hat{y}_{i,T_i+1}$ based on $s = \{1, \dots, N\}$, the individual estimator on $s = \{i\}$, and C-lasso estimates the s under the restriction that y_{i,T_i+1} is forecast by $\hat{y}_{i,T_i+1}(s)$ with $j \in s$ if and only if y_{j,T_j+1} is forecast by $\hat{y}_{j,T_j+1}(s)$ with $i \in s$. We implement C-Lasso as proposed by Su et al. (2016), for $K = 1, \dots, 5$ number of groups and tuning parameters $c_\lambda = \{0.125, 0.25, 0.5, 1, 2\}$, and the estimated grouping selected by an information criterion. The oracle estimator assumes that the ‘true’ group structure in the data generating process is known, and forecasts y_{i,T_i+1} by $\hat{y}_{i,T_i+1}(s)$ with s defined by the data generating process above.

4.2 Simulation results

Table 1 shows the mean squared forecast error over 1.000 replications of the four data generating processes with $R^2 = 0.4$ and $R^2 = 0.9$. The reported mean squared forecast errors of the oracle estimator, asymmetric grouping estimators, C-lasso, and the pooling estimator (Pool) are relative to the mean squared forecast error of the individual estimator.

Table 1: Monte Carlo Simulation Results

	$R^2 = 0.4$				$R^2 = 0.9$			
DGP	1	2	3	4	1	2	3	4
Oracle	0.084	0.214	0.514	1.000	0.088	0.214	0.514	1.000
AGE^S	0.398	0.517	0.487	0.509	0.397	0.609	0.661	0.691
AGE^L	0.378	0.523	0.491	0.510	0.375	0.599	0.683	0.714
AGE^H	0.393	0.540	0.507	0.512	0.388	0.556	0.702	0.705
C-Lasso	0.087	0.687	0.661	0.785	0.091	0.880	1.899	1.962
Pool	0.084	0.731	0.652	0.877	0.088	8.839	7.788	10.844

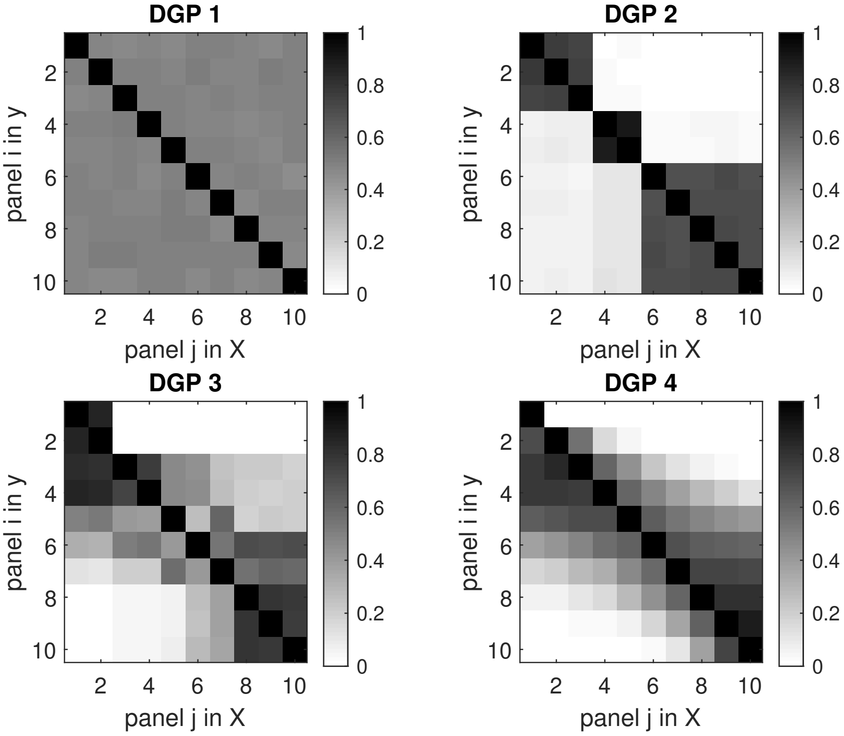
Note: this table shows the mean squared forecast error over 1,000 replications of four data generating processes defined in Section 4, with $R^2 = \{0.4, 0.9\}$ and $T = 20$ and $N = 10$. The mean squared forecast errors of the asymmetric grouping estimators defined in Algorithm 1 (AGE^S), Algorithm 2 (AGE^L), and Algorithm 3 (AGE^H), together with the oracle, C-Lasso, and pooled estimator (Pool) are relative to the mean squared forecast error of the individual estimator.

The asymmetric grouping estimator outperforms the pooled and individual estimator in almost all cases. When the coefficients are completely homogeneous, which is the case under the first data generating process, the pooled estimator is more accurate. In all other cases under consideration, the asymmetric grouping estimator has a lower mean squared forecast error. Even when the coefficients are completely heterogeneous, the individual estimator is outperformed by a wide margin.

