

Improved Average Estimation in Seemingly Unrelated Regressions[†]

Ali Mehrabani[‡] and Aman Ullah[§]

June 24, 2020

Abstract

In this paper, we propose an efficient weighted average estimator in Seemingly Unrelated Regressions. This average estimator shrinks a generalized least squares (GLS) estimator towards a restricted GLS estimator, where the restrictions represent possible parameter homogeneity specifications. The shrinkage weight is inversely proportional to a weighted quadratic loss function. The approximate bias and second moment matrix of the average estimator using the large-sample approximations are provided. We give the conditions under which the average estimator dominates the GLS estimator on the basis of their mean squared errors. We illustrate our estimator by applying it to a cost system for U.S. Commercial banks, over the period from 2000 to 2018. Our results indicate that on average most of the banks have been operating under increasing returns to scale. We find that over the recent years, scale economies are a plausible reason for the growth in average size of banks and the tendency toward increasing scale is likely to continue.

Key Words: Stein-type Shrinkage Estimator; Asymptotic Approximations; SUR; GLS

JEL Classification: C10, C12, C13, C30, C52.

1 Introduction

Seemingly unrelated regressions (SUR) was introduced by Zellner, 1962 and is one of the econometric developments that has been widely used in applied work. The relative ease of estimation, applying a large class of modeling and testing problems, and the availability of data representing a sample of cross section units observed over several time periods are related to the popularity of this model. Zellner, 1962 proposed a generalized least squares (GLS) estimator for estimating the coefficients of a set of SUR and established that it yields, at least asymptotically, to more efficient estimators than those obtained by single-equation least squares. See the surveys by Srivastava and Dwivedi, 1979; Fiebig, 2001 and the book by Srivastava and Giles, 1987 for a concise coverage of the literature in this area.

[†]The authors are thankful to the guest-editors and two referees for helpful and constructive comments on the subject matter of this paper.

[‡]Department of Economics, University of California, Riverside, CA 92521. E-mail: ali.mehrabani@email.ucr.edu

[§]Department of Economics, University of California, Riverside, CA 92521. E-mail: aman.ullah@ucr.edu

Shrinkage estimations in SUR was first introduced by Zellner and Vandaale, 1975, which extends the results of James and Stein, 1961 and Sclove, 1968 to multivariate regression equations and presents a technique of constructing an estimator whose risk is smaller than the risk of the GLS estimator. However, the resulting estimator depends on some unknown matrices and is not practical. Srivastava, 1973 investigates the properties of the estimator when consistent estimators are substituted for these unknown matrices. Maddala, 1991 reviewed the shrinkage estimators and showed that these estimators appear to perform better than both pooled and single-equation least squares estimators (see also, Maddala and Hu, 1996; Maddala et al., 1997; Choi and Li, 2000). Maddala et al., 2001 show the superior properties of shrinkage estimators among single-equation estimators and various averaging estimators in a heterogeneous panel data model under error homoscedasticity framework. In univariate equation models, recently, Hansen, 2016 introduces shrinkage for general parametric models by shrinking maximum likelihood estimators (MLE) toward a restricted MLE. Hansen, 2016 shows the dominance of the shrinkage estimator over the MLE in terms of having lower asymptotic risk when the shrinkage dimension exceeds two, using a local to zero asymptotic framework. Wang et al., 2019 propose a Mallows pooling averaging estimator for heterogeneous panel data models and conclude that the pooling averaging estimator is preferred when the panel is heterogenous and the signal-to-noise ratio is moderate or large. For more on averaging estimators see Ullah and Wang, 2013.

In the analysis of SUR, a question that often practitioners are faced is whether to assume parameter homogeneity or parameter heterogeneity. On one hand, the parameter heterogeneity assumption results in consistent estimators and violation of this assumption causes misleading estimates (see, for example, Robertson and Symons, 1992; Pesaran and Smith, 1995; Su and Chen, 2013; Durlauf et al., 2001; Browning and Carro, 2007). On the other hand, the parameter homogeneity assumption causes higher efficiency but could be at the cost of estimation bias and inconsistency of estimators, which is supported by an increasing number of studies due to a better forecast performance of the estimators under this assumption (see, for example, Maddala, 1991; Maddala and Hu, 1996; Baltagi and Griffin, 1984; Baltagi et al., 2000; Hoogstrate et al., 2000). This question and the results of the mentioned research show the typical bias-variance trade-off that needs to be considered in choosing the restrictions. While efficiency is important, robustness is also critical, since researchers prefer as few ad hoc restrictions as possible. In the present scenario the efficient estimator depends on more stringent condition of homogeneity and therefore is less robust to the heterogeneity restriction. Therefore, this efficiency-robustness trade-off (bias-variance trade-off) calls for thorough examination. A natural approach is to consider a pre-test estimator, but it is proven unable to solve the efficiency-robustness issue (Leeb and Pötsche, 2005).

A more useful approach considered here is an averaging estimator (Stein-type shrinkage), that is a weighted average of the robust and efficient (the unrestricted GLS and the restricted GLS) estimators. The weight is inversely related to a weighted quadratic loss function, which measures the weighted distance between the unrestricted and the restricted GLS estimators. The first and second moments of our proposed average estimator are derived using Nagar, 1959 large-sample approximations. Furthermore, we show the dominance properties in terms of mean squared error

(MSE) of the estimators, which ensures that the proposed estimator is robust against arbitrary deviations from the restrictions. This is an advantage of our method relative to the “local asymptotic” argument that some previous studies rely on (see, for example, Hansen, 2016). Our dominance property ensures that the proposed averaging estimator is robust against arbitrary deviations from the restrictions, while previous estimators in the literature consider mainly very small violations of the restrictions. Further discussion in this area has been generally theoretical, but we present here, an important empirical question and show the advantages of the averaging estimator over the previous estimators considered. Essentially, we apply our estimator to estimate cost efficiency of U.S. Commercial banks using a cost system method over the period from 2000 to 2018. Since bank size is an important factor of production environment, following the literature, we use it to partition bank technologies. However, these partitions are user-specified, and the estimates can be misleading because of false parameter heterogeneity assumptions. Therefore, we use the average estimator introduced in the paper to estimate the cost efficiency, as it optimally balances the trade-off between bias and variance efficiency of the restricted and the unrestricted GLS estimators. We find that on average majority of banks have been operating under increasing returns to scale over the sample period. We also find more signs of cost efficiency for Large Banks (banks with asset size more than \$500 Million dollars) and Small Banks (banks with asset size less than \$100 Million dollars) relative to Medium banks (banks with asset size between \$500 and \$100 Million dollars). This finding is important for gauging costs and benefits of any policy intervention to control the size of banks.

The paper is organized as follows. Section 2 describes the model and the assumptions. In Section 3, we introduce the estimators. We give the bias, MSE matrix and the risk of the average estimator using the large-sample approximations in Section 4. Section 5 reports some Monte Carlo simulations to evaluate the accuracy of the approximations. Results from our empirical example are presented in Section 6. Finally, Section 7 contains some concluding remarks, and proofs are given in Appendix A.

2 The Model and Notation

Consider the following m seemingly unrelated linear regressions

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_i + \mathbf{u}_i, \quad i = 1, 2, \dots, m, \quad (2.1)$$

where $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{iT})'$ is a $T \times 1$ vector of observations on the dependent variable y_{it} , with T being the number of observations, \mathbf{X}_i is a $T \times k$ matrix of observations on the k vector of regressors including the intercept (that is, $x_{it,1} = 1$)¹, $\boldsymbol{\beta}_i$ is a $k \times 1$ vector of unknown coefficients and $\mathbf{u}_i = (u_{i1}, u_{i2}, \dots, u_{iT})'$ is a $T \times 1$ vector of disturbances, for $i = 1, 2, \dots, m$. It is convenient to stack the m equations above in the following form

¹Note that we do not assume that X_i 's are the same, nor do we assume they are different across equations. In other words, our model supports complete heterogeneity, partial heterogeneity, and complete homogeneity of regressors.

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_m \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & & \\ \vdots & \ddots & & \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{X}_m \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} + \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_m \end{bmatrix}, \quad (2.2)$$

or compactly as

$$\underset{mT \times 1}{\mathbf{y}} = \underset{mT \times mk}{\mathbf{X}} \underset{mk \times 1}{\boldsymbol{\beta}} + \underset{mT \times 1}{\mathbf{u}}. \quad (2.3)$$

We assume,

Assumption 1: The $mT \times 1$ vector of disturbances, \mathbf{u} , has a zero conditional mean

$$\mathbb{E}(\mathbf{u} | \mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_m) = \mathbf{0}.$$

Assumption 2: The disturbances are uncorrelated across observations but correlated across equations,

$$\mathbb{E}(\mathbf{u}_i \mathbf{u}_j' | \mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_m) = \sigma_{ij} \mathbf{I}_T,$$

or

$$\mathbb{E}(\mathbf{u} \mathbf{u}' | \mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_m) = \boldsymbol{\Omega} = \begin{bmatrix} \sigma_{11} \mathbf{I}_T & \sigma_{12} \mathbf{I}_T & \dots & \sigma_{1m} \mathbf{I}_T \\ \sigma_{21} \mathbf{I}_T & \sigma_{22} \mathbf{I}_T & \dots & \sigma_{2m} \mathbf{I}_T \\ & & \ddots & \\ \sigma_{m1} \mathbf{I}_T & \sigma_{m2} \mathbf{I}_T & \dots & \sigma_{mm} \mathbf{I}_T \end{bmatrix} = \boldsymbol{\Sigma} \otimes \mathbf{I}_T,$$

where \mathbf{I}_T is the $T \times T$ identity matrix,

$$\boldsymbol{\Sigma}_{m \times m} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1m} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2m} \\ & & \ddots & \\ \sigma_{m1} & \sigma_{m2} & \dots & \sigma_{mm} \end{bmatrix},$$

and we assume $\boldsymbol{\Omega}$ is positive definite.

Assumption 3: The disturbances are normally distributed with mean zero and variance-covariance matrix $\boldsymbol{\Omega}$.

