ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics

Bachelor thesis Econometrics and Operations Research
Quantitative Finance

Generalized least squares pooling averaging: A simulation study[*]

Author
M. van der Linden
(441844)

Supervisor
W. Wang

Second assessor
J.W.N. Reuvers

July 7, 2019

Abstract In this paper we propose a generalized least squares (GLS) pooling averaging estimator to address the bias-efficiency trade-off dilemma in estimating heterogeneous linear panel data regression models. By means of a Monte Carlo simulation study, we provide insights in the performance of our GLS pooling averaging estimator compared to the existing ordinary least squares (OLS) pooling averaging in different situations, specifically in the presence of cross-unit error correlations. Our simulation results show that GLS pooling averaging convincingly outperforms OLS pooling averaging in panels with a large number of observations per unit and high error correlations, particularly under moderate noise. We further find that pooling averaging is more robust than individual OLS estimation under endogeneity.

[*] The views stated in this thesis are those of the author and not necessarily those of Erasmus School of Economics or Erasmus University Rotterdam.

Contents

1 Introduction

The empirical analysis of panel data is of growing importance due to its increasing availability. In heterogeneous panel data regression models, the potentially large number of coefficients that need to be estimated can lead to a substantial loss in efficiency. However, assuming homogeneity between units by applying pooled coefficient estimation could lead to a substantial bias in the individual coefficient estimates and therefore to incorrect empirical inference making. This typical trade-off between bias and efficiency describes an important dilemma in panel data model estimation.

Existing literature proposes multiple approaches for the estimation of heterogeneous coefficients in panel data models. For example, Swamy (1970) are the first to specifically address this problem, for which they propose the random coefficient model (RCM) in which the average effect of (potentially) heterogeneous random coefficients is estimated using feasible generalized least squares (GLS). They prove that this estimator is consistent and asymptotically efficient. Pesaran and Smith (1995) propose the mean group estimator, which, opposed to Swamy (1970), equally averages the random coefficients from each individual regression for each unit. Another approach is considered by Su et al. (2016), who propose their classifier-Lasso (C-Lasso) to determine latent group structures in panel data models in which coefficients of units are homogeneous within groups and heterogeneous between groups. They further show that the C-Lasso estimator enables consistent estimation of the number of groups.

Wang et al. (2019) propose to average the estimators or forecasts of different pooling strategies in a linear panel data regression model with fixed, but potentially heterogeneous coefficients, with appropriate weights. By selecting the weights with Mallows C_p-criterion, following Hansen (2007), their Mallows pooling averaging estimator (MPA) makes an explicit bias-variance tradeoff which is optimal with respect to the mean square (forecast) error (MS(F)E). This is different from other methods, which focus on other criteria than the MS(F)E, like unbiasedness or consistency, and therefore cannot guarantee a minimum MS(F)E. The MPA estimator further opposes the average-effect estimators of Swamy (1970) and Pesaran and Smith (1995) by leading to potentially different unit-specific coefficient estimates, while the average-effect estimators lead to a common coefficient estimate for all units. Unit-specific estimates are particularly relevant in the explanation or forecasting of the individual units. An other advantage of the MPA estimator is that it enables heterogeneity to exist in any pattern, not requiring a correct specification of the group structure. By means of a Monte Carlo simulation study, Wang et al. (2019) compare the performance of the MPA estimator to, among others, the methods mentioned above for multiple combinations of data generating processes (DGPs). They find that the MPA estimator often outperforms the other estimators in heterogeneous panels with a moderate signal-to-noise ratio and also performs well in homogeneous panels or panels with a lower signal-to-noise ratio.

For the estimation of the individual coefficients, Wang et al. (2019) consider ordinary least squares (OLS), which is equivalent to applying system OLS in a seemingly unrelated regression (SUR) model. The estimation of SUR models for panel data is proposed by Zellner (1962), who prove the asymptotic efficiency of the feasible GLS estimator in this setting, of which system OLS is a specific case. The relative performance of feasible GLS compared to OLS depends on the variance-covariance structure of the error terms. In the absence of error correlations, the finite sample properties favour the OLS estimator, which is also more robust. Feasible GLS is more efficient in the case of cross-unit correlation between the error terms: higher correlations lead to greater efficiency gains by (feasible) GLS relative to OLS. Wang et al. (2019) assume uncorrelated error terms and consider the corresponding consistent variance estimators of the OLS estimators, for which the optimality of the MPA estimator is established. Liu et al. (2016) propose to average feasible GLS estimators of models with heteroskedastic error terms, but in a single equation setting. In this case, the weight selection by Mallows C_p-criterion is also proven to be optimal with respect to the MS(F)E.

In our research, we consider the MPA estimator of Wang et al. (2019), but examine its application in the presence of contemporaneous cross-unit error correlations. We propose an extension of the MPA estimator to (feasible) GLS and compare the performance of this GLS MPA estimator to that of the OLS MPA estimator of Wang et al. (2019). We ask ourselves whether GLS MPA outperforms OLS MPA in terms of MS(F)E due to the gain in efficiency by GLS in the presence of contemporaneous cross-unit error correlations—and, if so, in which situations. To the best of our knowledge, no literature exists on pooling averaging estimation under GLS in panel data models nor on the performance of OLS MPA under cross-unit error correlations. We expect that GLS MPA outperforms OLS MPA in samples with a large number of observations per unit and with high cross-unit error correlations. Further, in our case, the variance estimators of the OLS estimators considered by Wang et al. (2019) are not consistent. Hence, we also examine the performance of OLS MPA when applying a variance estimator which is consistent in the presence of cross-unit error correlations.

To answer our research question, we perform an extensive Monte Carlo simulation study with different DGPs. Wang et al. (2019) find that the bias-efficiency trade-off of OLS MPA is jointly determined by multiple factors, including the signal-to-noise ratio, the degree of cross-unit coefficient heterogeneity, the number of observations per unit, and the number of explanatory variables. In our case, the size of the cross-unit error correlations also determines the relative efficiency of GLS compared to OLS and the sample size for each unit is particularly important for feasible GLS. Therefore, we compare the squared risk, which is closely related to the MS(F)E, of GLS MPA, OLS MPA, (individual) OLS and feasible GLS, and pooling forecasts for different DGP combinations with respect to coefficient heterogeneity, sample size, cross-unit error correlations, and signal-to-noise ratio. To compare the robustness of the estimators with respect to the MS(F)E, we also consider DGP combinations in which the exogeneity condition is violated.

We must note that the optimality results of the MPA estimator by Wang et al. (2019) are not established for the (feasible) GLS estimator nor in presence of cross-unit error correlations. Our research is not intended to provide a theoretical derivation of the (potential) optimality properties of GLS MPA and OLS MPA in this situation, but we merely attempt to gain insights into the relative performances of these estimators by means of our simulation study.

Our simulation results show that GLS MPA convincingly outperforms OLS MPA in panels with a large number of observations per unit and high cross-unit error correlations. In particular, for a moderate signal-to-noise ratio and a large sample size, GLS MPA outperforms all considered estimators. For a moderate number of observations, OLS MPA often performs best. Further, in accordance with the findings of Wang et al. (2019), we find that none of the considered estimators performs best in all situations. For instance, the pooling estimator performs best in most homogeneous panels and (individual) OLS or feasible GLS outperform MPA in completely heterogeneous panels in many situations. The considered variance estimators for OLS MPA and GLS MPA often give similar results, suggesting that the importance of their consistency is negligible. In case of endogeneity and a large number of observations, GLS MPA outperforms all estimators in almost all heterogeneous panels, but often the difference with other estimators is small. MPA generally outperforms OLS under endogeneity, proving to be more robust with respect to MS(F)E in this situation.

The remainder of this paper is as follows. In Section 2 we explain our model set-up and derive and explain the estimators. Specifically, we derive the pooling averaging estimators under OLS and (feasible) GLS and provide an intuitive theoretical MS(F)E comparison. Further, we explain the Mallows criterion for the weight selection and our considered variance estimators. Section 3 provides an explanation of our Monte Carlo simulation and in Section 4 we discuss the simulation results. Section 5 concludes this paper with an overview and discussion of our findings and some suggestions for future research.