The forecasts of the asymmetric grouping estimator are more accurate than the C-lasso in all heterogeneous data generating processes. The C-lasso estimates the 'true' underlying panel grouping in the data generating process, while the asymmetric grouping estimator estimates the optimal bias-variance trade-off. The latter approach results in substantial improvements in forecast accuracy, especially when the signal-to-noise ratio is high. A comparison between the oracle estimator and the asymmetric grouping estimator shows that, even when the underlying grouping is correctly identified, the asymmetric grouping estimator performs better when the panels are sufficiently heterogeneous.

The accuracy of the approximate asymmetric group estimators in Algorithm 2 and Algorithm 3 is close to the standard asymmetric grouping

Figure 3: Simulations: asymmetric panel grouping

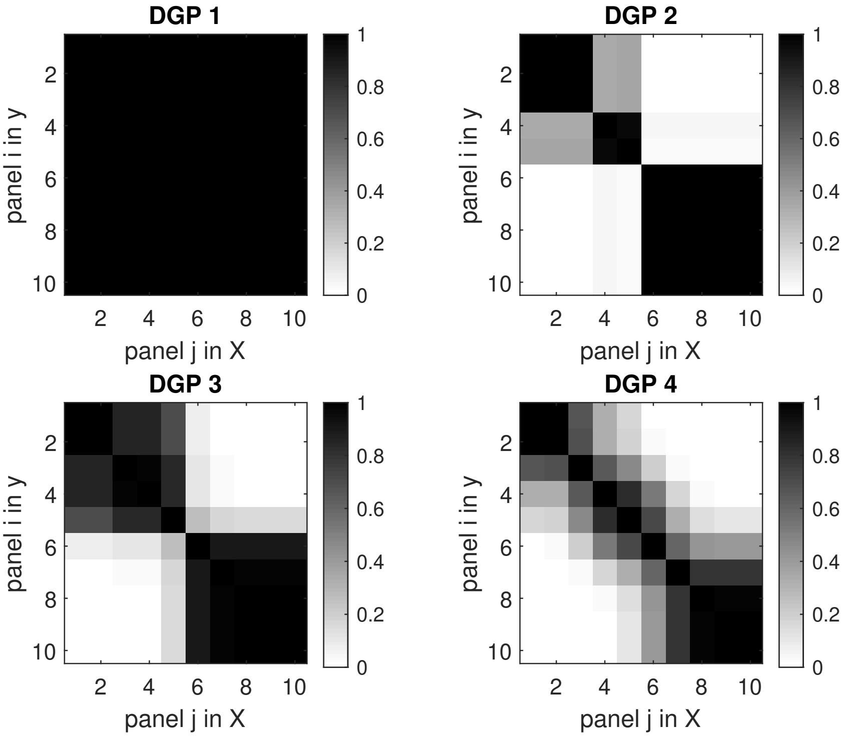


Percentage replications $\hat{y}_{i,T_i+1}(s)$ is based on $j \in s$ by the asymmetric grouping estimator defined in Algorithm 1 (AGE^S), for the four data generating processes with $R^2 = 0.9$.

estimator defined by Algorithm 1. Although the standard asymmetric group estimator performs slightly better under most settings, the difference is small in any setting, and the approximate algorithms also outperform the benchmark methods in all settings but the completely homogeneous coefficient data generating process.

Figure 3 shows the percentage of replications the panel forecasts in the rows are constructed with the panel data in the columns, for the standard asymmetric grouping estimator defined in Algorithm 1. The forecast for $y_{i,T+1}$ in row i is based on a combination of panels s that contains the panel in column j. The group probabilities are shown for the four different data generating processes with $R^2 = 0.9$. The experiments with $R^2 = 0.4$ and the approximate grouping estimators show similar results.

Figure 4: Simulations: symmetric panel grouping



Percentage replications $\hat{y}_{i,T_i+1}(s)$ is based on $j \in s$ by C-lasso, for the four different data generating processes with $R^2 = 0.9$.

We find for different data generating processes different groupings. Figure 3 shows a diagonal structure of the group probabilities for DGP 1. Due to the homogeneous coefficients, a symmetric grouping is most accurate. Each panel has a probability close 50% to be used for another panel. The group probabilities for DGP 2 also suggest a symmetric grouping. Since the panels 1-3, panels 4 and 5, and panels 6-10 have the same data generating process, DGP 2 reveals a block structure.

The Monte Carlo experiments show that the asymmetric grouping estimator can indeed identify asymmetric grouping structures. A close look at the graphs show that they are not completely symmetric with respect to the diagonal. The estimated grouping for DGP 3 shows that for forecasting panel 1 and 2 the panels 3 and 4 are not used, while for forecasting panels 3 and 4 the panels 1 and 2 are used. This asymmetric group structure is caused by the higher noise level in panel 3 and 4 compared to the noise level in panel 1 and 2. We also find evidence for asymmetric grouping in DGP 4, where for instance the first panel is used to forecast in panels 1-8, while forecasts for panel 1 only use data in the panel itself.