We define some notations below, which will be used in the following sections. So let

$$\underset{mT \times mT}{\mathbf{Q}} = \boldsymbol{\Omega}^{-1} (\mathbf{I}_{mT} - \boldsymbol{\Psi}), \quad (2.4)$$

where

$$\mathbf{\Psi}_{mT \times mT} = \mathbf{X}(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}. \quad (2.5)$$

If we partition \mathbf{Q} in the sub-matrices of $T \times T$ as below

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} & \cdots & \mathbf{Q}_{1m} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} & \cdots & \mathbf{Q}_{2m} \\ \vdots & \vdots & & \vdots \\ \mathbf{Q}_{m1} & \mathbf{Q}_{m2} & \cdots & \mathbf{Q}_{mm} \end{bmatrix}, \quad (2.6)$$

then we define

$$\mathbf{\Pi} = \begin{bmatrix} \mathbf{Q}'_{11} & \mathbf{Q}'_{12} & \cdots & \mathbf{Q}'_{1m} \\ \mathbf{Q}'_{21} & \mathbf{Q}'_{22} & \cdots & \mathbf{Q}'_{2m} \\ \vdots & \vdots & & \vdots \\ \mathbf{Q}'_{m1} & \mathbf{Q}'_{m2} & \cdots & \mathbf{Q}'_{mm} \end{bmatrix}. \quad (2.7)$$

Also, we define

$$\mathbf{\Phi}_{T \times T} = \sum_{i=1}^m \mathbf{\Psi}_{ii}, \quad (2.8)$$

where $\mathbf{\Psi}_{ii}$ is the i th diagonal $T \times T$ sub-matrix of $\mathbf{\Psi}$ which is partitioned as below

$$\mathbf{\Psi} = \begin{bmatrix} \mathbf{\Psi}_{11} & \mathbf{\Psi}_{12} & \cdots & \mathbf{\Psi}_{1m} \\ \mathbf{\Psi}_{21} & \mathbf{\Psi}_{22} & \cdots & \mathbf{\Psi}_{2m} \\ \vdots & \vdots & & \vdots \\ \mathbf{\Psi}_{m1} & \mathbf{\Psi}_{m2} & \cdots & \mathbf{\Psi}_{mm} \end{bmatrix}. \quad (2.9)$$

3 Estimators

Our goal is to estimate the vector of slope parameters, $\boldsymbol{\beta}$, in equation (2.3). We consider three estimators of the slope parameters. The first estimator is the Zellner, 1962 GLS estimator (the unrestricted GLS estimator), which is the standard estimator in SUR. The second estimator is a restricted GLS estimator that ignores the slope parameters heterogeneity and estimates a pooled model. The third estimator, called the average estimator, is a weighted average of the restricted and the unrestricted GLS estimators where the weight is proportional to a weighted quadratic loss function.

3.1 Unrestricted Estimator

The typical estimator of the slope parameters in SUR is a feasible GLS estimator defined as

$$\hat{\beta} = (\mathbf{X}'\hat{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\Omega}^{-1}\mathbf{y} = \beta + (\mathbf{X}'\hat{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\Omega}^{-1}\mathbf{u}, \quad (3.1)$$

where $\hat{\Omega}$ is an estimator of Ω , which can be calculated as

$$\hat{\Omega} = \hat{\Sigma} \otimes \mathbf{I}_T, \quad (3.2)$$

where $\hat{\Sigma}$ is an estimator of Σ , such that its (i, j) th element, s_{ij} , estimates σ_{ij} , using a single-equation estimator of β_i , defined as $\check{\beta}_i = (\mathbf{X}'_i\mathbf{X}_i)^{-1}\mathbf{X}'_i\mathbf{y}_i$, $i = 1, 2, \dots, m$. Hence, s_{ij} is equal to

$$s_{ij} = (\mathbf{y}_i - \mathbf{X}_i\check{\beta}_i)'(\mathbf{y}_j - \mathbf{X}_j\check{\beta}_j)/T = \mathbf{u}'_i\mathbf{M}_i\mathbf{M}_j\mathbf{u}_j/T, \quad (3.3)$$

where $\mathbf{M}_i = \mathbf{I}_T - \mathbf{X}_i(\mathbf{X}'_i\mathbf{X}_i)^{-1}\mathbf{X}'_i$ is an idempotent projection matrix.

3.2 Restricted Estimator

The restricted estimator is defined under the parameter homogeneity assumption across equations, which can be written as

$$\beta_1 = \beta_2 = \dots = \beta_m = \bar{\beta},$$

where $\bar{\beta}$ is a weighted average of the slope parameters, β_i 's, defined as

$$\bar{\beta} = (\mathbf{J}'\mathbf{X}'\Omega^{-1}\mathbf{X}\mathbf{J})^{-1}\mathbf{J}'\mathbf{X}'\Omega^{-1}\mathbf{X}\beta, \quad (3.4)$$

in which $\mathbf{J}' = (\mathbf{I}_k, \mathbf{I}_k, \dots, \mathbf{I}_k)$ is a $k \times mk$ matrix, where \mathbf{I}_k denotes the $k \times k$ identity matrix.

Equivalently, the parameter homogeneity assumption can be formulated as a restriction matrix as

$$\begin{bmatrix} \beta_1 - \bar{\beta} \\ \beta_2 - \bar{\beta} \\ \vdots \\ \beta_m - \bar{\beta} \end{bmatrix} = \beta - \mathbf{J}\bar{\beta} = \left(\mathbf{I}_{mk} - \mathbf{J}(\mathbf{J}'\mathbf{X}'\Omega^{-1}\mathbf{X}\mathbf{J})^{-1}\mathbf{J}'\mathbf{X}'\Omega^{-1}\mathbf{X} \right) \beta = \mathbf{R}\beta = \mathbf{0}, \quad (3.5)$$

where $\mathbf{R} = \mathbf{I}_{mk} - \mathbf{J}(\mathbf{J}'\mathbf{X}'\Omega^{-1}\mathbf{X}\mathbf{J})^{-1}\mathbf{J}'\mathbf{X}'\Omega^{-1}\mathbf{X}$ is an idempotent matrix.

Hence, we can derive the restricted estimator from the following minimization

$$\text{Minimize}_{\text{s.t. } \beta} (\mathbf{y} - \mathbf{X}\beta)' \Omega^{-1} (\mathbf{y} - \mathbf{X}\beta), \quad \text{subject to } \mathbf{R}\beta = \mathbf{0}.$$

The solution to the above minimization can be formulated as a feasible restricted GLS estimator

in below

$$\begin{aligned}\tilde{\beta} &= \hat{\beta} - (\mathbf{X}'\hat{\Omega}^{-1}\mathbf{X})^{-1}\hat{\mathbf{R}}' \left[\hat{\mathbf{R}}(\mathbf{X}'\hat{\Omega}^{-1}\mathbf{X})^{-1}\hat{\mathbf{R}}' \right]^{-1} \hat{\mathbf{R}}\hat{\beta} = (\mathbf{I}_{mk} - \hat{\mathbf{R}})\hat{\beta} \\ &= \mathbf{J}(\mathbf{J}'\mathbf{X}'\hat{\Omega}^{-1}\mathbf{X}\mathbf{J})^{-1}\mathbf{J}'\mathbf{X}'\hat{\Omega}^{-1}\mathbf{X}\hat{\beta},\end{aligned}\tag{3.6}$$

where $\hat{\mathbf{R}} = \mathbf{I}_{mk} - \mathbf{J}(\mathbf{J}'\mathbf{X}'\hat{\Omega}^{-1}\mathbf{X}\mathbf{J})^{-1}\mathbf{J}'\mathbf{X}'\hat{\Omega}^{-1}\mathbf{X}$ is an estimate of \mathbf{R} .

3.3 Average Estimator

We define the average estimator as below

$$\hat{\beta}_A = \left(1 - \frac{\tau}{\mathbf{D}}\right)\hat{\beta} + \frac{\tau}{\mathbf{D}}\tilde{\beta},\tag{3.7}$$

where \mathbf{D} is a weighted quadratic loss function defined as

$$\mathbf{D} = (\hat{\beta} - \tilde{\beta})'\mathbf{W}(\hat{\beta} - \tilde{\beta}),\tag{3.8}$$

with \mathbf{W} an arbitrary symmetric positive definite weight matrix with elements of order $O(T)$, and τ is a positive characterizing parameter. We will defer describing the optimal choice for this parameter in the next section.²

The idea behind the average estimator defined above is that when the difference between the restricted and the unrestricted GLS estimators is small (\mathbf{D} is small), the average estimator gives a higher weight to the restricted GLS estimator, as it is the most efficient estimator. However, when the difference between the restricted and the unrestricted GLS estimators is substantial, the bias of the restricted GLS estimator, resulting from ignoring the parameter heterogeneity, could be more than its variance efficiency gain, so the average estimator assigns a higher weight to the unrestricted GLS estimator.

4 Large-Sample Approximate Bias and MSE

We employ the large-sample approximations method developed by Nagar, 1959, to analyze the bias, mean squared error matrix (MSEM) and risk of the average estimator.

Theorem 1: *Under Assumptions 1–3, the bias of the average estimator up to order $O(T^{-1})$ is*

$$\text{Bias}(\hat{\beta}_A) = \mathbb{E}(\hat{\beta}_A - \beta) = -\frac{\tau}{\phi}\mathbf{R}\beta,\tag{4.1}$$

²The weight can be replaced by a positive part weight, that is, when $\tau/\mathbf{D} < 0$, we assign a zero weight to the restricted estimator. It is easy to verify that the MSE of the positive part is smaller. However, it will not change the approximations, so for simplicity we do not impose it at this stage. Nevertheless, the Monte Carlo and the application results are reported using the positive part weights.

and the MSEM of the average estimator up to order $O(T^{-2})$ is

$$\begin{aligned} MSEM(\hat{\beta}_A) &= \mathbb{E} \left[(\hat{\beta}_A - \beta)(\hat{\beta}_A - \beta)' \right] = MSEM(\hat{\beta}) + \frac{\tau^2}{\phi^2} \mathbf{R}\beta\beta'\mathbf{R}' - \frac{2\tau}{\phi} \mathbf{R}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{R}' \\ &\quad + 2\frac{\tau}{\phi^2} \left[\mathbf{R}\beta\beta'\mathbf{R}'\mathbf{W}\mathbf{R}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{R}' + \mathbf{R}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{R}'\mathbf{W}\mathbf{R}\beta\beta'\mathbf{R}' \right], \end{aligned} \quad (4.2)$$

and for the symmetric positive definite weight matrix \mathbf{W} of order $O(T)$, the risk of the average estimator up to order $O(T^{-1})$ is

$$\begin{aligned} Risk(\hat{\beta}_A) &= \mathbb{E} \left[(\hat{\beta}_A - \beta)' \mathbf{W} (\hat{\beta}_A - \beta) \right] = Risk(\hat{\beta}) + \frac{\tau}{\phi} \left[\tau - 2 \left[tr(\mathbf{P}) - 2\frac{\phi_p}{\phi} \right] \right] \\ &\leq Risk(\hat{\beta}) + \frac{\tau}{\phi} \left[\tau - 2 \left[tr(\mathbf{P}) - 2\lambda_{max}(\mathbf{P}) \right] \right], \end{aligned} \quad (4.3)$$

where $\phi = \beta'\mathbf{R}'\mathbf{W}\mathbf{R}\beta = O(T)$, $\phi_p = \beta'\mathbf{R}'\mathbf{W}^{1/2}\mathbf{P}\mathbf{W}^{1/2}\mathbf{R}\beta = O(T)$, $\lambda_{max}(\cdot)$ denotes the maximum eigenvalue, and $\mathbf{P} = \mathbf{W}^{1/2}\mathbf{R}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{R}'\mathbf{W}^{1/2}$.

Proof: Appendix A.

We note that,

$$MSEM(\hat{\beta}) = \left(1 + \frac{m}{T}\right) (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1} - \frac{1}{T} (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1} \mathbf{H} (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1} + o(T^{-2}), \quad (4.4)$$

see [Srivastava, 1970](#) for a proof, hence

$$Risk(\hat{\beta}) = \left(1 + \frac{m}{T}\right) tr \left[\mathbf{W} (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1} \right] - \frac{1}{T} tr \left[\mathbf{W} (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1} \mathbf{H} (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1} \right] + o(T^{-1}), \quad (4.5)$$

where $\mathbf{H} = \mathbf{X}'(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Phi})\mathbf{X} - \mathbf{X}'\boldsymbol{\Pi}\mathbf{X}$.