2 Estimators

2.1 Model set-up

We consider the linear panel data model with heterogeneous coefficients for each unit i at period $t = 1, \ldots, T$,

$$y_i = X_i \beta_i + \varepsilon_i, \quad i = 1, \ldots, N, \tag{2.1.1}$$

where $y_i = (y_{i1}, \ldots, y_{iT})'$ is a $T \times 1$ vector of dependent variables, $X_i = (x_{i1}', \ldots, x_{iT}')'$ is a $T \times K$ matrix of explanatory variables, $\beta_i = (\beta_{i1}, \ldots, \beta_{iK})'$ a $K \times 1$ vector of fixed regression coefficients, and $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iT})'$ a $T \times 1$ vector of error terms. The first coefficient, β_{i1}, corresponds to the intercept, such that $x_{1,it} = 1$ for all i, t. Further, the coefficients β_i are assumed to be fixed, but they can vary over units i. This can be written more compactly for all N units as the SUR model

$$y = X\beta + \varepsilon, \tag{2.1.2}$$

where $y = (y_1', \ldots, y_N')'$, $X = \mathrm{diag}(X_1, \ldots, X_N)$, $\beta = (\beta_1', \ldots, \beta_N')'$, and $\varepsilon = (\varepsilon_1', \ldots, \varepsilon_N')'$.

Wang et al. (2019) assume that the error terms ε_i are uncorrelated conditional on X_i for all i. In this paper, we allow for contemporaneous correlation between the error terms to compare the performance of the OLS and GLS pooling averaging estimators explained in Section 2.2 and 2.3, respectively. That is, we assume that

$$\varepsilon \sim \mathrm{D}(0, \Sigma \otimes I_T) = \mathrm{D}(0, \Omega) \tag{2.1.3}$$

where D denotes some distribution, Σ is the $N \times N$ contemporaneous variance-covariance matrix of the error terms with typical elements $\sigma_{ij} = \mathrm{E}[\varepsilon_{it}\varepsilon_{jt}]$, for $i, j = 1, \ldots, N$ and for all t, and I_T the $T \times T$ identity matrix. In practice, the presence of cross-unit error correlations can be tested, for instance, by the Lagrange Multiplier test proposed by Baltagi et al. (2012). We further assume that the explanatory variables are strictly exogenous, within and across units: $\mathrm{E}[x'_{it}\varepsilon_{jt}] = 0$ for $i, j = 1, \ldots, N$ and for all t; and correct specification of the model in (2.1.2) to ensure that both the OLS and GLS estimators are consistent.

2.2 OLS pooling averaging estimator

We first consider the two extremes in panel data model estimation by OLS: estimating the coefficients of the individual units by (system) OLS in the SUR model, which allows for complete heterogeneity; and pooled estimation of the coefficients, which completely ignores possible heterogeneity. We further consider restricting some of the coefficients to be equal while allowing others to vary across different units. By averaging the restricted estimators, Wang et al. (2019) propose an estimator that makes an explicit trade-off between the gained efficiency from pooling and bias due to potential heterogeneity.

Provided that each individual model is correctly specified and all regressors are within-unit exogenous, unbiased estimation of the individual coefficients by system OLS in (2.1.2) is obtained by $\hat{\beta} = (\hat{\beta}'_1, \ldots, \hat{\beta}'_N)' = (X'X)^{-1}X'y$. This is equivalent to equation-by-equation OLS estimation, $\hat{\beta}_i = (X'_iX_i)^{-1}X'_iy_i$ for all i, if there are no restrictions imposed on the coefficients. However, the OLS estimator can be inefficient due to the large number of coefficients and because no cross-unit variation is taken into consideration.

Pooled estimation of the coefficients, which completely ignores possible heterogeneity across units, restricts the coefficient β_i to be equal for all i. This is equivalent to transforming the matrix X to $X_* = (X'_1, \ldots, X'_N)'$, and then applying OLS as above, such that $b = (X'_*X_*)^{-1}X'_*y$, a $K \times 1$ vector. The pooled estimator for all units is $\hat{\beta}_{\text{pool}} = (\imath \otimes b)$, where $\imath$ is the $N \times 1$ unit vector. This method can lead to a substantial bias in the (individual) coefficient estimates, but is generally more efficient.

A compromise between the two estimators above is to introduce a specific pooling strategy by restricting some of the coefficients to be the same. By imposing equality restrictions of the form

$$R_m\beta = 0 \tag{2.2.1}$$

on the coefficients vector β in (2.1.2), we get the restricted OLS estimator

$$\hat{\beta}_{(m)} = P_m\hat{\beta} = (I_{NK} - (X'X)^{-1}R'_m(R_m(X'X)^{-1}R'_m)^{-1}R_m)\hat{\beta}, \tag{2.2.2}$$

where R_m and P_m are the restriction matrix and the resulting projection matrix, respectively, under pooling strategy m.

Wang et al. (2019) propose to average estimators or forecasts of M different pooling strategies $P_m\hat{\beta}$ with appropriate weights to obtain an optimal bias-efficiency trade-off, defining their pooling averaging estimator as

$$\hat{\beta}(w) = \sum_{m=1}^{M} w_m\hat{\beta}_{(m)} = \sum_{m=1}^{M} w_mP_m\hat{\beta} = P(w)\hat{\beta}. \tag{2.2.3}$$

Here, $P(w) = \sum_{m=1}^{M} w_mP_m$ is an $NK \times NK$ matrix and $w = (w_1, \ldots, w_M)'$ is a vector of weights, such that $w \in \mathcal{W} = \{w \in [0,1]^M : \sum_{m=1}^{M} w_m = 1\}$. Wang et al. (2019) show that weight selection by minimizing the appropriate Mallows C_p-criterion is optimal with respect to the squared risk, or MS(F)E, resulting in their MPA estimator. Because the possible number of imposing restrictions on β, and thus the number of possible pooling strategies, can be large depending on the number of units and the number of explanatory variables, a preliminary step of model screening is desirable. Wang et al. (2019) use the classifier-Lasso (C-Lasso) proposed by Su et al. (2016) for this purpose, under which the optimality of the weight selection still holds. We give a more detailed explanation of the weight selection and preliminary model screening in Section 2.5.

2.3 GLS pooling averaging estimator

Contrary to Wang et al. (2019), we allow for contemporaneous error correlations across different units. In this particular case, GLS estimation can be more efficient than OLS estimation. The GLS estimator is given by

$$\begin{aligned}\hat{\beta}_{\text{GLS}} &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y \\ &= (X'(\Sigma^{-1} \otimes I_T)X)^{-1}X'(\Sigma^{-1} \otimes I_T)y.\end{aligned} \tag{2.3.1}$$

Consistency of the GLS estimator requires strict exogeneity of the explanatory variables, following the assumption in Section 2.1. For feasible estimation, it must further hold that $\Sigma \succ 0$ and that $X'\Omega^{-1}X$ is nonsingular. Because the matrix Ω is generally unknown in practice, we substitute it with a consistent estimate to obtain a feasible GLS estimator, which we explain in Section 2.6. We note that for an accurate estimate of Ω, a sufficiently large sample size is required. Therefore, the feasible GLS estimator may have the potential to outperform the OLS estimator only for larger T.

Because GLS is equivalent to OLS in a transformed model, we can show that the OLS pooling averaging estimator proposed by Wang et al. (2019) and explained in Section 2.2 can be extended to GLS. That is, letting A be an invertible transformation matrix such that

$$\text{var}(A\varepsilon) = A\,\text{var}(\varepsilon)A' = A\Omega A' = I_{NT},$$

then it holds that $\Omega = (A'A)^{-1}$. In this case, applying OLS in the transformed model

$$Ay = AX\beta + A\varepsilon$$

gives the efficient GLS estimator

$$\begin{aligned}\hat{\beta}_{\text{GLS}} &= ((AX)'AX)^{-1}(AX)'Ay \\ &= (X(A'AX))^{-1}X'(A'A)y \\ &= (X'\Omega^{-1}X)^{-1})^{-1}X'\Omega^{-1}y,\end{aligned}$$

which is the best linear unbiased estimator (BLUE), as proven by Zellner (1962). This can be extended to estimation under coefficient restrictions. Given the restriction $R_m\beta = 0$ and following (2.2.2), we get the restricted GLS estimator

$$\begin{aligned}\hat{\beta}_{(m),\text{GLS}} &= (I_{NK} - ((AX)'(AX))^{-1}R_m'(R_m((AX)'(AX))^{-1}R_m')^{-1}R_m)\hat{\beta}_{\text{GLS}} \\ &= (I_{NK} - (X'\Omega^{-1}X)^{-1}R_m'(R_m(X'\Omega^{-1}X)^{-1}R_m')^{-1}R_m)\hat{\beta}_{\text{GLS}} \\ &= P_{m,\text{GLS}}\hat{\beta}_{\text{GLS}}.\end{aligned} \tag{2.3.2}$$