Figure 4 shows the group probabilities for the C-lasso. The completely homogeneous data generating process is correctly identified, and also the weakly heterogeneous group structure in DGP 2 is correctly estimated in most Monte Carlo replications. However, for the strongly heterogeneous panel data in DGP 3 and DGP4, the grouping estimates of C-lasso are strongly biased and very different from the estimated asymmetric groupings in Figure 3.

5 Empirical application

To illustrate the usefulness of our asymmetric grouping estimator we consider panel forecasts of market leverage of U.S. firms. Market leverage forecasts are a key input in corporate capital structure decisions. The bankruptcy costs and tax savings related to the capital structure of firms, are not only of interest to the firms itself, but also to policymakers and financial market agents trying to understand the market risk. Frank and Goyal (2009) construct a panel data set to examine the important predictors for market leverage. They identify six core predictors that account for 27% of the variation in leverage, while the remaining analyzed predictors only add a further 2%. Smith et al. (2019) study the same data in a pooled regression model with Bayesian variable selection under breaks. We study forecast performance conditional on the six core predictors for market leverage in a panel regression model that allows for heterogeneity across panels.

5.1 Data and methods

We use the data of Frank and Goyal (2009), and refer to their paper for a detailed definition of all variables. To have a balanced panel, we run the forecast exercise for the 172 publicly traded American firms without missing observations between 1963 and 2003. We consider a heterogeneous panel regression model,

$$LV_{it} = \alpha_i + x'_{i,t-1}\beta_i + \varepsilon_{it}, \quad i = 1, \dots, N, \quad t = 1, \dots, T, \tag{43}$$

where LV_{it} is the market-based leverage ratio measure, total debt to market assets, of firm i at year t, $x'_{i,t-1}$ is a 6×1 vector of lagged core predictors, and $N = 172$ and $T = 41$. The core predictors of market leverage are: median industry leverage, market-to-book assets ratio, tangibility, profits, log of assets, and expected inflation.

We use an expanding window to produce 15 forecasts for each panel, from 1989 to 2003. The forecasts are constructed based on four different methods: the individual estimator, pooled estimator, C-lasso, and asymmetric grouping estimators. We follow the C-Lasso settings in Su et al. (2016), with $K = 1, \dots, 5$ number of groups and tuning parameters $c_\lambda = \{0.125, 0.25, 0.5, 1, 2\}$, and the estimated grouping selected by an information criterion. Each estimation window mean-centers and scales the dependent and independent variables to a variance of one. This means that we estimate group-specific effects, up to a panel-specific fixed effect and scaling factor.

The forecasting algorithms proposed in Section 2.4 target data sets with different numbers of panels. To illustrate the performance of each algorithm, we apply them to subsets of the data as if there is only information available of a small number of firms. Algorithm 1 is applied to the 10 firms with the largest asset value at 2003, Algorithm 2 to the 25 firms with the largest asset value at 2003, and Algorithm 3 to all 172 firms. For each set of panels we calculate the average mean squared forecast error over the panels for the forecast from 1989 to 2003, from 1994 to 2003, and 1999 to 2003.

5.2 Results

Table 2 shows the mean squared forecast errors for different panel data sets and different hold-out samples. Values below one favor the grouping methods over forecasts based on the individual panel estimator. We find that asymmetric grouping improves upon the benchmark methods in all but one

Table 2: Mean squared forecast errors market leverage

	$N = 10$			$N = 25$			$N = 172$		
τ	15	10	5	15	10	5	15	10	5
AGE	$\underline{0.731}$	$\underline{0.798}$	$\underline{0.804}$	$\underline{0.880}$	$\underline{0.901}$	0.866	$\underline{0.925}$	$\underline{0.948}$	$\underline{0.942}$
C-Lasso	0.841	0.869	0.851	1.007	0.929	$\underline{0.848}$	1.055	1.000	1.044
Pool	0.841	0.869	0.851	1.019	0.947	0.871	1.204	1.083	1.129

Note: this table shows the mean squared forecast errors of the asymmetric group estimator (AGE), C-Lasso, and the pooling estimator (Pool), relative to the mean squared forecast error of the individual estimator. For $N = 10$, AGE forecasts with Algorithm 1, for $N = 25$ with Algorithm 2, and for $N = 172$ with Algorithm 3. The number of forecast periods is indicated by τ. The minimum values in each column are underlined.

setting. The mean squared forecast error based on only the five forecast periods for 1999 to 2003 in the panel with the 25 largest firms is minimized by C-lasso. The symmetric grouping estimated by C-lasso performs at least as good as the pooling estimator, in all settings under consideration. However, both symmetric grouping estimators perform worse than the individual estimator in several cases.