From [Theorem 1](#), it follows that the average estimator dominates the unrestricted GLS estimator in terms of having a smaller risk, when the second term on the right-hand side of equation (4.3) is negative, which will hold when

$$0 < \tau < 2 \left[tr(\mathbf{P}) - 2\frac{\phi_p}{\phi} \right], \quad (4.6)$$

given $tr(\mathbf{P}) > 2\phi_p/\phi$. As the upper bound of the condition above depends on the slope parameters, one could replace it with an infimum value. Let d be defined as

$$d = \frac{tr(\mathbf{P})}{\lambda_{max}(\mathbf{P})}, \quad (4.7)$$

which lies in the range $d \in [0, (m-1)k]$, as \mathbf{P} is a non-zero positive semi-definite matrix. Therefore,

when $d > 2$, an infimum value for the upper bound is $2 \left[\text{tr}(\mathbf{P}) - 2 \lambda_{max}(\mathbf{P}) \right]$.³ Therefore, given $d > 2$, an equivalent condition for the condition in (4.6) can be written as

$$0 < \tau \leq 2 \left[\text{tr}(\mathbf{P}) - 2 \lambda_{max}(\mathbf{P}) \right]. \quad (4.8)$$

In other words, when $d > 2$, and τ satisfies the condition in equation (4.8), the risk of the average estimator is less than the risk of the unrestricted GLS estimator up to the order of interest. In addition, as the choice of the characteristic parameter is user-specified, its optimal value, τ_{opt} , that minimizes the upper bound of the risk of the average estimator (the last term in equation (4.3)), up to order $O(T^{-1})$, provided $d > 2$, is

$$\tau_{opt} = \text{tr}(\mathbf{P}) - 2 \lambda_{max}(\mathbf{P}). \quad (4.9)$$

Since the optimal τ depends on the unknown value of $\mathbf{\Omega}$, one could substitute it with its estimated value, and use an estimate of τ_{opt} , as below

$$\hat{\tau}_{opt} = \text{tr}(\hat{\mathbf{P}}) - 2 \lambda_{max}(\hat{\mathbf{P}}),$$

where $\hat{\mathbf{P}} = \mathbf{W}^{1/2} \hat{\mathbf{R}} (\mathbf{X}' \hat{\mathbf{\Omega}}^{-1} \mathbf{X})^{-1} \hat{\mathbf{R}}' \mathbf{W}^{1/2}$, is an estimate of \mathbf{P} .

Corollary 1.1: *Under Assumptions 1–3, when $d > 2$, then up to order $O(T^{-1})$ we have*

$$\text{Risk}(\hat{\beta}_A) \leq \text{Risk}(\hat{\beta}) - \frac{\tau_{opt}^2}{\phi} < \text{Risk}(\hat{\beta}), \quad (4.10)$$

where $\hat{\beta}_A$ is the average estimator with $\hat{\tau}_{opt}$.

Proof: Appendix A.

Two arbitrary choices of \mathbf{W} are $T \mathbf{I}_{mk}$, and $\mathbf{X}' \mathbf{\Omega}^{-1} \mathbf{X}$, where the former one in the risk gives the mean squared error and the latter one, provides the (in-sample) mean squared forecast error (MSFE).

Corollary 1.2: *Under Assumptions 1–3, when $0 < \tau < 2 \left[(m-1)k - 2 \right]$, then up to order $O(T^{-1})$, we have*

$$\text{MSFE}(\hat{\beta}_A) = \text{MSFE}(\hat{\beta}) + \frac{\tau}{\phi} \left[\tau - 2 \left[(m-1)k - 2 \right] \right] < \text{MSFE}(\hat{\beta}).$$

The optimal value of τ that minimizes the MSFE of the average estimator, provided $(m-1)k > 2$, is

$$\tau_{opt,F} = (m-1)k - 2, \quad (4.11)$$

³See equation (A.19) in Appendix.

and the associated optimal MSFE of the average estimator up to order $O(T^{-1})$ is

$$MSFE_{opt}(\widehat{\beta}_A) = MSFE(\widehat{\beta}) - \frac{[(m-1)k-2]^2}{\phi}, \quad (4.12)$$

where

$$MSFE(\widehat{\beta}) = (1 + \frac{m}{T})mk - \frac{1}{T} \text{tr} [\mathbf{H}(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}].$$

Proof: Appendix A.

5 Monte Carlo Simulation

The results below are the simulation results of the model of Section 2, where $x_{it,1} = 1$ and the remaining regressors are independently generated from the standard normal distribution. The sample size varies from $T = 100$, $m = 3, 6$, and $k = 3, 5$, leading to four combinations of m , and k . u_{1t} is generated as $IIDN(0, 1)$, while $u_{it} = cu_{1t} + v_{it}$, for $i = 2, \dots, m$, where $v_{it} \sim IIDN(0, 1)$ and $c = 0.5$. We consider two DGPs for generating β_i 's, the first one is under a complete heterogeneity in coefficients where we assume that

$$\text{DGP1: } \beta_i = \bar{\beta} + (i \times \delta)/m, \quad i = 1, 2, \dots, m,$$

with $\bar{\beta} = (1, 1, \dots, 1)'$, and the second DGP is under a weak heterogeneity where we assume that

$$\text{DGP2: } \beta_{i1}, \beta_{i2} = \begin{cases} 1 + (i \times \delta)/m, & \text{if } i = 1, \dots, [m/2] \\ 1.2, & \text{if } i = [m/2] + 1, \dots, m \end{cases}, \beta_{il} = 2, \quad l \in \{3, \dots, k\},$$

where $[m/2]$ denotes the largest integer value that is smaller than $m/2$, and δ takes values on a 10-point grid on $[0, 1]$.

The results of 1,000 monte carlo simulations are given in Figures 1–4, where the vertical axes measure the relative mean squared error (RMSE) of the unrestricted GLS estimator, the restricted GLS estimator and the average estimator to the unrestricted GLS estimator. Hence, the RMSE of the unrestricted GLS estimator is equal to one. The horizontal axes measure the degree of parameter heterogeneity, δ , which is set between zero and one with 0.1 grid value.

The Monte Carlo results support our theoretical findings of the previous section. The figures show that the RMSE of the average estimator for the whole parameter heterogeneity is below that of the unrestricted estimator. This shows the superiority of the average estimator relative to the unrestricted GLS estimator.

The RMSE of the average estimator in DGP1 of a complete heterogeneous SUR, is smaller than that of the restricted GLS estimator except for very small values of parameter heterogeneity (δ). This is expected because as δ takes higher values, the bias of the restricted GLS estimator

increases, which then results in higher MSE. In DGP2 where the SUR is characterized by some degrees of homogeneity, the RMSE of the restricted GLS estimator remains smaller than that of the unrestricted GLS estimator for larger values of δ relative to DGP1. In this case, the unrestricted GLS estimator can be inferior to the restricted GLS estimator even with the presence of weak degrees of heterogeneity. This is because although the unrestricted GLS estimator is unbiased, it is inefficient, especially under small sample sizes, and a high number of regressors. In contrast, the restricted GLS estimator properly makes the use of cross equation variations and thus provides a more accurate results.

In general, we find that the average estimator performs robustly well in SUR with various degrees of heterogeneity. When there is a strong heterogeneity, the average estimator prevails. When there is a relatively weak heterogeneity, the average estimator tends to gain more from the efficiency of the restricted GLS estimator by assigning a high weight to this estimator and thus still remains one of the best choices.

6 Application: Returns to Scale in US Banking Industry

In this section we apply the average estimator studied in the previous sections to regressions of the cost system for US commercial banks. We are interested in estimating the returns to scale (RTS) for these banks over the past recent years.

Over the past few years, the number of U.S. commercial banks fell by almost 70%, where in 1984 the total number of U.S. commercial banks was 14,391 and dropped to 4,773 in 2018. Over the same period of time, the average asset value of U.S. banks (adjusted for inflation), which is also a measure of bank size, increased by about ten times, from 140 million dollars in 1984 to 1,400 million dollars in 2018 (See Figure 5). To support this bank size expansion, bank executives and analysts claim that due to the changes in regulation (such as the permission of interstate branching and combination of banks) and because of technological and financial innovations (such as communication technologies, the securitization and sale of bank loans) over the past few years, the cost of production for larger banks has reduced and encouraged banks to grow larger and/or merge.

On the other hand, critics contend that this decrease in the number of operating banks, and having banks with large assets not only impact the market competition, but also result in agency problems and disproportionate benefits of government policies in favor of large banks. In particular, the financial crisis of 2007 focused attention on large financial institutions considered as “too-big-to-fail”. These together have brought attention of policy makers for regulatory limits on bank size. However, any policy intervention needs to consider the potential efficiency benefits of operating at a large scale. Therefore, estimation of scale economies and RTS is essential for analyzing the costs and benefits of any policy intervention to control the size of banks.

The estimation of scale economies and RTS for U.S. banking industry has stimulated a substantial body of studies. Older empirical studies that used data from the 1980s and 1990s did not find scale economies in banking industry except for very small banks. But recent

research that used data from the 2000s, and more modern methods for estimating the banking models, has found considerably more evidence of scale economies in banking. These studies include (Hughes et al., 1996, 2000, 2001; Berger and Mester, 1997; Hughes and Mester, 1998, 2013; Wheelock and Wilson, 2001; Feng and Serletis, 2009). Most of the studies in the literature partition banks based on their asset size in groups and estimate each group independently from the other groups. However, not only there is no reason to believe these categories are set based on banks underlying technology, at the same time, it is hard to believe that these groups are not affected by some unknown factors that could have resulted in correlations between groups. Therefore, there are two issues that need to be carefully considered by researchers in the estimation. First, estimating cost efficiency using all observations as one group has the advantage of smaller variance but at the same time, it means ignoring the potential heterogeneity bias due to difference in technology. Second, partitioning banks in different groups could also suffer from the heterogeneity bias as the asset size caps could be specified incorrectly. In addition, single-equation estimators are inefficient compared to pooled estimators which ignore heterogeneity. Hence, there is a trade-off between bias and variance efficiency between these two estimators. As the average estimator introduced in the previous sections results in the optimal balance between bias and variance efficiency, we recommend using this estimator in estimation of the returns to scale for banking industry to obtain robust and efficient estimators.

6.1 The Model

We follow the so called “intermediation approach” framework of Sealey and Lindley, 1977, which is broadly employed in the literature. According to this approach, a bank’s balance sheet is assumed to capture the essential structure of a bank’s core business. Inputs are considered to be liabilities (core deposits and purchased funds), physical capital and labor. The inputs result in the bank’s productions which are assets (other than the physical, includes loans and trading securities).

With regard to variable specification, we define five inputs and five outputs that are the ones used in the literature. We define the following output quantities: consumer loans (y_1), real estate loans (y_2), loans to business and other institutions (y_3), federal funds sold and securities purchased under agreements to resell (y_4), and other assets (y_5). The input variables are: labor quantities (x_1), premises and fixed assets (x_2), purchased funds (x_3), interest-bearing transaction accounts (x_4), and non-transaction accounts (x_5). For each input x_j , its price w_j is obtained by dividing its total expenses by the corresponding input quantities.