Opposed to the restricted OLS estimator in (2.2.2), the matrix $P_{m,\text{GLS}}$ depends on the matrix Ω, which is unknown in practice. We therefore replace it with the same consistent estimate as for the estimator $\hat{\beta}_{\text{GLS}}$ in (2.3.1) to get a feasible estimator. We can then extend the OLS pooling averaging estimator in (2.2.3) to GLS in a similar manner, giving

$$\hat{\beta}_{\text{GLS}}(w) = \sum_{m=1}^{M} w_m \hat{\beta}_{(m),\text{GLS}} = \sum_{m=1}^{M} w_m P_{m,\text{GLS}} \hat{\beta}_{\text{GLS}} = P_{\text{GLS}}(w)\hat{\beta}_{\text{GLS}}. \tag{2.3.3}$$

For clarity we proceed by denoting $\hat{\beta}_{\text{OLS}}(w) = P_{\text{OLS}}(w)\hat{\beta}_{\text{OLS}}$ and $\hat{\beta}_{\text{GLS}}(w) = P_{\text{GLS}}(w)\hat{\beta}_{\text{GLS}}$ as the pooling averaging estimators under OLS and GLS, respectively, and, for notational convenience, $\hat{\beta}(w) = P(w)\hat{\beta}$ when no distinction is necessary.

2.4 MS(F)E comparison under OLS and GLS estimation

We follow a part of the theoretical derivation of Wang et al. (2019) to motivate the application of pooling averaging. Specifically, we motivate a pooling averaging estimator under GLS instead of the OLS by comparing their MS(F)E, or squared risk. This provides insights into the bias-efficiency trade-off of an estimator or a forecast.

The following derivations are meant to give an intuitive motivation, assuming that the weights of the pooling averaging estimator are fixed. Because in practice these weights are estimated from the data, meaning that they are correlated with the explanatory variables and coefficient estimators, the practical implications of our assumption are less clear. However, more formal derivations are beyond the scope of our research and we refer to Wang et al. (2019) for the implications of random weights. The same holds for the covariance matrix Ω, required for the GLS estimator, which in practice needs to be estimated from the data to obtain a feasible GLS estimator. In the following derivations, we assume that Ω is known.

The MSFE of forecast $\hat{y}$ of the model defined in (2.1.2) can be written as

$$\begin{aligned} \text{MSFE}^*(\hat{y}) &= \text{E}[(y - \hat{y})(y - \hat{y})'] \\ &= \text{E}[(X\beta + \varepsilon - X\hat{\beta})(X\beta + \varepsilon - X\hat{\beta})'] \\ &= \text{E}[(X(\beta - \hat{\beta}) + \varepsilon)(X(\beta - \hat{\beta}) + \varepsilon))'] \\ &= \text{E}[X(\beta - \hat{\beta})(\beta - \hat{\beta})'X'] + \Omega, \end{aligned} \tag{2.4.1}$$

where $\hat{\beta}$ is some estimator for the true coefficient vector β and the last step holds (asymptotically) due to the exogeneity condition. We omit the Ω term in further analysis because it is fixed and identical under both OLS and GLS and therefore irrelevant for the comparison of the estimators. Defining $||\theta||_H^2 = \theta' H \theta$ for any vector θ and nonnegative definite matrix H and by taking the trace of the last expression in (2.4.1), we find the following alternative expression for the MSFE:

$$\begin{aligned} \text{MSFE}(\hat{y}) &= \text{tr}(\text{E}[X(\beta - \hat{\beta})(\beta - \hat{\beta})'X']) \\ &= \text{E}[\text{tr}(X(\beta - \hat{\beta})(\beta - \hat{\beta})'X')] \\ &= \text{E}[\text{tr}((\beta - \hat{\beta})'X'X(\beta - \hat{\beta}))] \\ &= \text{E}[(\beta - \hat{\beta})'X'X(\beta - \hat{\beta})] \\ &= \text{E}\,||\hat{\beta} - \beta||_{X'X}^2. \end{aligned} \tag{2.4.2}$$

Here, we follow the cyclic property of the trace and the property that the trace of a scalar expectation equals the expectation of that trace. The last expression of (2.4.2) defines the

MSFE that we consider in further analysis, following Wang et al. (2019). This expression is particularly relevant in the appropriate Mallows C_p-criterion discussed in Section 2.5, as it is equivalent to the (scaled) squared risk, or expected squared loss.

The scaling matrix $X'X$ in (2.4.1) gives the MSFE expression for forecast accuracy: the expected "scaled sum of squared errors", as defined by Mallows (1973); while replacing $X'X$ by the $NT \times NT$ identity matrix I_{NT} gives the MSE expression for the accuracy of the coefficient estimator $\hat{\beta}$. To generalize the squared risk to both forecasting and coefficient estimation accuracy, we denote the scaling matrices $X'X$ and I_{NT} by H. The choice of H depends on the application, specifically for the weight selection of the pooling averaging estimator by the Mallows C_p-criterion.

For the pooling averaging estimators $\hat{\beta}(w)$ in (2.2.3) and (2.3.3), for OLS and GLS, respectively, assuming fixed weights w and fixed Ω and using the trace properties mentioned above, we can write the (asymptotic) MS(F)E, or squared risk, as

$$\begin{aligned}
R_H(w) &= \mathrm{E}\,||\hat{\beta}(w) - \beta||_H^2 \\
&= \mathrm{E}[(\hat{\beta}(w) - \beta)'H(\hat{\beta}(w) - \beta)] \\
&= \mathrm{E}[\mathrm{tr}((\hat{\beta}(w) - \beta)'H(\hat{\beta}(w) - \beta))] \\
&= \mathrm{tr}(\mathrm{E}[(\hat{\beta}(w) - \beta)(\hat{\beta}(w) - \beta)']H) \\
&= \mathrm{tr}(\mathrm{var}(\hat{\beta}(w))H) + \mathrm{tr}((\mathrm{E}[\hat{\beta}(w)] - \beta)(\mathrm{E}[\hat{\beta}(w)] - \beta)'H) \\
&= \mathrm{tr}(\mathrm{var}(\hat{\beta}(w)H) + \mathrm{tr}(\mathrm{E}[\hat{\beta}(w)] - \beta)'H(\mathrm{E}[\hat{\beta}(w)] - \beta)) \\
&= \mathrm{tr}(\mathrm{var}(\hat{\beta}(w))H) + ||\,\mathrm{E}[\hat{\beta}(w)] - \beta||_H^2 \\
&= \mathrm{tr}(P(w)\,\mathrm{var}(\hat{\beta})P'(w)H) + ||P(w)\beta - \beta||_H^2.
\end{aligned} \tag{2.4.3}$$

The last step holds because both the OLS and GLS estimator for β are unbiased. This expression shows the bias-efficiency trade-off of the estimator $\hat{\beta}(w)$, with the left term accounting for the variance and the right term for the bias. As pointed out by Wang et al. (2019), the last expression in (2.4.3) further shows that the MS(F)E depends on the degree of heterogeneity of the true coefficients β, the variance of the individual estimator $\hat{\beta}$, the scaling matrix H, and the combination of weights and projection matrices in $P(w)$. The MS(F)E of the estimators also depends on the signal-to-noise ratio because the variances of both the GLS and the OLS estimator of $\hat{\beta}$ depend on the error variances.

We specifically consider the presence of contemporaneous cross-unit error correlations, meaning that the variance of the OLS estimator is $(X'X)^{-1}X'\Omega X(X'X)^{-1}$ and the variance of the GLS estimator is $(X'\Omega^{-1}X)^{-1}$. In this case the GLS estimator is BLUE, following Zellner (1962), and it holds that

$$\mathrm{var}(\hat{\beta}_{\mathrm{OLS}}) - \mathrm{var}(\hat{\beta}_{\mathrm{GLS}}) \succcurlyeq 0.$$

Depending on the weights, the last expression in (2.4.3) shows that the GLS pooling averaging estimator potentially has a smaller MS(F)E than the OLS pooling averaging estimator due to the lower variance of $\hat{\beta}_{\mathrm{GLS}}$.