The asymmetric grouping estimator outperforms benchmark methods for all panel sizes. Although the sample covariance matrix of the regressors show substantial variation across panels, the integer programming approach and the screening approach achieve competitive forecast performance. Especially for the panel including all 172 firms the improvements in forecast accuracy are substantial, with the asymmetric grouping estimator the only method outperforming the individual estimator. Note that the mean squared forecast error estimates in the large panel are based on 172 $\times$ τ forecasts and therefore most reliable. Unreported results show that for $N = 25$ the performance of the asymmetric grouping estimator based on Algorithm 3 instead of 2 is slightly worse. For $N = 10$ the asymmetric grouping estimator based on Algorithms 2 and 3 perform even worse than the C-Lasso and Pooled estimator. Hence, in small samples it is to be preferred to avoid approximations of the cross-validation approach.

Figure 5 shows the estimated panel groupings by the asymmetric grouping estimator. In every setting we find forecasts for firms that do not use

Table 3: Statistics group size market leverage forecasts

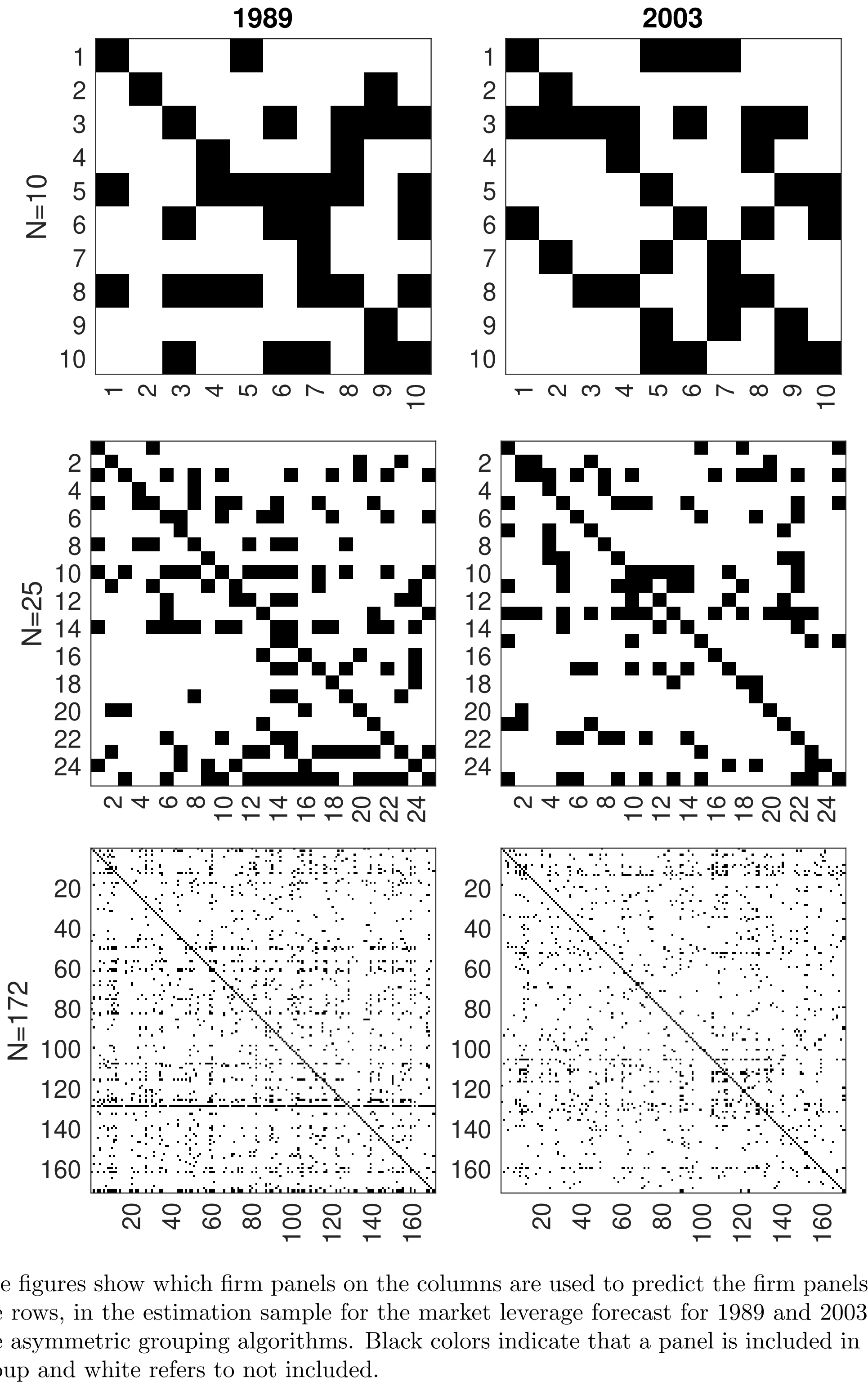
		$N = 10$		$N = 25$		$N = 172$	
		1989	2003	1989	2003	1989	2003
AGE	min	1	1	1	1	1	1
	max	7	7	16	15	153	52
	mean	3.6	3.5	6.1	5.2	12.3	9.2
	median	3.0	3.5	5.0	4.0	7.0	6.5
	std	2.3	1.6	4.3	3.5	16.7	9.3
C-Lasso	min	10	10	25	7	34	64
	max	10	10	25	18	138	108