For modeling the cost of banking industry, we consider a translog cost function and normalize it, so that the homogeneity (in input prices) property is automatically satisfied. We allow for individual (fixed) effects by adding intercepts in each regression, to control for specific group characteristics, heterogeneity in skills and so on. Hence, the cost equation for each group $i = 1, 2, \dots, m$, is

considered as

$$\begin{aligned} \ln(C_{it}/w_{5,it}) &= \beta_{0,i} + \sum_{n=1}^4 \beta_{n,i} \ln(w_{n,it}/w_{5,it}) + \sum_{l=1}^5 \gamma_{l,i} \ln(y_{l,it}) + \frac{1}{2} \sum_{n=1}^5 \sum_{l=1}^5 \gamma_{nl,i} \ln(y_{n,it}) \ln(y_{l,it}) \\ &+ \frac{1}{2} \sum_{n=1}^4 \sum_{l=1}^4 \eta_{nl,i} \ln(w_{n,it}/w_{5,it}) \ln(w_{l,it}/w_{5,it}) + \sum_{l=1}^4 \sum_{n=1}^5 \delta_{ln} \ln(w_{l,it}/w_{5,it}) \ln(y_{n,it}) + u_{it}, \quad t = 1, 2, \dots, T_i, \end{aligned} \quad (6.1)$$

where T_i is the number of observations in group i (the number of banks operating within group i), and C_{it} is the total cost of bank t , in group i , defined as

$$C_{it} = w_{1,it} x_{1,it} + w_{2,it} x_{2,it} + w_{3,it} x_{3,it} + w_{4,it} x_{4,it} + w_{5,it} x_{5,it}, \quad t = 1, 2, \dots, T_i, \quad \text{and } i = 1, \dots, m. \quad (6.2)$$

The cost function is symmetric which requires the imposition of the following restrictions on the parameters as below

$$\begin{aligned} \eta_{jq,i} &= \eta_{qj,i} \\ \gamma_{nl,i} &= \gamma_{ln,i}. \end{aligned} \quad (6.3)$$

RTS is defined as the inverse of the sum of cost elasticities. If we define the output elasticity of the model for output j of bank t in group i , as $Ecy_{j,it} = \frac{\partial \ln(C_{it})}{\partial \ln(y_{j,it})}$ and the sum of cost elasticities as $Ecy_{it} = \sum_{j=1}^5 \frac{\partial \ln(C_{it})}{\partial \ln(y_{j,it})}$, then RTS of bank t in group i is defined as $RTS_{it} = 1/Ecy_{it}$. Also, the RTS for group i is a vector of $T_i \times 1$ defined as

$$RTS_i = \begin{bmatrix} RTS_{i1} \\ RTS_{i2} \\ \vdots \\ RTS_{iT_i} \end{bmatrix}, \quad (6.4)$$

which is used for calculating mean, quartiles, and deciles of RTS for group i with T_i banks, see section 6.3.

A bank with $RTS > 1$, has increasing returns to scale, that is for one percent increase in all outputs, cost is increased by less than one percent, and the bank is operating below its efficient scale size ($RTS = 1$) when $RTS < 1$.

6.2 The Data

The data we use is obtained from the Reports of Income and Condition (Call Reports)⁴, over the period from 2000 to 2018. We omit observations where negative values for assets, equity, outputs, and prices are reported. The summary of the data for years 2000 and 2018 is in Table 1⁵.

Following [Feng and Serletis, 2009](#) and others in the literature, we classify the banks into three groups which is mainly based on the standard asset size categories that are used by the Federal Financial Institutions Examination Council (FFIEC). Banks with over \$500 million in total assets are classified as Large banks, banks with assets between \$100 million and \$500 million are classified as Medium banks, and banks with under \$100 million in assets are classified as Small banks. In order to have a consistent partitions over time, the asset size caps in each year are justified upward by the growth in the CPI. Table 2 presents the number and share of banks in each group with their corresponding asset ranges for years 2000 and 2018.

6.3 Estimation

We estimate model of equation (6.1) using the average estimation method developed in the previous sections for each year separately. Basically, our SUR at each year consists of three cost equations representing Large, Medium, and Small bank groups, and the observations for each regression are the operating banks data under each bank group. Since the sample size for each group is different, we face a SUR with unequal number of observations and to estimate the variance-covariance matrix (Ω), we consider the following procedures recommended in the literature (See [Schmidt, 1977](#); [Baltagi et al., 1989](#)):

1. Ignore the extra observations in estimating Ω ;
2. Use the extra observations to estimate variances. This procedure has the disadvantage of producing estimates of Ω that are not positive definite;
3. Use the extra observations to estimate variances, and modifying the estimates of covariances using the method of [Srivastava and Zaatar, 1973](#);
4. Use all observations in estimation, following the method of [Hocking and Smith, 1968](#).

It is known in the literature that the results of the above procedures are much the same. Likewise, we find that the procedures above, generate similar results, so we only report the results of method 3.

After estimating equation (6.1), we obtain the sum of cost elasticities for each bank by

$$Ecy_{it} = \sum_{j=1}^5 \frac{\partial \ln(C_{it})}{\partial \ln(y_{j,it})} = \sum_{j=1}^5 \gamma_{j,i} + \sum_{j=1}^5 \sum_{n=1}^5 \gamma_{nj,i} \ln(y_{n,it}) + \sum_{j=1}^5 \sum_{n=1}^4 \delta_{nj} \ln(w_{n,it}/w_{5,it}), \quad (6.5)$$

⁴The data from 2000–2010 is downloaded from the Federal Reserve Bank of Chicago website, and the rest of the data from 2011–2018 is downloaded from the FFIEC Central Data Repository’s Public Data Distribution website.

⁵The summary of the data for other years are not reported to save the space, but it is available upon request.

where Ecy_{it} is the cost elasticity of bank t in group i , and the parameters are replaced with their estimated values. Then, the RTS is calculated following equation (6.4). We also obtain the RTS using the unrestricted GLS estimator, and the restricted GLS estimator.

The results of years 2000, 2009 and 2018 ⁶ are reported in Table 3, which presents the extreme deciles, quartiles and means of our estimated RTS using the three estimation methods. The patterns of mean, extreme deciles and quartiles for all years are plotted in Figure 6. The results base on the restricted estimator over the most recent years suggest increasing RTS at each decile. However, these results are not economically reasonable. On the other hand, we find evidence of decreasing RTS for almost 50% of Small and Medium banks using the unrestricted estimator over the most recent years. These contradicting results show the importance of the average estimator which is used to respond to this model uncertainty.

Comparing the RTS of banks using the average estimator over the sample period shows that, on average majority of banks have increasing returns to scale. In most recent years, the results exhibit more signs of cost efficiency for Large and Small banks, such that all of Large banks have increasing RTS and only less than 25% of Small banks have exhausted their cost efficiency. However, we find that more than 50% of Medium banks are operating under decreasing returns to scale near the end of the sample. The results are consistent with some recent studies (e.g., References [Feng and Serletis, 2009](#); [Hughes and Mester, 2013](#); [Wheelock and Wilson, 2001](#); [Henderson et al., 2015](#); [Mailkov et al., 2015](#)) although we are not aware of any study from 2011 to 2018.

7 Conclusions

In this paper, we introduce an averaging estimator for a SUR model. The introduced estimator is a weighed average of an unrestricted GLS estimator which is the [Zellner, 1962](#) estimator and a restricted GLS estimator. The weight is inversely related to a quadratic loss function which measures the weighted distance between the unrestricted and the restricted GLS estimators. The bias, MSE matrix, and risk of the average estimator using the large-sample approximations of [Nagar, 1959](#) are derived. The superiority conditions of the average estimator in terms of the weighted mean squared error is given for any user-specific symmetric positive definite weight matrix, and is not limited to the case where the weight is the inverse of the variance-covariance matrix of the unrestricted GLS estimator.

We also provide some Monte Carlo results which support our theoretical claims. Finally, as our estimator is motivated by economic theory, we use U.S. Commercial banking data, and estimate a cost system for the banking industry to show how our estimator can be used in the applied work. We also estimate the cost system using single-equation least squares and a pooled estimator, and compare them with our proposed average estimator. We found more reliable estimation results with the cost system using our average estimator than the other estimators. We found that on average majority of banks have been operating under increasing returns to scale over the sample

⁶We only report the results for these three years to save the space. However, the results for other years are available upon request.

period from 2000 to 2018.

References

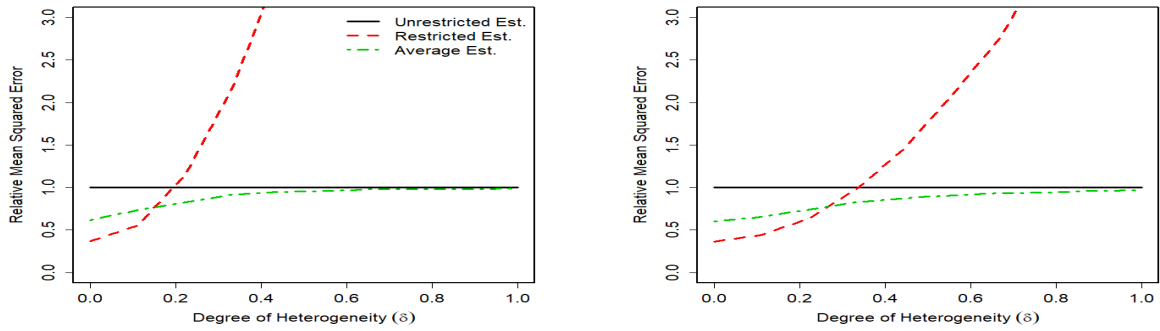
- Abadir, Karim M., and Jan R. Magnus. 2005. *Matrix Algebra*. New York: Cambridge University Press.
- Baltagi, Badi H., and James M. Griffin. 1984. Short and long run effects in pooled models. *International Economic Review* 25(3): 631–45.
- Baltagi, Badi H., James M. Griffin, and Weiwen Xiong. 2000. To pool or not to pool: Homogeneous versus heterogeneous estimations applied to cigarette demand. *The Review of Economics and Statistics* 82(1): 117–26.
- Baltagi, Badi H., Susan Garvin and Stephen Kerman. 1989. Further Monte Carlo Evidence on Seemingly Unrelated Regressions with Unequal Number of Observations. *Annales D'Economie et de Statistique* 14: 103–15.
- Berger, Allen N., and Loretta J. Mester. 1997. Inside the black box: What explains differences in the efficiencies of financial institutions? *Journal of Banking and Finance* 21(7): 895–947.
- Browning, Martin, and Jesus Carro. 2007. Heterogeneity and microeconometrics modeling. In *Advances in Economics and Econometrics*. Edited by R. Blundell, W. Newey and T. Persson. Cambridge: Cambridge University Press, vol. 3, pp. 47–74.
- Choi, Hak, and Hongyi Li. 2000. Economic development and growth convergence in China. *Journal of International Trade and Economic Development* 9(1): 37–54.
- Durlauf, Steven N., Andros Kourtellos, and Artur Minkin. 2001. The local Solow growth model. *European Economic Review* 45(4–6): 928–40.
- Feng, Guohua, and Apostolos Serletis. 2009. Efficiency and productivity of the US banking industry, 1998–2005: Evidence from the Fourier cost function satisfying global regularity conditions. *Journal of Applied Econometrics* 24(1): 105–38.
- Fiebig, Denzil G. 2001. Seemingly Unrelated Regression. *A Companion to Theoretical Econometrics*. Edited by B. H. Baltagi. Oxford: Blackwell. chp. 5.
- James, W., and Charles Stein. 1961. Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* 1: 361–80.
- Hansen, Bruce E. 2016. Efficient Shrinkage in Parametric Models. *Journal of Econometrics* 190(1): 115–32.
- Henderson, Daniel J., Subal C. Kumbhakar, Qi Li, and Christopher F. Parameter. 2015. Smooth coefficient estimation of a seemingly unrelated regression. *Journal of Econometrics* 189(1): 148–62.