2.5 Mallows pooling averaging estimator

As proposed by Wang et al. (2019), we select the weight vector w of the pooling averaging estimators by minimizing the appropriate C_p-criterion by Mallows (1973), resulting in the Mallows pooling averaging (MPA) estimator. Wang et al. (2019) prove that, in the case of OLS MPA, the selected weights provide asymptotically optimal results in terms of the squared risk, as defined in (2.4.3). We must note that the optimality is not established in the presence of cross-unit error correlations and, to our knowledge, no prove exists for the GLS pooling averaging estimator and its variance estimators.

For the OLS pooling averaging estimator, weight selection by Mallows criterion is further motivated by Hansen (2007) and Hansen (2008) for averaging nested models with discrete weights. Wan et al. (2010) extend the optimality theory to non-nested models and continuous weights. Liu et al. (2016) prove the optimality for GLS averaging in a single equation model with heteroskedastic error terms, which, however, deviates from our assumptions about the error distribution.

In the case of a linear panel data regression model, the generalization by Wang et al. (2019) of the Mallows criterion proposed by Hansen (2007) is defined as

$$C_H(w) = ||P(w)\hat{\beta} - \hat{\beta}||_H^2 + 2\,\mathrm{tr}(P'(w)H\,\mathrm{var}(\hat{\beta})) - ||\hat{\beta} - \beta||_H^2, \tag{2.5.1}$$

which they show to be an unbiased estimator of the squared risk. The choice of the scaling matrix H depends on the application: if the focus lies on forecasting, $H = X'X$, and if the focus lies on the coefficient estimation, $H = I_{NK}$, the identity matrix. In our Monte Carlo simulation study, we focus on the forecasting performance of the pooling averaging estimator in the spirit of Wang et al. (2019) to make our results comparable.

Selection of the weights for GLS MPA or OLS MPA is achieved by minimizing the criterion with the appropriate projection matrices, $P_{\mathrm{OLS}}(w)$ or (the feasible) $P_{\mathrm{GLS}}(w)$, and the estimates $\hat{\beta}_{\mathrm{OLS}}$ and $\hat{\beta}_{\mathrm{GLS}}$ and their variances. That is,

$$\hat{w} = \arg\min_{w \in \mathcal{W}} C_H(w). \tag{2.5.2}$$

The last term of (2.5.1) is irrelevant for the minimization, because it is independent of the weights w. Minimizing this criterion ensures asymptotic optimality of the pooling averaging estimator under the assumptions and regularity conditions addressed by Wang et al. (2019), which we refer to for a thorough explanation and extensive proofs. However, as noted above, the optimality is not proven for the GLS pooling averaging estimator and further depends on the structure of the variance of the error terms Ω.

Because in practice the variance $\mathrm{var}(\hat{\beta})$ is unknown, we need to replace it with an estimate. Wang et al. (2019) propose three variance estimators which are all inconsistent in the presence of cross-unit error correlations. We explain our considered variance estimators in Section 2.6, but must note that optimality of the weight selection by (2.5.2) is not proven for our considered robust OLS variance estimator and both GLS variance estimators.

As noted in Section 2.2, we apply the C-Lasso proposed by Su et al. (2016) for our preliminary step of model screening. Following Wang et al. (2019), we select M pooling strategies by estimating the group memberships of each unit with the C-Lasso. That is, we select M panel structure models in which each unit belongs to a group in which all units share the same coefficients, including the intercept. Each pooling strategy m corresponds to a specific maximum number of possible groups equal to m, for $m = 1, \dots, M$ and $M \leq N$. As Wang et al. (2019) show, the optimality of the weight selection by Mallows criterion still holds when screening out the poor performing pooling strategies by the C-Lasso, provided that the other conditions and assumptions are met.

2.6 Variance estimators for OLS and feasible GLS

In this section, we explain the applicable variance estimators for both the OLS and the feasible GLS estimators of β, and what one should take into consideration when applying these with respect to consistency and the optimality of MPA. For notational convenience, we define the $N \times 1$ residual vector $\hat{\varepsilon}_t = [\hat{\varepsilon}_{1t}, \dots, \hat{\varepsilon}_{Nt}]'$ and the $N \times NK$ block diagonal matrix $X_t = \text{diag}(x_{1t}, \dots, x_{Nt})$ with $1 \times K$ vectors of explanatory variable, for $t = 1, \dots, T$.

For OLS, we consider the robust variance estimator

$$\hat{\text{var}}(\hat{\beta}_{\text{OLS}}) = (X'X)^{-1}(\sum_{t=1}^{T} X_t'\hat{\varepsilon}_t\hat{\varepsilon}_t'X_t)(X'X)^{-1}, \tag{2.6.1}$$

where X is defined as in (2.1.2). This variance estimator is consistent in the presence of cross-unit correlation and cross-unit heteroskedasticity of the error terms, following the distributional assumptions in (2.1.3). As an alternative, we consider an OLS variance estimator that is only consistent in the presence of cross-unit error heteroskedasticity. This estimator is given by

$$\hat{\text{var}}(\hat{\beta}_{\text{OLS}}) = \text{diag}(\hat{\sigma}_1^2(X_1'X_1)^{-1}, \dots, \hat{\sigma}_N^2(X_N'X_N)^{-1}). \tag{2.6.2}$$

In this expression, $\hat{\sigma}_i^2 = \hat{\sigma}_{ii} = \frac{1}{T}\sum_{t=1}^{T} \hat{\varepsilon}_i'\hat{\varepsilon}_i$, where $\hat{\varepsilon}_i$ is the $T \times 1$ residual vector and X_i is defined as in (2.1.1), for all i. The estimator in (2.6.2) is equivalent to the 'between-individual heteroscedasticity' consistent variance estimator considered by Wang et al. (2019).

For GLS, we first address the estimation of Σ to obtain the feasible GLS estimator. We consistently estimate the contemporaneous variance of the error terms by

$$\hat{\Sigma} = \frac{1}{T}\sum_{t=1}^{T} \hat{\varepsilon}_t\hat{\varepsilon}_t', \tag{2.6.3}$$

which has typical elements $\hat{\sigma}_{ij} = \frac{1}{T}\sum_{i=1}^{T} \hat{\varepsilon}_{it}\hat{\varepsilon}_{jt}$ for $i, j = 1, \dots, N$ and where $\hat{\varepsilon}_{it}$ are the residuals obtained by a preliminary OLS estimation step. Following this estimator, we construct $\hat{\Omega} = \hat{\Sigma} \otimes I_T$. As for OLS, we consider two estimators for the variance of the feasible GLS estimator. The robust variance estimator is

$$\hat{\text{var}}(\hat{\beta}_{\text{GLS}}) = (X'\hat{\Omega}^{-1}X)^{-1}(\sum_{t=1}^{T} X_t'\hat{\Sigma}^{-1}\hat{\varepsilon}_t^*\hat{\varepsilon}_t^{*\prime}\hat{\Sigma}^{-1}X_t)(X'\hat{\Omega}^{-1}X)^{-1}, \tag{2.6.4}$$

where $\hat{\varepsilon}_t^*$ is an $N \times 1$ vector with feasible GLS residuals and X defined as in (2.1.2). As an alternative, we consider the non-robust and statistically most efficient variance estimator for feasible GLS,

$$\hat{\text{var}}(\hat{\beta}_{\text{GLS}}) = (X'\hat{\Omega}X)^{-1}. \tag{2.6.5}$$

However, this estimator is only consistent when it holds that $\text{E}[X_t'\Sigma^{-1}\varepsilon_t\varepsilon_t'\Sigma^{-1}X_t] = \text{E}[X_t'\Sigma^{-1}X_t]$ for all t, which is not satisfied when the error terms within units i are heteroskedastic.

Under the regularity conditions discussed by Wang et al. (2019), weight selection following the Mallows criterion by (2.5.2) is optimal for uncorrelated error terms. Applying the cross-unit heteroskedastic variance estimator of OLS in (2.6.2) is proven to be optimal. However, the optimality is not established for the other variance estimators we discuss in this section. By means of our Monte Carlo simulation study we attempt to gain insights in how the different variance estimators perform for both OLS MPA and GLS MPA. The theoretical (potentially optimality) properties are beyond the scope of this research.