This table shows statistics of the estimated panel group size in the asymmetric grouping estimator (AGE) and the symmetric grouping estimator (C-Lasso).

information from other panels. This boils down to using the individual estimator. There are no estimated groupings that include all panels, the pooled estimator. Apart from the individual estimators, there are no settings in which the asymmetric grouping estimator estimates a symmetric grouping. Table 3 shows the wide variety in panel group sized across panel forecasts. The estimated panel groupings also show substantial variation over time. The groupings estimated to forecast market leverage in 1989 are different from groupings for market leverage in 2003. When we increase the number of available panels, the estimated groupings also increase.

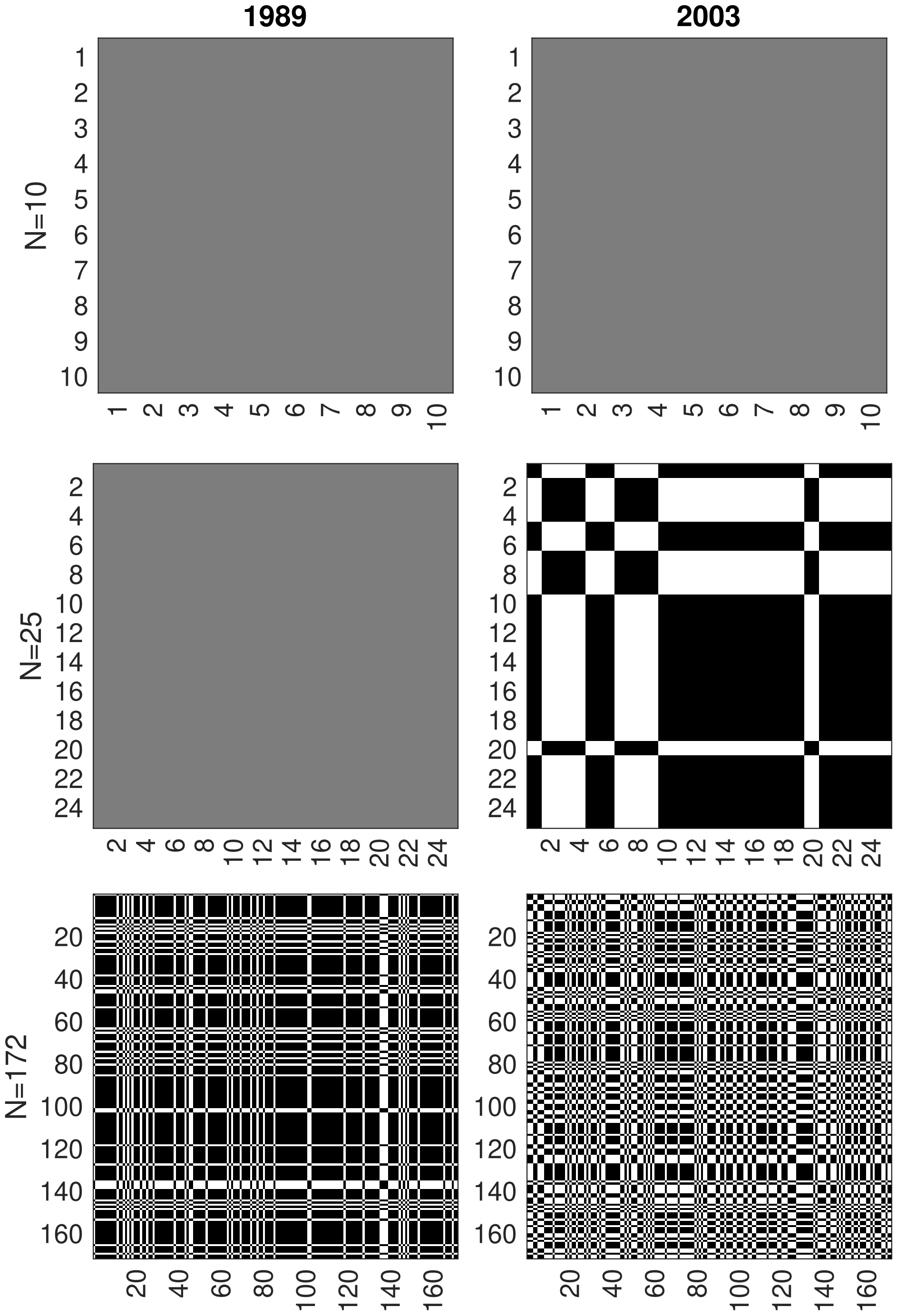
Figure 6 shows the estimated symmetric panel groupings by C-lasso. The symmetric grouping estimator does not find heterogeneity in the small panel of 10 firms and estimates only one group. The same holds for the estimation window for the market leverage forecasts in 1989. C-lasso finds two groups of panels with different predictor coefficients in the other settings. Table 3 shows the panel group sizes. The increase in forecast performance of the asymmetric grouping estimators relative to the symmetric grouping estimator in Table 2, may be explained by the fact that a symmetric grouping is not able to balance the bias-variance trade-off in an optimal way.