- Hocking, Ronald R., and William B. Smith. 1968. Estimation of Parameters in the Multivariate Normal Distribution with Missing Observations. *Journal of the American Statistical Association* 63: 154–73.
- Hoogstrate, André J., Franz C. Palm, and Gerald A. Pfann. 2000. Pooling in dynamic panel-data models: An application to forecasting GDP growth rates. *Journal of business and Economic Studies* 18(3): 274–83.
- Hughes, Joseph P., and Loretta J. Mester. 1998. Bank capitalization and cost: Evidence of scale economies in risk management and signaling. *The Review of Economics and Statistics* 80(2): 314–25.
- Hughes, Joseph P., and Loretta J. Mester. 2013. Who said large banks don't experience scale economies? Evidence from a risk-return-driven cost function. *Journal of Financial Intermediation* 22(4): 559–85.
- Hughes, Joseph P., William Lang, Loretta J. Mester, and Choon-Geol Moon. 1996. Efficient banking under interstate branching. *Journal of Money, Credit, and Banking* 28(4): 1045–71.
- Hughes, Joseph P., William Lang, Loretta J. Mester, and Choon-Geol Moon. 2000. Recovering risky technologies using the almost ideal demand system: An application to US banking. *Journal of Financial Services Research* 18: 5–27.
- Hughes, Joseph P., Loretta J. Mester, and Choon-Geol Moon. 2001. Are scale economies in banking elusive or illusive?: Evidence obtained by incorporating capital structure and risk-taking into models of bank production. *Journal of Banking and Finance* 25(12): 2169–208.
- Leeb, Hannes, and Benedikt M. Pötscher. 2005. Model Selection and Inference: Facts and Fiction. *Econometric Theory* 21(1): 21–59.
- Maddala, G. S., and Wanhong Hu. 1996. The pooling problem. In *The Econometrics of Panel Data*. Edited by L. Matyas and P. Sevestre. Advanced Studies in Theoretical and Applied Econometrics, vol 33. Springer, Dordrecht, pp. 307–22.
- Maddala, G. S., Hongyi Li, and V. K. Srivastava. 2001. A Comparative Study of Different Shrinkage Estimators for Panel Data Models. *Annals of Economics and Finance* 2: 1–30.
- Maddala, G. S., Robert P. Trost, Hongyi Li, and Frederick Joutz. 1997. Estimation of short run and long run elasticities of energy demand from panel data using shrinkage estimators. *Journal of Business and Economic Statistics* 15(1): 90–100.
- Maddala, G. S. 1991. To pool or not to pool: That is the question. *Journal of Quantitative Economics* 7: 255–64.
- Mailkov, Emir, Diego Restrepo-Tobon, and Subal C. Kumbhakar. 2015. Estimation of banking technology under credit uncertainty. *Empirical Economics* 49: 185–211.

- Nagar, A. L. 1959. The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations. *Econometrica* 27(4): 575–95.
- Pesaran, M. Hashem, and Ron Smith. 1995. Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics* 68(1): 79–113.
- Robertson, Donald, and James Symons. 1992. Some strange properties of panel data estimators. *Journal of Applied Econometrics* 7(2): 175–89.
- Schmidt, Peter 1977. Estimation of Seemingly Unrelated Regressions with Unequal Numbers of Observations. *Journal of Econometrics* 5(3): 365–77.
- Sclove, Stanley L. 1968. Improved estimators for coefficients in linear regression. *Journal of American Statistical Association* 63(322): 596–606.
- Sealey, C.W., and James T. Lindley. 1977. Inputs, outputs, and a theory of production and cost at depository financial institutions. *Journal of Finance* 32(4): 1251–66.
- Srivastava, Virendra K. 1970. The efficiency of estimating seemingly unrelated regression equations. *Annals of the Institute of Statistical Mathematics* 483-93.
- Srivastava, Virendra K., and David E. A. Giles. 1987. *Seemingly Unrelated Regression Models: Estimation and Inference*. New York: Marcel Dekker.
- Srivastava, Virendra K., and Tryambkeshwar D. Dwivedi. 1979. Estimation of seemingly unrelated regression equations: A brief survey. *Journal of Econometrics* 10(1): 15–32.
- Srivastava, Virendra K. 1973. The Efficiency of an Improved Method of Estimating Seemingly Unrelated Regression Equations. *Journal of Econometrics* 1(4): 341–50.
- Srivastava, J. N., and M. K. Zaatar. 1973. A Monte Carlo Comparison of Four Estimators of Dispersion Matrix of a Bivariate Normal Population, Using Incomplete Data. *Journal of the American Statistical Association* 68(341): 180–83.
- Su, Liangjun, and Qihui Chen. 2013. Testing homogeneity in panel data models with interactive fixed effects. *Econometric Theory* 29(6): 1079–135.
- Ullah, Aman, and Huansha Wang. 2013. Parametric and Nonparametric Frequentist Model Selection and Model Averaging. *Econometrics* 1: 157–79.
- Wang, Wendun, Xinyu Zhang, and Richard Paap. 2019. To Pool or Not to Pool: What is a Good Strategy? *Journal of Applied Econometrics* 34(5): 724–45.
- Wheelock, David C., and Paul W. Wilson. 2001. New evidence on returns to scale and product mix among US commercial banks. *Journal of Monetary Economics* 47(3): 653–74.

Zellner, Arnold, and Walter Vandaele. 1975. Bayes-Stein estimators for k-means, regression and simultaneous equation models in *Studies in Bayesian Econometrics and Statistics*. Edited by S. E. Fienberg and A. Zellner. Amsterdam: North-Holland Publishing Company, pp. 627–53.

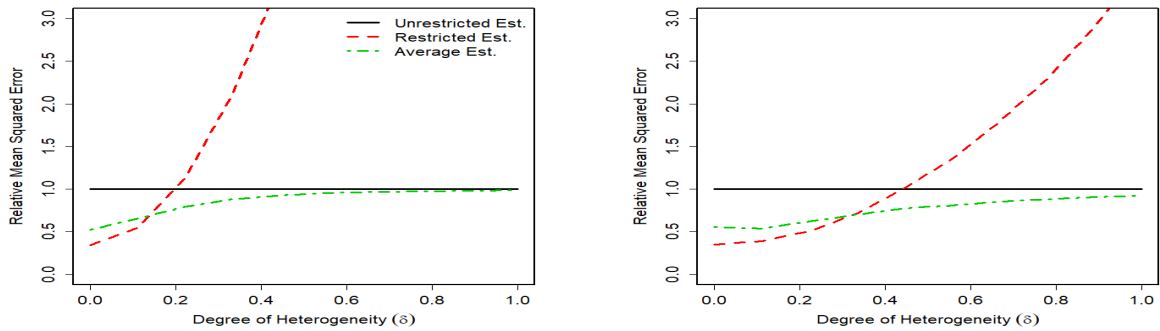
Zellner, Arnold. 1962. An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias. *Journal of the American Statistical Association* 57(298): 348–68.



(a) DGP1

(b) DGP2

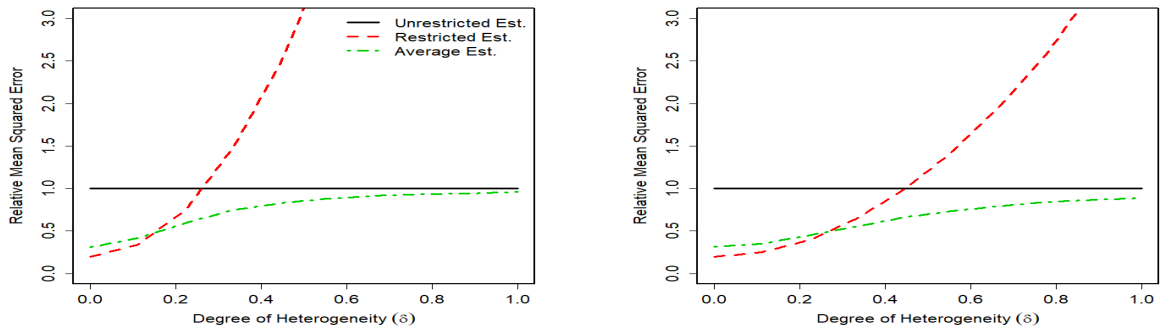
Figure 1: RMSE of Unrestricted, Restricted and Average Estimators, for $T = 100$, $m = 3$, $k = 3$



(a) DGP1

(b) DGP2

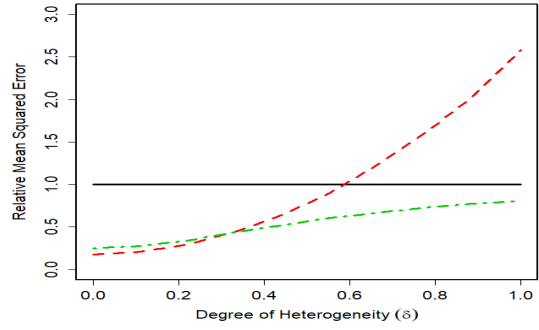
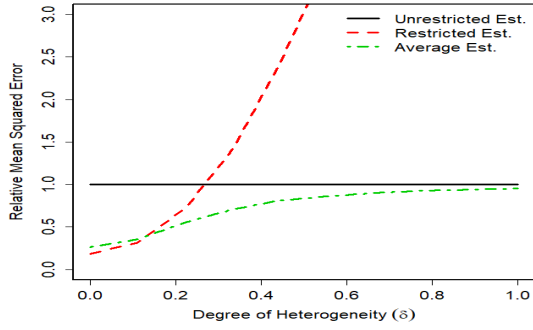
Figure 2: RMSE of Unrestricted, Restricted and Average Estimators, for $T = 100$, $m = 3$, $k = 5$



(a) DGP1

(b) DGP2

Figure 3: RMSE of Unrestricted, Restricted and Average Estimators, for $T = 100$, $m = 6$, $k = 3$



(a) DGP1

(b) DGP2

Figure 4: RMSE of Unrestricted, Restricted and Average Estimators, for $T = 100$, $m = 6$, $k = 5$

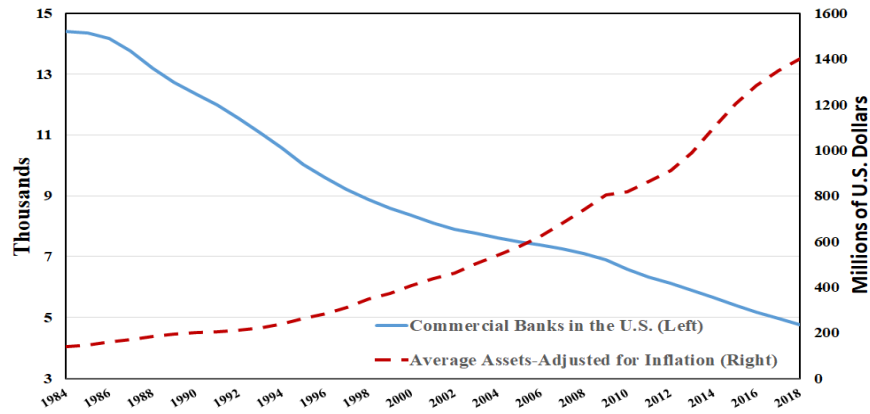


Figure 5: United States Commercial Banks, Number and Average Assets (source: Federal Reserve Bank of St. Louis FRED database)

Table 1: Summary Statistics.