3 Monte Carlo simulation set-up

3.1 Data generating processes

In this section, we explain the design of the data generating processes (DGPs) for our simulation study. The bias-efficiency trade-off of the MPA estimators depends, among other things, on the the degree of cross-unit coefficient heterogeneity and the variance of the individual OLS or feasible GLS estimator, which is shown in the last expression of (2.4.3). The size of the cross-unit error correlations determines the relative efficiency gain by GLS compared to OLS. Further, in a finite sample, the performance of feasible GLS is particularly dependent on the sample size. Therefore, to compare the estimators, we consider multiple DGP combinations with respect to coefficient heterogeneity, signal-to-noise ratio, sample size, and cross-unit error correlations. We also consider DGPs in which the exogeneity condition is violated within units to compare the robustness with respect to MS(F)E of the estimators in this situation.

Following Wang et al. (2019), we consider the panel data model

$$y_{it} = \beta_{i1} + \beta_{i2}x_{2,it} + \beta_{i3}x_{3,it} + \varepsilon_{it}, \quad i = 1, \ldots, N; \quad t = 1, \ldots, T, \tag{3.1.1}$$

for which we separate four DGPs by their degree of coefficient heterogeneity. We define DGP 1 with completely homogeneous coefficients as

$$\beta_{ik} = 1 \quad \text{for all } i \text{ and } k = 1, 2, 3.$$

Further, letting $\lceil\cdot\rceil$ and $\lfloor\cdot\rfloor$ denote the ceiling and the floor function, respectively, we define DGP 2 with weakly heterogeneous coefficients as

$$\beta_{i1}, \beta_{i2} = \begin{cases} 1, \ i = 1, \ldots, \lfloor N/2 \rfloor \\ 3, \ i = \lceil N/2 \rceil, \ldots, N, \end{cases} \qquad \beta_{i3} = \begin{cases} 1, \ i = 1, \ldots, \lfloor N/3 \rfloor \\ 3, \ i = \lceil N/3 \rceil, \ldots, N; \end{cases}$$

DGP 3 with strongly heterogeneous coefficients as

$$\beta_{i1}, \beta_{i2} = \begin{cases} 1, \ i = 1, \ldots, \lfloor N/4 \rfloor \\ 2, \ i = \lceil N/4 \rceil, \ldots, \lfloor N/2 \rfloor \\ 3, \ i = \lceil N/2 \rceil, \ldots, \lfloor 3N/4 \rfloor \\ 4, \ i = \lceil 3N/4 \rceil, \ldots, N, \end{cases} \qquad \beta_{i3} = \begin{cases} 1, \ i = 1, \ldots, \lfloor N/5 \rfloor \\ 2, \ i = \lceil N/5 \rceil, \ldots, \lfloor 2N/5 \rfloor \\ 3, \ i = \lceil 2N/5 \rceil, \ldots, \lfloor 3N/5 \rfloor \\ 4, \ i = \lceil 3N/5 \rceil, \ldots, N; \end{cases}$$

and DGP 4 with completely heterogeneous coefficients as

$$\beta_{ik} = 0.1 \times i \times k \quad \text{for all } i \text{ and } k = 1, 2, 3.$$

We generate the explanatory variables $x_{k,it}$, for $k = 2, 3$ and for all i, t, independently from a standard normal distribution. Further, we generate the error terms ε_{it}, for all i, t, independently from a normal distribution with mean zero and, to obtain a predetermined signal-to-noise ratio R^2, unit dependent variances. Contemporaneous cross-unit error correlations are obtained by means of the Cholesky decomposition of a predetermined correlation matrix with typical off-diagonal elements ρ. The actual size of the cross-unit correlations in our DGPs can deviate from ρ, being lower between units with higher degrees of heterogeneity and for units that have larger coefficients. That is, we consider heterogeneous cross-unit error correlations by construction. It holds, though, that in a DGP constructed with 2ρ the cross-unit correlations are twice as high as in a DGP constructed with ρ. This results from the fact that we apply the Cholesky decomposition of the correlation matrix and not of the variance-covariance matrix, while we consider DGPs in which the variances across units can differ due to cross-unit heterogeneous coefficients. To obtain DGP samples with endogenous variables, we consider a similar approach as for the generation of cross-unit error correlations. For the derivation of the error variances and a detailed explanation of the generation of cross-unit error correlations and endogenous variables, we refer to Appendix A.1, A.2, and A.3, respectively.

We consider DGP combinations of $N = 10$ units, moderate and large sample sizes $T = 20, 80$, cross-unit error correlations constructed with $\rho = 0, 0.4, 0.8$, and high and moderate signal-to-noise ratios $R^2 = 0.9, 0.4$. For the DGPs with endogenous variables, we only consider $T = 80$ and apply the Cholesky decomposition of a correlation matrix which has a moderate correlation of 0.4 between the explanatory variables and errors.

3.2 Evaluation and estimation

To gain insight in the performance of GLS MPA relative to OLS MPA, we perform a Monte Carlo simulation study with 100 replications for the different DGP combinations explained in Section 3.1. For each DGP combination, we consider the GLS MPA, OLS MPA, (individual) OLS and feasible GLS, and pooling estimators. Following Wang et al. (2019), we evaluate the performance of each estimator in terms of their (sample) squared risk:

$$\hat{R}_H(\hat{\beta}) = \frac{1}{100}\sum_{j=1}^{100} ||\hat{\beta}_j - \beta||_H^2, \tag{3.2.1}$$

where $\hat{\beta}_j$ denotes a specific estimate in the DGP combination of replication j and β the true DGP coefficient vector. For the estimation by MPA and the evaluation of all estimators, we let $H = X'X$, focusing on forecast accuracy, to make our results comparable with those of Wang et al. (2019).

Because the consistency of the variance estimators for OLS and GLS depends on the structure of the variance of the error terms, we consider both the cross-unit error correlations robust and the non-robust estimators explained in Section 2.6 for the MPA estimators. For GLS MPA, the assumption of within-unit homoscedasticity for the consistency of the efficient variance estimator in (2.6.5) is always met. This is due to the design of our simulation DGPs. However, because the minimization of Mallows criterion in (2.5.1) may produce different results for different variance estimators, we consider both the robust and non-robust variance estimator.

Considering the model screening and group assignment by the C-Lasso, we follow Wang et al.

(2019) and Su et al. (2016) by applying the tuning parameter $\lambda = c_\lambda S_y^2 T^{-1/3}$, where S_y^2 is the sample variance of the dependent variable y in (2.1.2). Because the C-Lasso is computationally expensive, all C-Lasso group estimations are performed with $c_\lambda = 0.25$. This value provides the best results on average in terms of risk in preliminary simulations of 50 replication for a few DGP combinations over the grid $c_\lambda = 0.0625, 0125, 025, 0.5, 1$. We realize that this procedure may be a point of discussion, which we note in Section 5. Further, we let the maximum number of groups in the group selection estimation for each pooling strategy be equal to $m = 1, \ldots, N-1$ and therefore leave out the individual OLS or feasible GLS estimator.

The simulation study is implemented in software package MATLAB[1], including some adjusted versions of the original C-Lasso optimization code by Su et al. (2016). An overview of the programs and files can be found in Appendix A.4.

4 Results

4.1 Moderate sample size

In this subsection we consider the simulation results for a moderate number of $T = 20$ observations for each of the $N = 10$ units, which are presented in Table 1. Note that all numbers are divided by the squared risks of the (individual) OLS estimator, which is why the OLS estimator is not included in the table. Of the 24 different DGP combinations, GLS MPA only performs best twice: for high cross-unit error correlations in the strongly heterogeneous panels of DGP 3.

Considering the results for a high signal-to-noise ratio, $R^2 = 0.9$, GLS MPA outperforms OLS MPA in the strongly and completely heterogeneous panels of DGP 3 and 4 for high cross-unit error correlations, $\rho = 0.8$. This is, however, only by a small margin. As expected, the pooled estimator performs best in all the combinations of the homogeneous panel of DGP 1, but performs by far worst in the heterogeneous panels. The OLS MPA estimator performs best in the other situations. We further notice that the feasible GLS estimator performs better than the OLS estimator for high error correlations, particularly in the homogeneous panel. Interestingly, all estimators fail to outperform OLS in the completely heterogeneous panel of DGP 4 for high error correlations, resulting in squared risk ratios larger than one.