Figure 5: Asymmetric panel grouping of firm market leverage



The figures show which firm panels on the columns are used to predict the firm panels on the rows, in the estimation sample for the market leverage forecast for 1989 and 2003 by the asymmetric grouping algorithms. Black colors indicate that a panel is included in the group and white refers to not included.

Figure 6: Symmetric panel grouping of firm market leverage



The figures show which firm panels on the columns are used to predict the firm panels on the rows, in the estimation sample for the market leverage forecast for 1989 and 2003 by C-lasso. Black colors indicate that a panel is included in the group and white refers to not included.

In sum, we conclude that the asymmetric grouping estimator turns out to be a useful estimator for forecasting market leverage. The resulting grouping structure clearly suggests that asymmetric grouping is often better for forecasting than symmetric grouping.

6 Conclusion

Exploiting cross-sectional information in panel data potentially improves forecast accuracy when the number of observations in each panel is small. This paper constructs panel-specific forecasts based on a asymmetric grouping estimator, that allows for an asymmetric bias-variance trade-off across panels. The estimator can be extended to the setting where the number of panels is large. We show that asymmetric grouping is optimal in terms of mean squared forecast under a broad range of conditions. A simulation study and an empirical application support these findings. Although the asymmetric estimator is only discussed in a linear setting, the estimator may also be useful in nonlinear panel models. A clear disadvantage of applying the methods in nonlinear panel data models is however that the cross-validation may take much more computing time which limits the practical applicability of the approach.

References

Ando, T. and Bai, J. (2016). Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics*, 31(1):163–191.

Baltagi, B. H., Bresson, G., and Pirotte, A. (2008). To pool or not to pool? In *The econometrics of panel data*, pages 517–546. Springer.

Bester, C. A. and Hansen, C. B. (2016). Grouped effects estimators in fixed effects models. *Journal of Econometrics*, 190(1):197–208.

Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184.

Chib, S. (2008). Panel data modeling and inference: a bayesian primer. In *The econometrics of panel data*, pages 479–515. Springer.

Danilov, D. and Magnus, J. R. (2004). On the harm that ignoring pretesting can cause. *Journal of Econometrics*, 122(1):27–46.

Desbordes, R., Koop, G., and Vicard, V. (2018). One size does not fit all... panel data: Bayesian model averaging and data poolability. *Economic Modelling*, 75:364–376.

Frank, M. Z. and Goyal, V. K. (2009). Capital structure decisions: which factors are reliably important? *Financial management*, 38(1):1–37.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Science & Business Media.

Jin, S. and Su, L. (2013). A nonparametric poolability test for panel data models with cross section dependence. *Econometric Reviews*, 32(4):469–512.

Juhl, T. and Lugovskyy, O. (2014). A test for slope heterogeneity in fixed effects models. *Econometric Reviews*, 33(8):906–935.

Kasahara, H. and Shimotsu, K. (2009). Nonparametric identification of finite mixture models of dynamic discrete choices. *Econometrica*, 77(1):135–175.

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

Maddala, G. S., Trost, R. P., Li, H., and Joutz, F. (1997). Estimation of short-run and long-run elasticities of energy demand from panel data using shrinkage estimators. *Journal of Business & Economic Statistics*, 15(1):90–100.

Matsypura, D., Thompson, R., and Vasnev, A. L. (2018). Optimal selection of expert forecasts with integer programming. *Omega*, 78:165–175.

McLachlan, G. and Peel, D. (2000). Finite mixture models, willey series in probability and statistics.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, pages 1–32.

Pesaran, M. H. and Yamagata, T. (2008). Testing slope homogeneity in large panels. *Journal of Econometrics*, 142(1):50–93.

Smith, S. C., Timmermann, A., and Zhu, Y. (2019). Variable selection in panel models with breaks. *Journal of Econometrics*.

Su, L., Shi, Z., and Phillips, P. C. (2016). Identifying latent structures in panel data. *Econometrica*, 84(6):2215–2264.