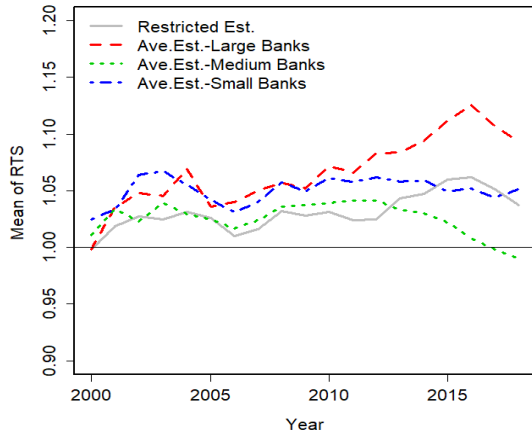
Variable	Min		Max	
	2000	2018	2000	2018
C	136.49	151.78	18,369,359.04	26,757,643.39
w_1	7015.14	15321.43	163358.47	261240.11
w_2	6.75	9.61	12375.00	43469.79
w_3	3.20	0.02	206.53	48.82
w_4	0.26	0.02	265.36	153.96
w_5	0.46	0.06	49.21	16.82
y_1	1.25	1.00	42,638,250.00	173,922,000.00
y_2	0.25	170.50	168,465,250.00	686,161,250.00
y_3	45.75	125.50	178,056,500.00	457,517,750.00
y_4	89.50	0.25	144,188,250.00	703,099,250.00
y_5	39.00	42.50	86,346,000.00	704,384,250.00

Variable	Mean		STD	
	2000	2018	2000	2018
C	24,454.49	43,803.47	331,091.51	643,872.50
w_1	26175.32	49386.16	7084.42	15332.39
w_2	230.31	265.78	351.89	920.70
w_3	33.22	3.51	6.38	2.75
w_4	14.79	2.43	9.01	3.87
w_5	22.87	3.46	4.79	1.90
y_1	57,014.23	268,717.11	754,295.15	4,722,816.78
y_2	175,843.57	881,915.93	2,961,928.59	15,632,429.24
y_3	206,836.22	929,043.01	2,423,963.16	10,526,805.47
y_4	192,820.92	862,701.61	2,985,314.66	16,392,980.08
y_5	85,019.93	842,793.39	1,554,035.20	16,868,329.90

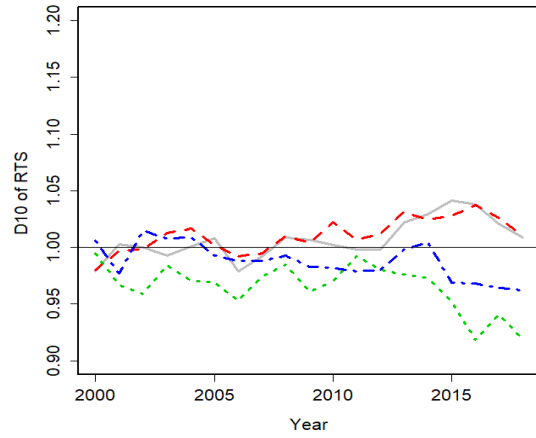
Note: All variables except $w_1 - w_5$, are measured in thousands of dollars.

Table 2: Banks Asset Size Classes

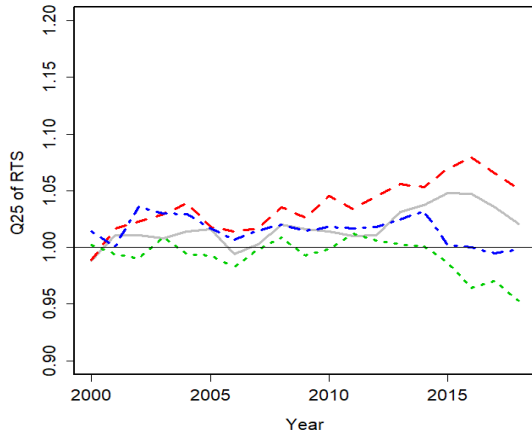
Bank Groups	Asset Size	Number of Banks	Share of Banks
	(in Millions of Dollars of the year)		
2000			
Large Banks	Assets \geq 500	739	8.9
Medium Banks	$100 \leq$ Assets $<$ 500	2946	35.5
Small Banks	Assets $<$ 100	4620	55.6
2018			
Large Banks	Assets \geq 729	988	19.8
Medium Banks	$146 \leq$ Assets $<$ 729	2298	46.1
Small Banks	Assets $<$ 146	1704	34.1



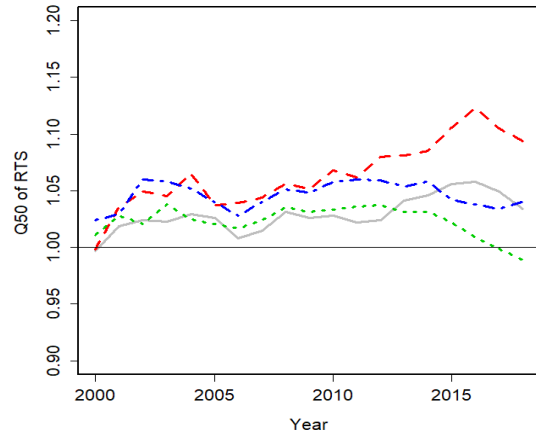
(a) Mean



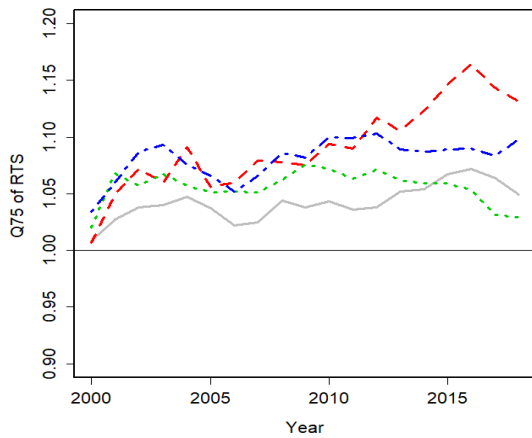
(b) 10th Decile



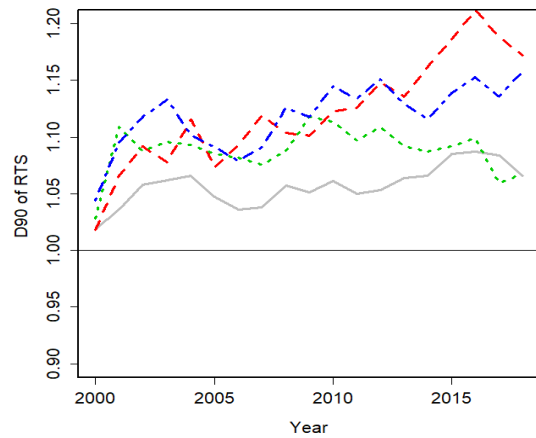
(c) 25th Quartile



(d) 50th Quartile



(e) 75th Quartile



(f) 90th Decile

Figure 6: RTS Estimates Using The Restricted and The Average Estimator over period 2000–2018

Table 3: Summary of estimates of returns to scale

Bank Size	Estimator	Estimates of RTS					
		D_{10}	Q_{25}	Q_{50}	Q_{75}	D_{90}	Mean
2000							
Large Banks	Restricted	0.979	0.988	0.997	1.008	1.019	0.998
	Unrestricted	0.971	0.985	0.998	1.009	1.021	0.996
Medium Banks	Average Estimator	0.980	0.989	0.998	1.007	1.018	0.998
	Unrestricted	0.969	0.992	1.017	1.043	1.071	1.019
Small Banks	Average Estimator	0.995	1.002	1.011	1.020	1.028	1.011
	Unrestricted	0.997	1.014	1.034	1.055	1.075	1.035
	Average Estimator	1.006	1.014	1.024	1.034	1.044	1.025
2009							
Large Banks	Restricted	1.007	1.016	1.026	1.038	1.051	1.028
	Unrestricted	1.001	1.025	1.053	1.080	1.108	1.055
Medium Banks	Average Estimator	1.004	1.026	1.051	1.075	1.101	1.052
	Unrestricted	0.956	0.990	1.032	1.079	1.125	1.038
Small Banks	Average Estimator	0.961	0.993	1.031	1.076	1.118	1.037
	Unrestricted	0.979	1.013	1.049	1.085	1.124	1.050
	Average Estimator	0.983	1.015	1.048	1.082	1.118	1.049
2018							
Large Banks	Restricted	1.009	1.021	1.034	1.049	1.066	1.037
	Unrestricted	1.009	1.052	1.100	1.140	1.184	1.098
Medium Banks	Average Estimator	1.010	1.052	1.094	1.132	1.172	1.093
	Unrestricted	0.915	0.950	0.988	1.030	1.070	0.989
Small Banks	Average Estimator	0.919	0.953	0.989	1.029	1.068	0.990
	Unrestricted	0.960	0.997	1.042	1.104	1.166	1.054
	Average Estimator	0.962	0.998	1.040	1.098	1.158	1.051

Note: Decile, quartile and mean estimates of RTS for the restricted, unrestricted, and average estimators. D_{10} , D_{90} , Q_{25} , Q_{50} , and Q_{75} are the lower decile, the upper decile, the lower quartile, median and upper quartile, respectively.