We find similar results in the case of a moderate signal-to-noise ratio, $R^2 = 0.4$. However, compared to the individual OLS estimator, GLS MPA and OLS MPA perform better than for $R^2 = 0.9$ in the strongly heterogeneous panels of DGP 3 and 4, and worse in the weaker heterogeneous panels. In contrast to the theoretical and simulation results of Wang et al. (2019), the pooling estimator does not compete with MPA in heterogeneous panels for a lower signal-to-noise ratio. However, we find a large decrease in its risk ratio relative to $R^2 = 0.9$. Further, we notice that the squared risk of GLS scales linearly with that of OLS for the different signal-to-noise ratios, such that the squared risk ratios are constant over $R^2 = 0.9, 0.4$. The same holds for the pooling estimator in the homogeneous panels of DGP 1. Interestingly, in the presence of error correlations, OLS MPA under the inconsistent non-robust variance estimator often performs better than under the consistent robust estimator, particularly for the lower signal-to-noise ratio.

[1] All programs are written in MATLAB R2019a.

Table 1: Squared risk comparison correctly specified model: GLS MPA, OLS MPA, feasible GLS, and pooling forecasts for the different DGPs described in Section 3.1, for $N = 10$ and $T = 20$ under different cross-unit error correlations constructed by ρ and signal-to-noise ratios R^2. All numbers are divided by the squared risk of the individual OLS estimators.

			GLS MPA		OLS MPA		GLS	Pool
		DGP	rb[a]	non-rb[b]	rb[c]	non-rb[d]		
$R^2 = 0.9$								
	$\rho = 0$	1	0.660	0.619	0.455	0.384	1.128	0.093*
		2	0.600	0.561	0.407	0.363*	1.133	14.510
		3	0.804	0.794	0.682*	0.683	1.123	12.956
		4	0.886	0.887	0.815*	0.841	1.124	13.547
	$\rho = 0.4$	1	0.580	0.550	0.511	0.423	0.908	0.205*
		2	0.586	0.552	0.435	0.382*	1.054	19.388
		3	0.796	0.792	0.723*	0.730	1.055	17.596
		4	0.982	0.988	0.929*	0.965	1.081	19.136
	$\rho = 0.8$	1	0.443	0.425	0.586	0.494	0.612	0.314*
		2	0.530	0.504	0.483	0.428*	0.865	39.480
		3	0.767	0.765*	0.772	0.776	0.908	39.352
		4	1.309	1.314	1.310	1.342	1.019	50.897
$R^2 = 0.4$								
	$\rho = 0$	1	0.668	0.628	0.469	0.398	1.128	0.093*
		2	0.648	0.618	0.489	0.463*	1.133	1.160
		3	0.655	0.630	0.531	0.505*	1.123	1.046
		4	0.640	0.615	0.516	0.496*	1.124	1.092
	$\rho = 0.4$	1	0.596	0.566	0.524	0.432	0.908	0.205*
		2	0.637	0.613	0.517	0.498*	1.054	1.557
		3	0.684	0.666	0.586	0.562*	1.055	1.423
		4	0.673	0.651	0.579	0.561*	1.081	1.528
	$\rho = 0.8$	1	0.452	0.432	0.591	0.498	0.612	0.314*
		2	0.571	0.555	0.567	0.551*	0.865	3.110
		3	0.698	0.690*	0.705	0.724	0.908	3.087
		4	0.728	0.721	0.697	0.696*	1.019	3.902

[a] Applying the robust variance estimator in (2.6.4).
[b] Applying the non-robust variance estimator in (2.6.5).
[c] Applying the robust variance estimator in (2.6.1).
[d] Applying the non-robust variance estimator in (2.6.2).
* Best forecast for a particular DGP with respect to risk.

A possible explanation for the fact that GLS MPA does not convincingly outperform OLS MPA is the sample size. As noted in Section 2.3, a sufficiently large sample size is required for an accurate estimate of the error variance Ω for the feasible GLS estimator. At the same time, in some cases where OLS MPA outperforms GLS MPA, feasible GLS outperforms OLS. This opposes our (intuitive) reasoning in Section 2.4, in which we compare the MS(F)E of the pooling averaging estimators and argue that the efficiency gained by GLS could lead to a better performing GLS pooling averaging estimator. However, in most of these occurrences, feasible GLS performs only slightly better than OLS or GLS MPA performs only slightly worse than

OLS MPA. More accurate results obtained by a larger number of simulation replications could perhaps answer if our reasoning about the efficiency gain in GLS MPA by GLS is false.

Our results for the OLS MPA and pooling estimators, in the absence of error correlations, deviate slightly from the results of Wang et al. (2019). Concerning OLS MPA, a possible explanation is that they consider two other variance estimators for the weight selection by Mallows criterion, while we consider their between-unit heteroscedasticity consistent (non-robust) and our cross-unit error correlation consistent (robust) variance estimators. Further, we determine the tuning parameter of the C-Lasso by a few screening simulations and not within each replication. With respect to both the OLS MPA and pooling estimators, a reason for the differences could be a deviation in generating the simulation samples, for instance, in our construction of a predetermined signal-to-noise ratio. Further, we base our simulation results on 100 replications, while Wang et al. (2019) consider 1,000 replications. Although our simulation results are generally stable, slight deviations may be accounted to this fact.

4.2 Large sample size

In this subsection we explain the simulation results for a large number of $T = 80$ observations for each of the $N = 10$ units. These are presented in Table 2. The GLS MPA estimator performs best in 9 of the 24 DGP combinations, which is, as expected, substantially more than for $T = 20$. GLS MPA performs particularly well for high error correlations, while OLS MPA often performs best in the absence of error correlations. Similar to the results for a moderate sample size in Section 4.1, we notice that in some situations OLS MPA outperforms GLS MPA, while feasible GLS outperforms OLS. In accordance with the results for $T = 20$, it holds that in most of these cases feasible GLS performs only slightly better than OLS or GLS MPA performs only slightly worse than OLS MPA. Further, the robust and non-robust variance estimators generally give similar results for both MPA estimators. Some explanations for deviations from the simulation results of Wang et al. (2019), in the absence of error correlations, are discussed in Section 4.1.

For a high signal-to-noise ratio, $R^2 = 0.9$, GLS MPA outperforms OLS MPA in the homogeneous and heterogeneous panels of DGP 1, 3, and 4 respectively, for moderate cross-unit error correlations, $\rho = 0.4$. Under high error correlations, GLS MPA outperforms all other considered estimators in the homogeneous panels of DGP 1 and the heterogeneous panels of DGP 2 and 3. The pooling estimator performs best in the homogeneous panels of DGP 1 for $\rho = 0, 0.4$, but, opposed to what we expect, slightly worse than GLS MPA for $\rho = 0.8$. In the completely heterogeneous panel of DGP 4, the feasible GLS estimator outperforms all others for high error correlations and is slightly worse than the OLS estimator for zero and moderate error correlations. In all other situations, particularly in the absence of error correlations, OLS MPA performs best.

Considering a moderate signal-to-noise ratio, $R^2 = 0.4$, OLS MPA outperforms GLS MPA in the absence of error correlations, as expected. However, there is no clear situation with respect to coefficient heterogeneity in which GLS MPA performs better than OLS MPA for moderate error correlations. GLS MPA outperforms OLS MPA in the homogeneous and strongly heterogeneous panels of DGP 1 and 3, while OLS MPA performs better in the weakly heterogeneous and completely heterogeneous panels of DGP 2 and 4. We must note that the differences are small.

Table 2: Squared risk comparison correctly specified model: GLS MPA, OLS MPA, feasible GLS, and pooling forecasts for the different DGPs described in Section 3.1, for $N = 10$ and $T = 80$ under different cross-unit error correlations constructed by ρ and signal-to-noise ratios R^2. All numbers are divided by the squared risk of the individual OLS estimators.