Swamy, P. A. (1970). Efficient inference in a random coefficient regression model. *Econometrica: Journal of the Econometric Society*, pages 311–323.

Wang, W., Phillips, P. C., and Su, L. (2018). Homogeneity pursuit in panel data models: Theory and application. *Journal of Applied Econometrics*, 33(6):797–815.

Wang, W., Zhang, X., and Paap, R. (2015). What is a good strategy for parameter estimation and forecasting in panel regressions? *Working paper*.

Zellner, A. (1971). *An introduction to Bayesian inference in econometrics*, volume 156. Wiley New York.

A Proof of Lemma 1

The mean squared forecast error equals

$$\rho_i(s) = E\left[\left(x'_{i,T_i+1}\beta_i - x'_{i,T_i+1}U_s^{-1}\sum_{l\in s}Q_l\beta_l - x'_{i,T_i+1}U_s^{-1}\sum_{l\in s}X'_l\varepsilon_l\right)^2\right] \quad (44)$$

$$= E\left[\left(x'_{i,T_i+1}U_s^{-1}\sum_{l\in s}Q_l(\beta_l-\beta_i) - x'_{i,T_i+1}U_s^{-1}\sum_{l\in s}X'_l\varepsilon_l\right)^2\right] \quad (45)$$

$$= \left(x'_{i,T_i+1}U_s^{-1}\sum_{l\in s}Q_l(\beta_l-\beta_i)\right)^2 + E\left[\left(x'_{i,T_i+1}U_s^{-1}\sum_{l\in s}X'_l\varepsilon_l\right)^2\right], \quad (46)$$

where the second line uses $\beta_i = U_s^{-1}U_s\beta_i$, and the third line uses $E\left[X'_l\varepsilon_l\right] = 0$. The second term in (46) equals

$$E\left[x'_{i,T_i+1}U_s^{-1}\sum_{l\in s}(X'_l\varepsilon_l)\sum_{l\in s}(\varepsilon'_lX_l)U_s^{-1}x_{i,T_i+1}\right] = \quad (47)$$

$$\sum_{l\in s}x'_{i,T_i+1}U_s^{-1}X'_lE[\varepsilon_l\varepsilon'_l]X_lU_s^{-1}x_{i,T_i+1} = \quad (48)$$

$$\sum_{l\in s}\sigma_l^2x'_{i,T_i+1}U_s^{-1}Q_lU_s^{-1}x_{i,T_i+1}, \quad (49)$$

where the second line uses $E[\varepsilon_i\varepsilon'_j] = 0$ for $i \neq j$, and the third $E[\varepsilon_i\varepsilon'_i] = \sigma_i^2I$.

B Proof of Theorem 1

Using Lemma 1, the mean squared forecast errors for panel i satisfy,

$$\rho_i(\{ij\}) - \rho_i(\{i\}) = x'_{i,T_i+1}U_s^{-1}A_iU_s^{-1}x_{i,T_i+1}, \text{ with} \quad (50)$$

$$A_i = Q_j(\beta_i-\beta_j)(\beta_i-\beta_j)'Q_j + \sigma_i^2Q_i + \sigma_j^2Q_j - \sigma_i^2U_sQ_i^{-1}U_s, \quad (51)$$

from which follows that $\rho_i(\{ij\}) - \rho_i(\{i\}) > 0$ if $A_i \succ 0$. In the same way,

$$\rho_j(\{ij\}) - \rho_j(\{j\}) = x'_{j,T_j+1} U_s^{-1} A_j U_s^{-1} x_{j,T_j+1}, \text{ with} \tag{52}$$
$$A_j = Q_i(\beta_i - \beta_j)(\beta_i - \beta_j)' Q_i + \sigma_i^2 Q_i + \sigma_j^2 Q_j - \sigma_j^2 Q_s Q_j^{-1} Q_s, \tag{53}$$

which gives $\rho_j(\{ij\}) - \rho_j(\{j\}) < 0$ if $A_j \prec 0$. Combining the two cases,

$$\sigma_i^2 U_s Q_i^{-1} U_s - Q_j(\beta_i - \beta_j)(\beta_i - \beta_j)' Q_j \prec \sigma_i^2 Q_i + \sigma_j^2 Q_j \tag{54}$$
$$\prec \sigma_j^2 U_s Q_j^{-1} U_s - Q_i(\beta_i - \beta_j)(\beta_i - \beta_j)' Q_i. \tag{55}$$