A Appendix A

Lemma 1. Under assumptions 1–3, we have the followings

$$\widehat{\Omega}^{-1} = \underbrace{\Omega^{-1}}_{O_p(1)} - \underbrace{\Omega^{-1}\Delta\Omega^{-1}}_{O_p(T^{-1/2})} + \underbrace{\Omega^{-1}\Delta\Omega^{-1}\Delta\Omega^{-1}}_{O_p(T^{-1})} - \underbrace{\Omega^{-1}\Delta\Omega^{-1}\Delta\Omega^{-1}\Delta\Omega^{-1}}_{O_p(T^{-3/2})} + O_p(T^{-2}), \quad (\text{A.1})$$

$$(\mathbf{X}'\widehat{\Omega}^{-1}\mathbf{X})^{-1} = \underbrace{(\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}}_{O_p(T^{-1})} + \underbrace{(\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}\Delta\Omega^{-1}\mathbf{X}(\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}}_{O_p(T^{-3/2})} + O_p(T^{-2}), \quad (\text{A.2})$$

$$\mathbf{X}'\widehat{\Omega}^{-1}\mathbf{u} = \underbrace{\mathbf{X}'\Omega^{-1}\mathbf{u}}_{O_p(T^{1/2})} - \underbrace{\mathbf{X}'\Omega^{-1}\Delta\Omega^{-1}\mathbf{u}}_{O_p(1)} + \underbrace{\mathbf{X}'\Omega^{-1}\Delta\Omega^{-1}\Delta\Omega^{-1}\mathbf{u}}_{O_p(T^{-1/2})} + O_p(T^{-1}), \quad (\text{A.3})$$

$$\widehat{\mathbf{R}} = \mathbf{R} + \mathbf{R}_{-1/2} + O_p(T^{-1}), \quad (\text{A.4})$$

where $\Delta = \widehat{\Omega} - \Omega$, and $\mathbf{R}_{-1/2} = \mathbf{J}(\mathbf{J}'\mathbf{X}'\Omega^{-1}\mathbf{X}\mathbf{J})^{-1}\mathbf{J}'\mathbf{X}'\Omega^{-1}\Delta\Theta\mathbf{X} = O_p(T^{-1/2})$, with the suffix showing the order of magnitude in probability, and $\Theta = \Omega^{-1} - \Omega^{-1}\mathbf{X}\mathbf{J}(\mathbf{J}'\mathbf{X}'\Omega^{-1}\mathbf{X}\mathbf{J})^{-1}\mathbf{J}'\mathbf{X}'\Omega^{-1}$.

Proof:

It can be easily verified that, under assumptions 1–3, $\Delta = O_p(T^{-1/2})$. Employing this condition, and using the standard geometric expansion for the inverse of a matrix ⁷, for large T , we have the followings

$$\begin{aligned} \widehat{\Omega}^{-1} &= (\Omega + \Delta)^{-1} = \Omega^{-1}[\mathbf{I}_{mT} + \Delta\Omega^{-1}]^{-1} \\ &= \Omega^{-1}\left[\mathbf{I}_{mT} - \Delta\Omega^{-1} + \Delta\Omega^{-1}\Delta\Omega^{-1} - \Delta\Omega^{-1}\Delta\Omega^{-1}\Delta\Omega^{-1} + \dots\right] \\ &= \underbrace{\Omega^{-1}}_{O_p(1)} - \underbrace{\Omega^{-1}\Delta\Omega^{-1}}_{O_p(T^{-1/2})} + \underbrace{\Omega^{-1}\Delta\Omega^{-1}\Delta\Omega^{-1}}_{O_p(T^{-1})} - \underbrace{\Omega^{-1}\Delta\Omega^{-1}\Delta\Omega^{-1}\Delta\Omega^{-1}}_{O_p(T^{-3/2})} + O_p(T^{-2}), \end{aligned}$$

which gives the results in equation (A.1). Now, by using equation (A.1), we have

$$\begin{aligned} (\mathbf{X}'\widehat{\Omega}^{-1}\mathbf{X})^{-1} &= \left[\mathbf{X}'\Omega^{-1}\mathbf{X} - \mathbf{X}'\Omega^{-1}\Delta\Omega^{-1}\mathbf{X} + \dots\right]^{-1} \\ &= (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\left[\mathbf{I}_{mk} - \mathbf{X}'\Omega^{-1}\Delta\Omega^{-1}\mathbf{X}(\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1} + \dots\right]^{-1} \\ &= \underbrace{(\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}}_{O_p(T^{-1})} + \underbrace{(\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}\Delta\Omega^{-1}\mathbf{X}(\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}}_{O_p(T^{-3/2})} + O_p(T^{-2}), \end{aligned}$$

⁷ $(\mathbf{I} + \mathbf{A})^{-1} = \mathbf{I} - \mathbf{A} + \mathbf{A}^2 - \mathbf{A}^3 + \dots$

also we have

$$\mathbf{X}'\widehat{\Omega}^{-1}\mathbf{u} = \underbrace{\mathbf{X}'\Omega^{-1}\mathbf{u}}_{O_p(T^{1/2})} - \underbrace{\mathbf{X}'\Omega^{-1}\Delta\Omega^{-1}\mathbf{u}}_{O_p(1)} + \underbrace{\mathbf{X}'\Omega^{-1}\Delta\Omega^{-1}\Delta\Omega^{-1}\mathbf{u}}_{O_p(T^{-1/2})} + O_p(T^{-1}).$$

By using the above results, we have

$$\begin{aligned} \widehat{\mathbf{R}} &= \mathbf{I}_{mk} - \mathbf{J}(\mathbf{J}'\mathbf{X}'\widehat{\Omega}^{-1}\mathbf{X}\mathbf{J})^{-1}\mathbf{J}'\mathbf{X}'\widehat{\Omega}^{-1}\mathbf{X} \\ &= \mathbf{I}_{mk} - \mathbf{J} \left[\underbrace{(\mathbf{J}'\mathbf{X}'\Omega^{-1}\mathbf{X}\mathbf{J})^{-1}}_{O_p(T^{-1})} + \underbrace{(\mathbf{J}'\mathbf{X}'\Omega^{-1}\mathbf{X}\mathbf{J})^{-1}\mathbf{J}'\mathbf{X}'\Omega^{-1}\Delta\Omega^{-1}\mathbf{X}\mathbf{J}(\mathbf{J}'\mathbf{X}'\Omega^{-1}\mathbf{X}\mathbf{J})^{-1}}_{O_p(T^{-3/2})} + O_p(T^{-2}) \right] \\ &\mathbf{J}' \times \left[\underbrace{\mathbf{X}'\Omega^{-1}\mathbf{X}}_{O_p(T)} - \underbrace{\mathbf{X}'\Omega^{-1}\Delta\Omega^{-1}\mathbf{X}}_{O_p(T^{1/2})} + O_p(1) \right] \\ &= \mathbf{R} + \mathbf{J}(\mathbf{J}'\mathbf{X}'\Omega^{-1}\mathbf{X}\mathbf{J})^{-1}\mathbf{J}'\mathbf{X}'\Omega^{-1}\Delta \left[\Omega^{-1} - \Omega^{-1}\mathbf{X}\mathbf{J}(\mathbf{J}'\mathbf{X}'\Omega^{-1}\mathbf{X}\mathbf{J})^{-1}\mathbf{J}'\mathbf{X}'\Omega^{-1} \right] \mathbf{X} + O_p(T^{-1}) \\ &= \mathbf{R} + \mathbf{R}_{-1/2} + O_p(T^{-1}). \end{aligned}$$

■

Proof. Theorem 1 :

Using the results of Lemma 1, in equation (3.1), we have

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}'\widehat{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\widehat{\Omega}^{-1}\mathbf{u} = \mathbf{A}_{-1/2} + \mathbf{A}_{-1} + O_p(T^{-3/2}), \quad (\text{A.5})$$

where $\mathbf{A}_{-1/2}$ and \mathbf{A}_{-1} are defined below, and the suffixes show the order of magnitude in probability,

$$\begin{aligned} \mathbf{A}_{-1/2} &= (\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}\mathbf{u} = O_p(T^{-1/2}), \\ \mathbf{A}_{-1} &= -(\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}\Delta\mathbf{Q}\mathbf{u} = O_p(T^{-1}), \end{aligned}$$

and $\mathbf{Q} = \Omega^{-1} - \Omega^{-1}\mathbf{X}(\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}$.

Using equation (A.5) in equation (3.8), we have

$$\begin{aligned} \frac{1}{\widehat{D}} &= \left[(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}})' \mathbf{W} (\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}) \right]^{-1} = \left[\widehat{\boldsymbol{\beta}}' \widehat{\mathbf{R}}' \mathbf{W} \widehat{\mathbf{R}} \widehat{\boldsymbol{\beta}} \right]^{-1} \\ &= \left[\left(\boldsymbol{\beta} + \mathbf{A}_{-1/2} + O_p(T^{-1}) \right)' \left[\mathbf{R} + \mathbf{R}_{-1/2} + O_p(T^{-1}) \right]' \mathbf{W} \left[\mathbf{R} + \mathbf{R}_{-1/2} + O_p(T^{-1}) \right] \left(\boldsymbol{\beta} + \mathbf{A}_{-1/2} + O_p(T^{-1}) \right) \right]^{-1} \\ &= \left[\phi + 2\boldsymbol{\beta}' \mathbf{R}' \mathbf{W} \mathbf{R} \mathbf{A}_{-1/2} + 2\boldsymbol{\beta}' \mathbf{R}' \mathbf{W} \mathbf{R}_{-1/2} \boldsymbol{\beta} + O_p(1) \right]^{-1} \\ &= \frac{1}{\phi} \left[1 + \frac{2}{\phi} \boldsymbol{\beta}' \mathbf{R}' \mathbf{W} \mathbf{R} \mathbf{A}_{-1/2} + \frac{2}{\phi} \boldsymbol{\beta}' \mathbf{R}' \mathbf{W} \mathbf{R}_{-1/2} \boldsymbol{\beta} + O_p(T^{-1}) \right]^{-1} \end{aligned}$$

$$= \frac{1}{\phi} \left[1 - \frac{2}{\phi} \boldsymbol{\beta}' \mathbf{R}' \mathbf{W} \mathbf{R} \mathbf{A}_{-1/2} - \frac{2}{\phi} \boldsymbol{\beta}' \mathbf{R}' \mathbf{W} \mathbf{R}_{-1/2} \boldsymbol{\beta} \right] + O_p(T^{-2}), \quad (\text{A.6})$$

where $\phi = \boldsymbol{\beta}' \mathbf{R}' \mathbf{W} \mathbf{R} \boldsymbol{\beta} = O(T)$, and the last equality above holds by using the standard geometric expansion. Also, the use has been made of equations (A.1)–(A.4). The terms with order $O_p(T^{-2})$ and smaller are dropped, because they will not enter in the calculation of the bias and MSE of the average estimator up to the orders of interest.

Employing equations (A.4) and (A.6) in equation (3.7), we obtain

$$\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta} = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \frac{\tau}{\phi} \left[1 - \frac{2}{\phi} \boldsymbol{\beta}' \mathbf{R}' \mathbf{W} \mathbf{R} \mathbf{A}_{-1/2} - \frac{2}{\phi} \boldsymbol{\beta}' \mathbf{R}' \mathbf{W} \mathbf{R}_{-1/2} \boldsymbol{\beta} \right] \left[\mathbf{R} \boldsymbol{\beta} + \mathbf{R} \mathbf{A}_{-1/2} + \mathbf{R}_{-1/2} \boldsymbol{\beta} \right] + O_p(T^{-2}). \quad (\text{A.7})$$

The bias of the average estimator using the above equation up to order $O(T^{-1})$ is

$$\mathbb{E}(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}) = \mathbb{E}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \frac{\tau}{\phi} \mathbf{R} \boldsymbol{\beta} + o_p(T^{-1}) = -\frac{\tau}{\phi} \mathbf{R} \boldsymbol{\beta} + o_p(T^{-1}), \quad (\text{A.8})$$

where the use has been made of

$$\mathbb{E}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \mathbb{E}(\mathbf{A}_{-1/2}) + \mathbb{E}(\mathbf{A}_{-1}) + o_p(T^{-1}) = \mathbf{0} + o_p(T^{-1}), \quad (\text{A.9})$$

because, both $\mathbf{A}_{-1/2}$ and \mathbf{A}_{-1} are odd functions of the error term which has a normal distribution.