		GLS MPA		OLS MPA		GLS	Pool
	DGP	rb[a]	non-rb[b]	rb[c]	non-rb[d]		
$R^2 = 0.9$							
$\rho = 0$	1	0.451	0.443	0.401	0.387	1.058	0.096*
	2	0.442	0.436	0.381	0.369*	1.063	61.788
	3	0.740	0.739	0.695	0.692*	1.061	55.883
	4	1.349	1.354	1.295	1.305	1.062	58.580
$\rho = 0.4$	1	0.379	0.370	0.474	0.426	0.808	0.202*
	2	0.428	0.421	0.407	0.386*	0.977	82.048
	3	0.679	0.678*	0.732	0.728	0.986	75.496
	4	1.407	1.407	1.456	1.472	1.014	82.277
$\rho = 0.8$	1	0.312*	0.305	0.563	0.491	0.477	0.313
	2	0.384	0.377*	0.466	0.435	0.770	166.224
	3	0.583	0.582*	0.706	0.699	0.822	167.482
	4	2.343	2.346	2.223	2.241	0.949*	218.513
$R^2 = 0.4$							
$\rho = 0$	1	0.458	0.450	0.404	0.389	1.058	0.096*
	2	0.484	0.480	0.443	0.435*	1.063	4.665
	3	0.757	0.756	0.721*	0.723	1.061	4.230
	4	0.731	0.733	0.680*	0.681	1.062	4.431
$\rho = 0.4$	1	0.379	0.370	0.479	0.430	0.808	0.202*
	2	0.452	0.447	0.452	0.440*	0.977	6.209
	3	0.766*	0.766	0.775	0.779	0.986	5.720
	4	0.760	0.763	0.741*	0.745	1.014	6.211
$\rho = 0.8$	1	0.315	0.308*	0.565	0.493	0.477	0.313
	2	0.393	0.388*	0.475	0.447	0.770	12.516
	3	0.643*	0.644	0.793	0.815	0.822	12.593
	4	0.861*	0.864	0.917	0.933	0.949	16.329

[a,b,c,d,*] See notes to Table 1.

Under high error correlations, GLS MPA outperforms all other considered estimators. The differences between GLS MPA and OLS MPA are substantially larger than for moderate error correlations. Considering the performance of the pooling estimator for a moderate signal-to-noise ratio, we get similar results as for $T = 20$. That is, the pooling estimator does not compete with the other estimators in all heterogeneous panels.

Next, we discuss the simulation results in the presence of endogeneity, which are presented in Table 3. In general, we find that GLS MPA outperforms all estimators in the heterogeneous panels and the pooling estimator performs best in the homogeneous panels. All estimators, except the pooling estimator, often have results close to individual OLS. Interestingly, OLS MPA does not outperform GLS MPA in any situation, while we expect OLS MPA to be better in the absence of error correlations, which is the case when the exogeneity condition is satisfied.

Table 3: Squared risk comparison endogenous variables model: GLS MPA, OLS MPA, feasible GLS, and pooling forecasts for the different DGPs described in Section 3.1, for $N = 10$ and $T = 80$ under different cross-unit error correlations constructed by ρ, and signal-to-noise ratios R^2. All numbers are divided by the squared risk of the individual OLS estimators.

		GLS MPA		OLS MPA		GLS	Pool
	DGP	rb[a]	non-rb[b]	rb[c]	non-rb[d]		
$R^2 = 0.9$							
$\rho = 0$	1	0.991	0.990	0.990	0.990	1.001	0.985[*]
	2	0.940	0.940[*]	0.944	0.943	1.005	8.968
	3	0.953[*]	0.953	0.959	0.959	1.007	11.294
	4	0.968	0.967[*]	1.022	1.022	1.002	4.427
$\rho = 0.4$	1	0.990	0.990	0.991	0.990	0.997	0.987[*]
	2	0.950	0.950[*]	0.959	0.958	0.996	9.182
	3	0.953[*]	0.953	0.969	0.969	0.996	11.696
	4	0.982	0.981[*]	1.027	1.028	0.998	4.479
$\rho = 0.8$	1	0.990	0.990	0.993	0.992	0.993	0.989[*]
	2	0.975	0.974[*]	0.981	0.980	0.992	9.523
	3	0.975	0.975[*]	0.994	0.994	0.992	12.370
	4	1.001	1.001	1.027	1.027	0.998[*]	4.562
$R^2 = 0.4$							
$\rho = 0$	1	0.896	0.895	0.887	0.884	1.011	0.829[*]
	2	0.666	0.663[*]	0.685	0.681	1.035	4.299
	3	0.766	0.765[*]	0.825	0.826	1.039	4.425
	4	0.766	0.764[*]	0.907	0.908	1.020	2.976
$\rho = 0.4$	1	0.885	0.883	0.903	0.894	0.965	0.851[*]
	2	0.683	0.679[*]	0.719	0.712	0.983	5.018
	3	0.789	0.788[*]	0.865	0.870	0.987	5.359
	4	0.810	0.807[*]	0.942	0.944	0.998	3.320
$\rho = 0.8$	1	0.877	0.875	0.920	0.906	0.908	0.873[*]
	2	0.782	0.779[*]	0.827	0.817	0.924	6.713
	3	0.828	0.827[*]	0.925	0.929	0.929	7.934
	4	0.917	0.916[*]	1.002	1.005	0.990	4.026

[a,b,c,d,*] See notes to Table 1.

However, the two MPA estimators generally outperform individual OLS. This indicates that MPA makes a better trade-off between bias and efficiency, also in the case when the individual OLS and feasible GLS estimators are biased due to endogeneity.

5 Conclusion

In this research, we propose an extension of the OLS MPA estimator proposed by Wang et al. (2019) to (feasible) GLS and compare the performance of this GLS MPA estimator to that of OLS MPA. We ask ourselves whether and in which situations GLS MPA performs better than OLS MPA in terms of MS(F)E due to the gain in efficiency by GLS in the presence of contemporaneous cross-unit error correlations. Because in this case the OLS variance estimators considered by Wang et al. (2019) are inconsistent, we also examine the performance of OLS MPA when applying a variance estimator which is consistent in case of cross-unit error correlations. We perform an extensive Monte Carlo simulation study in which we compare the squared risk of the GLS MPA, OLS MPA, (individual) OLS and feasible GLS, and pooling estimators for different DGP combinations with respect to coefficient heterogeneity, sample size, cross-unit error correlations, and signal-to-noise ratio. To compare the robustness with respect to the MS(F)E of the estimators, we also consider DGP combinations in which the exogeneity condition is violated.

Based on our simulation results, we conclude that GLS MPA convincingly outperforms OLS MPA in panels with a large number of observations per unit and high cross-unit error correlations. In particular, for a moderate signal-to-noise ratio, GLS MPA outperforms all considered estimators. The results for moderate error correlations do not lead to a clear conclusion, as the performances of OLS MPA and GLS MPA alternate over the degrees of heterogeneity in the considered panels. For a moderate number of observations, OLS MPA often performs best. We find that none of the considered estimators performs best in all situations, which is in accordance with the findings of Wang et al. (2019). The pooling estimator performs best in most homogeneous panels and (individual) OLS or feasible GLS outperform MPA in completely heterogeneous panels in many situations. Further, the differences between the considered variance estimators for OLS MPA and GLS MPA are often small, which suggests that the importance of their consistency is negligible. At last, in case of endogeneity and a large number of observations, GLS MPA outperforms all estimators in almost all heterogeneous panels, though often by a small margin. In general, MPA outperforms OLS under endogeneity, proving to be more robust with respect to MS(F)E in this case.

As a point of discussion, we must note that the simulation results are based on 100 replications because the C-Lasso is computationally very expensive with respect to time. Even though the results are stable, more accurate results may be obtained by more replications. Further, the tuning parameter of the C-Lasso used for group assignment to shrink the model space is determined by a few screening simulations because of the same argument of computational expenses. Although the results over the grid of tuning parameters proved to be quite similar, one could argue whether a different, more extensive screening would lead to different conclusions.

As a suggestion for future research, it is interesting to find out whether or not there exists a theoretical asymptotic optimality result for the GLS MPA estimator—and if so, under which conditions this optimality would hold true. We think that our simulation study provides some insight in these situations. One could further extend the simulation experiment by considering more versatile data generating processes.

References

Baltagi, B. H., Feng, Q., and Kao, C. (2012). A Lagrange Multiplier test for cross-sectional dependence in a fixed effects panel data model. *Journal of Econometrics*, 170(1):164–177.

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

Hansen, B. E. (2008). Least-squares forecast averaging. *Journal of Econometrics*, 146(2):342–350.

Liu, Q., Okui, R., and Yoshimura, A. (2016). Generalized least squares model averaging. *Econometric Reviews*, 35(8-10):1692–1752.

Mallows, C. L. (1973). Some comments on Cp. *Technometrics*, 15(4):661–675.