C Proof of Theorem 2

$$\hat{\beta}_i^{-t}(s) = \left(\sum_{l \in s/i} X_l' X_l + \sum_{j \neq t} x_{ij} x_{ij}' \right)^{-1} \left(\sum_{l \in s/i} X_l' y_l + \sum_{j \neq t} x_{ij} y_{ij} \right) \tag{56}$$
$$= \left(\sum_{l \in s/i} Q_l + Q_i^{-t} \right)^{-1} \left(\sum_{l \in s/i} Q_l Q_l^{-1} X_l' y_l + Q_i^{-t} (Q_i^{-t})^{-1} \sum_{j \neq t} x_{ij} y_{ij} \right)$$
$$= \left(\sum_{l \in s/i} Q_l + Q_i^{-t} \right)^{-1} \left(\sum_{l \in s/i} Q_l \hat{\beta}_l(\{l\}) + Q_i^{-t} \hat{\beta}_i^{-t}(\{i\}) \right) \tag{57}$$
$$= W_i^{-t}(s) \hat{\beta}_i^{-t}(\{i\}) + \sum_{j \neq i}^{N} W_{ij}(s) \hat{\beta}_j(\{j\}), \tag{58}$$

where $Q_l = X_l' X_l$ and $Q_i^{-t} = \sum_{j \neq t} x_{ij} x_{ij}'$. Define the matrices $W_i^{-t}(s) = \left(\sum_{l \in s/i} Q_l + Q_i^{-t} \right)^{-1} Q_i^{-t}$ and $W_{ij}(s) = \left(\sum_{l \in s/i} Q_l + Q_i^{-t} \right)^{-1} Q_j I[j \in s]$. It follows that

$$e_{it}(s) = W_i^{-t}(s)(y_{it} - x_{it}' \hat{\beta}_i^{-t}(\{i\})) + \sum_{j \neq i}^{N} W_{ij}(s)(y_{it} - x_{it}' \hat{\beta}_j(\{j\})). \tag{59}$$

When $\frac{1}{T_i}X_i'X_i \approx \frac{1}{T_j}X_j'X_j$ and $T_i = T_j$ for all i, j, we have $W_i^{-t}(s) \approx \frac{1}{k}$ and $W_{ij}(s) \approx \frac{1}{k}I[j \in s]$ with $k = |s|$. So under this assumption

$$v_{it}(s) = \frac{1}{k}(y_{it} - x_{it}'\hat{\beta}_i^{-t}(\{i\})) + \frac{1}{k}\sum_{j \neq i}^{N} I[j \in s](y_{it} - x_{it}'\hat{\beta}_j(\{j\})) \approx e_{it}(s).$$

D Proof of Theorem 3

The mean squared forecast error equals

$$\rho_i(s) = \left(x_{i,T_i+1}'U_s^{-1}\sum_{l \in s} Q_l(\beta_l - \beta_i)\right)^2 + \sum_{l \in s}\sigma_l^2 x_{i,T_i+1}'U_s^{-1}Q_lU_s^{-1}x_{i,T_i+1},$$

where the first term represents the forecast bias and the second term the forecast variance. Rewrite the forecast bias to

$$\sum_{l \in s} x_{i,T_i+1}'U_s^{-1}Q_l(\beta_l - \beta_i)(\beta_l - \beta_i)'Q_lU_s^{-1}x_{i,T_i+1} + \tag{60}$$

$$\sum_{l \in s}\sum_{k \in s/l} x_{i,T_i+1}'U_s^{-1}Q_l(\beta_l - \beta_i)(\beta_k - \beta_i)'Q_kU_s^{-1}x_{i,T_i+1} \tag{61}$$

where first line contains the squared bias terms and the second line the cross bias terms, which we denote by $C_i(s)$. We have

$$\begin{aligned}\rho_i(s) &= \sum_{l \in s} x_{i,T_i+1}'U_s^{-1}Q_l\left((\beta_l - \beta_i)(\beta_l - \beta_i)' + \sigma_l^2 Q_l^{-1}\right)Q_lU_s^{-1}x_{i,T_i+1} + C_i(s)\\ &= \sum_{l \in s} x_{i,T_i+1}'U_s^{-1}Q_lA_i(\{l\})Q_lU_s^{-1}x_{i,T_i+1} + C_i(s) \qquad (62)\end{aligned}$$

where $A_i(\{l\}) = (\beta_l - \beta_i)(\beta_l - \beta_i)' + \sigma_l^2 Q_l^{-1}$. Assume that $\frac{1}{T_i}X_i'X_i \approx \frac{1}{T_j}X_j'X_j$ and $T_i = T_j$ for all $i = 1, \dots, N$ and $j = 1, \dots, N$, it follows that

$$\rho_i(s) \approx \frac{\sum_{l \in s} T_l^2 \rho_i(\{l\}) + \sum_{k \in s/l} T_lT_k x_{i,T_i+1}'(\beta_l - \beta_i)(\beta_k - \beta_i)'x_{i,T_i+1}}{\sum_{l \in s} T_l^2}, \tag{63}$$

where we use that $\rho_i(\{l\}) = x_{i,T_i+1}'A_i(\{l\})x_{i,T_i+1}$.