The MSE matrix up to order $O(T^{-2})$ is

$$\mathbb{E} \left[(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta})' \right] = \mathbb{E} \left[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' + \boldsymbol{\Xi}_1 - \boldsymbol{\Xi}_2 - \boldsymbol{\Xi}_2' + \boldsymbol{\Xi}_3 + \boldsymbol{\Xi}_3' - \boldsymbol{\Xi}_4 - \boldsymbol{\Xi}_4' \right] + o_p(T^{-2}), \quad (\text{A.10})$$

where $\boldsymbol{\Xi}_1 - \boldsymbol{\Xi}_4$ are defined as below

$$\begin{aligned} \boldsymbol{\Xi}_1 &= \frac{\tau^2}{\phi^2} \mathbf{R} \boldsymbol{\beta} \boldsymbol{\beta}' \mathbf{R}', \\ \boldsymbol{\Xi}_2 &= \frac{\tau}{\phi} \mathbf{R} \boldsymbol{\beta} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})', \\ \boldsymbol{\Xi}_3 &= \frac{2\tau}{\phi^2} (\boldsymbol{\beta}' \mathbf{R}' \mathbf{W} \mathbf{R} \mathbf{A}_{-1/2} + \boldsymbol{\beta}' \mathbf{R}' \mathbf{W} \mathbf{R}_{-1/2} \boldsymbol{\beta}) \mathbf{R} \boldsymbol{\beta} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})', \\ \boldsymbol{\Xi}_4 &= \frac{\tau}{\phi} (\mathbf{R} \mathbf{A}_{-1/2} + \mathbf{R}_{-1/2} \boldsymbol{\beta}) (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'. \end{aligned}$$

Now, we obtain the expectations of $\boldsymbol{\Xi}_1 - \boldsymbol{\Xi}_4$.

$$\mathbb{E}(\boldsymbol{\Xi}_1) = \frac{\tau^2}{\phi^2} \mathbf{R} \boldsymbol{\beta} \boldsymbol{\beta}' \mathbf{R}', \quad (\text{A.11})$$

$$\mathbb{E}(\Xi_2) = \frac{\tau}{\phi} \mathbf{R} \boldsymbol{\beta} \mathbb{E}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \mathbf{0} + o_p(T^{-2}), \quad (\text{A.12})$$

$$\begin{aligned} \mathbb{E}(\Xi_3) &= \frac{2\tau}{\phi^2} \mathbf{R} \boldsymbol{\beta} \left[\boldsymbol{\beta}' \mathbf{R}' \mathbf{W} \mathbf{R} \mathbb{E}(\mathbf{A}_{-1/2} \mathbf{A}'_{-1/2}) + \mathbb{E}(\boldsymbol{\beta}' \mathbf{R}' \mathbf{W} \mathbf{R}_{-1/2} \boldsymbol{\beta} \mathbf{A}'_{-1/2}) \right] + o_p(T^{-2}) \\ &= \frac{2\tau}{\phi^2} \mathbf{R} \boldsymbol{\beta} \boldsymbol{\beta}' \mathbf{R}' \mathbf{W} \mathbf{R} (\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{R}' + o_p(T^{-2}), \end{aligned} \quad (\text{A.13})$$

where the last equality holds by noting that the second term on the right-hand side of the first equality in the above equation is an odd function of normal distributions, and utilizing the following two equations below

$$\mathbb{E}(\mathbf{A}_{-1/2} \mathbf{A}'_{-1/2}) = (\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbb{E}(\mathbf{u} \mathbf{u}') \boldsymbol{\Omega}^{-1} \mathbf{X} (\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} = (\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1}, \quad (\text{A.14})$$

$$\begin{aligned} \mathbf{R} (\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{R}' &= \left[\mathbf{I}_{mk} - \mathbf{J} (\mathbf{J}' \mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X} \mathbf{J})^{-1} \mathbf{J}' \mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X} \right] (\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \left[\mathbf{I}_{mk} - \mathbf{J} (\mathbf{J}' \mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X} \mathbf{J})^{-1} \mathbf{J}' \mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X} \right]' \\ &= (\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} - \mathbf{J} (\mathbf{J}' \mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X} \mathbf{J})^{-1} \mathbf{J}' \\ &= (\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{R}' = \mathbf{R} (\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1}. \end{aligned} \quad (\text{A.15})$$

Similarly, we have

$$\mathbb{E}(\Xi_4) = \frac{\tau}{\phi} \left[\mathbf{R} \mathbb{E}(\mathbf{A}_{-1/2} \mathbf{A}'_{-1/2}) + \mathbb{E}(\mathbf{R}_{-1/2} \boldsymbol{\beta} \mathbf{A}'_{-1/2}) \right] + o_p(T^{-2}) = \frac{\tau}{\phi} \mathbf{R} (\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{R}' + o_p(T^{-2}). \quad (\text{A.16})$$

Employing the results of equations (A.11)–(A.16), in equation (A.10), we obtain the MSE matrix of the average estimator up to order $O(T^{-2})$, as below

$$\begin{aligned} \text{MSEM}(\widehat{\boldsymbol{\beta}}_A) &= \text{MSEM}(\widehat{\boldsymbol{\beta}}) + \frac{\tau^2}{\phi^2} \mathbf{R} \boldsymbol{\beta} \boldsymbol{\beta}' \mathbf{R}' - \frac{2\tau}{\phi} \mathbf{R} (\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{R}' \\ &\quad + 2 \frac{\tau}{\phi^2} \left[\mathbf{R} \boldsymbol{\beta} \boldsymbol{\beta}' \mathbf{R}' \mathbf{W} \mathbf{R} (\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{R}' + \mathbf{R} (\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{R}' \mathbf{W} \mathbf{R} \boldsymbol{\beta} \boldsymbol{\beta}' \mathbf{R}' \right], \end{aligned} \quad (\text{A.17})$$

hence the risk of the average estimator up to order $O(T^{-1})$, can be written as

$$\begin{aligned} \text{Risk}(\widehat{\boldsymbol{\beta}}_A) &= \mathbb{E} \left[(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta})' \mathbf{W} \widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta} \right] = \text{tr} \left[\mathbf{W} \mathbb{E} \left[(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}) (\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta})' \right] \right] = \text{tr} \left[\mathbf{W} \text{MSEM}(\widehat{\boldsymbol{\beta}}_A) \right] \\ &= \text{Risk}(\widehat{\boldsymbol{\beta}}) + \frac{\tau^2}{\phi} - \frac{2\tau}{\phi} \text{tr} \left[\mathbf{W} \mathbf{R} (\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{R}' \right] + 4 \frac{\tau}{\phi^2} \boldsymbol{\beta}' \mathbf{R}' \mathbf{W} \mathbf{R} (\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{R}' \mathbf{W} \mathbf{R} \boldsymbol{\beta} \\ &\leq \text{Risk}(\widehat{\boldsymbol{\beta}}) + \frac{\tau^2}{\phi} - \frac{2\tau}{\phi} \text{tr} \left[\mathbf{W} \mathbf{R} (\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{R}' \right] + 4 \frac{\tau}{\phi} \lambda_{\max} \left[\mathbf{W}^{1/2} \mathbf{R} (\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{R}' \mathbf{W}^{1/2} \right], \end{aligned} \quad (\text{A.18})$$

where the last inequality holds because $\mathbf{W}^{1/2}\mathbf{R}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{R}'\mathbf{W}^{1/2}$ is symmetric, therefore

$$\frac{\boldsymbol{\beta}'\mathbf{R}'\mathbf{W}\mathbf{R}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{R}'\mathbf{W}\mathbf{R}\boldsymbol{\beta}}{\boldsymbol{\beta}'\mathbf{R}'\mathbf{W}\mathbf{R}\boldsymbol{\beta}} \leq \lambda_{max}(\mathbf{W}^{1/2}\mathbf{R}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{R}'\mathbf{W}^{1/2}), \quad (\text{A.19})$$

see [Abadir and Magnus, 2005](#), pp. 181–82. ■

Proof. Corollary 1.1:

Note that, using Lemma 1, we have

$$\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega} = O_p(T^{-1/2}), \quad (\text{A.20})$$

$$\widehat{\mathbf{R}} - \mathbf{R} = O_p(T^{-1/2}), \quad (\text{A.21})$$

and

$$\widehat{\mathbf{P}} = \mathbf{P} + \mathbf{P}_{-1/2} + O_p(T^{-1}), \quad (\text{A.22})$$

where

$$\begin{aligned} \mathbf{P}_{-1/2} &= \mathbf{W}^{1/2}\mathbf{R}_{-1/2}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{R}'\mathbf{W}^{1/2} + \mathbf{W}^{1/2}\mathbf{R}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{R}'_{-1/2}\mathbf{W}^{1/2} \\ &\quad + \mathbf{W}^{1/2}\mathbf{R}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{\Delta}\boldsymbol{\Omega}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{R}'\mathbf{W}^{1/2}. \end{aligned} \quad (\text{A.23})$$

Therefore, it is easy to see that $\widehat{\tau}_{opt} - \tau = O_p(T^{-1/2})$, hence the consistency of $\widehat{\tau}$ follows. Further, by replacing $\widehat{\tau}_{opt}$ by τ in equation (A.10), the results in equations (A.11), (A.13) and (A.16) remain unchange, and for equation (A.12) we have

$$\begin{aligned} \mathbb{E}(\widehat{\boldsymbol{\Xi}}_2) &= \mathbb{E}\left[\frac{\widehat{\tau}}{\phi}\mathbf{R}\boldsymbol{\beta}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right] \geq \mathbb{E}\left[\frac{\tau_{opt}}{\phi}\mathbf{R}\boldsymbol{\beta}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right] \\ &\quad + \frac{1}{\phi}\mathbb{E}\left[\left[\text{tr}(\mathbf{P}_{-1/2}) - 2\lambda_{max}(\mathbf{P}_{-1/2})\right]\mathbf{R}\boldsymbol{\beta}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right] + o_p(T^{-2}) = \mathbf{0} + o_p(T^{-2}), \end{aligned} \quad (\text{A.24})$$

where the last equality holds using equation (A.9) and noting that the terms on the right hand side of the inequality are odd functions of the error term which has a normal distribution. Also, the inequality above holds because \mathbf{P} , and $\mathbf{P}_{-1/2}$ are symmetric, so we have

$$\lambda_{max}(\widehat{\mathbf{P}}) \leq \lambda_{max}(\mathbf{P}) + \lambda_{max}(\mathbf{P}_{-1/2}) + o_p(T^{-1/2}), \quad (\text{A.25})$$

see [Abadir and Magnus, 2005](#), pp. 344.

Therefore, the result in equation (4.10) follows. ■

Proof. Corollary 1.2:

Using the results in the first equality of equation (A.18) when $\mathbf{W} = \mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}$, up to order

$O(T^{-1})$, we have

$$\text{MSFE}(\widehat{\boldsymbol{\beta}}_A) = \text{MSFE}(\widehat{\boldsymbol{\beta}}) + \frac{\tau}{\phi} \left[\tau - 2 \left[(m-1)k - 2 \right] \right], \quad (\text{A.26})$$

where the use has been made of equation (A.15), and noting that $\text{tr}(\mathbf{R}) = (m-1)k$. ■