Pesaran, M. H. and Smith, R. (1995). Estimating long-run relationships from dynamic heterogeneous panels. *Journal of econometrics*, 68(1):79–113.

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

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

Wan, A. T., Zhang, X., and Zou, G. (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics*, 156(2):277–283.

Wang, W., Zhang, X., and Paap, R. (2019). To pool or not to pool: What is a good strategy for parameter estimation and forecasting in panel regressions? *Journal of Applied Econometrics*, pages 1–22. https://doi.org/10.1002/jae.2696.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American statistical Association*, 57(298):348–368.

A Appendix

A.1 Predetermined DGP sample coefficient of determination

We show how to we obtain a predetermined coefficient of determination for the DGP samples of the simulation study. For each unit i it holds that

$$\begin{aligned} R_i^2 &= 1 - \frac{SS_{\text{res},i}}{SS_{\text{tot},i}} \\ &\xrightarrow{a} 1 - \frac{\text{var}(\varepsilon_{it})}{\text{var}(y_{it})}, \end{aligned}$$

where $SS_{\text{res},i}$ and $SS_{\text{tot},i}$ denote the residual sum of squares and the total sum of squares, respectively. Then, following the panel data model in (3.1.1), we can derive that

$$\begin{aligned} \text{var}(y_{it}) &= \text{var}(\beta_{i1} + \beta_{i2}x_{2,it} + \beta_{i3}x_{3,it} + \varepsilon_{it}) \\ &= \beta_{i2}^2 + \beta_{i3}^2 + \sigma_i^2, \end{aligned}$$

because $x_{k,it} \sim \text{i.i.d.}\,\text{N}(0,1)$, β_{ik} is fixed, $\text{E}[\varepsilon_{it}] = 0$, $\text{var}(\varepsilon_{it}) = \sigma_i^2$, and $\text{E}[x_{k,it}\varepsilon_{it}] = 0$ due to exogeneity. Then, it holds that

$$\begin{aligned} R_i^2 &= 1 - \frac{\sigma_i^2}{\text{var}(y_{it})} \\ &= 1 - \frac{\sigma_i^2}{\beta_{i2}^2 + \beta_{i3}^2 + \sigma_i^2}. \end{aligned}$$

For a simulation DGP sample with a predetermined R_i^2 we get the expression

$$\sigma_i^2 = \frac{(1 - R_i^2)}{R_i^2}(\beta_{i2}^2 + \beta_{i3}^2), \tag{A.1.1}$$

which we apply to generate the error terms for each unit such that

$$\varepsilon_{it} \sim \mathrm{N}(0, \sigma_i^2).$$

A.2 Generating cross-unit correlated error terms

We explain how we obtain cross-unit correlated error terms in our DGP samples of the simulation study. We define the $N \times N$ contemporaneous correlation matrix R, having typical off-diagonal elements ρ, and the $T \times N$ matrix with generated uncorrelated error terms $U = [u_1, \ldots, u_N]$. Then we construct contemporaneously correlated error terms by

$$E = [\varepsilon_1, \ldots, \varepsilon_N] = UL$$

where L is the upper triangular Cholesky factorization of R.

If the variances of all units are identical, which is only the case for the homogeneous panel model of DGP 1, the cross-unit correlation of the constructed error terms is identical for all units and equal to ρ. In other cases, ρ denotes the correlation of the first group of units in the particular DGP, which has the smallest coefficients. That is, the actual size of the cross-unit correlation is lower between units with higher degrees of heterogeneity and for units that have larger coefficients. The latter is due to the higher variance, in absolute terms, of their error terms, as we show in (A.1.1). However, as noted in Section 3.1, it holds that in a sample constructed with 2ρ the cross-unit correlations are twice as high as one constructed with ρ. This covariance structure originates from the fact that we use the Cholesky decomposition of the correlation matrix and not of the variance-covariance matrix, while we consider DGPs in which the variances across units may differ, resulting in heterogeneous cross-unit error correlations.

A.3 Generating endogenous variables

We explain how we obtain endogenous variables in our DGP samples for the simulation study. That is, having $\mathrm{E}[x_{kit}\varepsilon_{it}] \neq 0$ for $k = 2, 3$ and for all i, t. Defining the correlation matrix

$$R_{X\varepsilon} = \begin{pmatrix} 1 & 0 & \rho_{X\varepsilon} \\ 0 & 1 & \rho_{X\varepsilon} \\ \rho_{X\varepsilon} & \rho_{X\varepsilon} & 1 \end{pmatrix},$$

we correlate the explanatory variables with the error terms within each unit i for all t by

$$[x_{2,it}^*, x_{3,it}^*, \varepsilon_{it}^*] = [x_{2,it}, x_{3,it}, \varepsilon_{it}]J,$$

where $x_{2,it}^*$, $x_{3,it}^*$, and ε_{it}^* are the correlated DGP variables and J is the upper triangular Cholesky factorization of $R_{X\varepsilon}$. As noted in Section 3.1, we only consider $\rho_{X\varepsilon} = 0.4$.

A.4 MATLAB programs

We provide an overview of the programs implemented for our Monte Carlo simulation study in MATLAB (R2019a).[2] Further, we must note that the functions for the C-Lasso of Su et al. (2016) require the additional optimization software CVX in MATLAB.[3] The programs and files are ordered in multiple folders and explained below.

C-Lasso:

- *C_Lasso_groups.m*: Function for C-Lasso estimation of the group structure for different numbers of groups.
- *CLasso_tuningParam_Check.m*: Script for selecting a proper tuning parameter for the C-Lasso using *simulate.m*.
- *criterion.m*: Function for convergence criterion of the estimation algorithm for *PLS.m* (original by Su et al. (2016)).
- *get_groups.m*: Function which determines the groups where each unit should be assigned to following *PLS.m* (original by Su et al. (2016)).
- *PLS.m*: Function for penalized least squares of the C-Lasso to determine the group coefficients for a given number of groups (original by Su et al. (2016), requires CVX software).

MPA:

- *constr_P.m*: Function that constructs the restriction projection matrix of a specific pooling strategy (see (2.2.2) and (2.2.2)).
- *min_MPA_Cp.m*: Function that minimizes the Mallows criterion (see (2.5.1) and (2.5.2)).
- *MPA_Cp.m*: Function for the Mallows criterion for a linear panel data regression model (see (2.5.1)).
- *MPA.m*: Function for GLS MPA or OLS MPA estimation in a linear panel data regression model, including model shrinkage by the C-Lasso.

Simulation:

- *betas_dgp.mat*: File with DGP coefficients for DGP 1, 2, 3, and 4 (constructed by *constr_beta.m*).
- *constr_beta.m*: Function that constructs the DGP beta coefficient vectors for DGP 1, 2, 3, and 4.
- *genr_endogenous_variables.m*: Generates endogenous variables for a specific DGP, providing the X and y data, and the errors.
- *genr_errors.m*: Generates the error terms with specific variances, cross-unit correlations, based on beta coefficients of the DGP (used in *genr_variables.m* and *genr_endogenous_variables.m*).
- *genr_variables.m*: Generates (exogenous) variables for a specific DGP, providing the X and y data, and the errors.
- *s.mat*: File with random generator.

[2] The MATLAB programs, including random generators and original output, are provided on request.

[3] We use the standard bundle downloaded at http://cvxr.com/cvx/download/.

- *simulate_ endogenous.m*: Function for a simulation of specific DGP for endogenous variables, in which the linear panel data regression model is estimated by all estimators and their squared risks are computed.
- *simulate.m*: Function for a simulation of specific DGP for correctly specified model, in which the linear panel data regression model is estimated by all estimators and their squared risks are computed.
- *Simulation_ endogenous.m*: Script for the simulation for endogeneity, which can be regarded as an example.
- *Simulation.m*: Script for the simulation for correctly specified model, which can be regarded as an example.

Other:

- *FGLS.m*: Function for feasible GLS estimation in a linear panel data regression model.
- *L_ A.m*: Function for computing the squared loss of an estimate or forecast.
- *Pool.m*: Function for pooled estimation in a linear panel data regression model.
- *SUR.m*: Function for system OLS estimation in a linear panel data regression model.

Results:

- Folder with *.mat* files of simulation results for $T = 20$ and $T = 80$, for the different DGPs 1, 2, 3, and 4, and the R^2 and ρ combinations, including the endogeneity results for $T = 80$.