No More Limited Mobility Bias: Exploring the Heterogeneity of Labor Markets*

Miren Azkarate-Askasua† Miguel Zerecero‡

June 6, 2024

Abstract

We propose a bootstrap method for correcting the small-sample bias of variance components that accommodates general heteroskedasticity and serial correlation of the errors. Our approach is suited to correct variance decompositions and the bias of multiple quadratic forms of the same linear model without increasing the computational cost. We show with Monte Carlo simulations that our bootstrap procedure is effective in correcting the bias and find that is faster than other methods in the literature. Using administrative data for France, we correct variance decompositions per labor markets defined as commuting zone and occupation combinations. We find that the correlation between worker and firm effects is increasing in commuting zone population and the slope is stable to the corrections.

JEL Codes: C13, C23, C55, J30, J31

Keywords: Limited mobility bias, bias correction, variance components, fixed effects.

*This paper previously circulated as *Correcting Small Sample Bias in Linear Models with Many Covariates*. We thank Christian Hellwig for his guidance throughout this project. We thank Fabrice Collard, Thomas Crossley, Patrick Fève, Simen Gaure, Silvia Goncalves, Cristina Gualdani, Koen Jochmans, Elia Lapenta, Tim Lee, Thierry Magnac, Nour Meddahi, Jean-Marc Robin and Uta Schönberg for helpful comments. We gratefully acknowledge financial support from TSE to get access to the data. This work is supported by a public grant overseen by the French National Research Agency (ANR) as part of the 'Investissements d'avenir' program (reference: ANR-10-EQPX-17 – Centre d'accès sécurisé aux données – CASD). Zerecero acknowledges the financial support from CONACYT (reference: 329103/383874). Azkarate-Askasua acknowledges the support from the German Research Foundation (through the CRC-TR-224 project B06). All errors are our own. First version: May 2018.

†Azkarate-Askasua: University of Mannheim (azkarate-askasua@uni-mannheim.de)

‡Zerecero: University of California, Irvine (mzerecer@uci.edu).

1 Introduction

The model of log wages introduced by Abowd, Kramarz, and Margolis (1999), AKM from now on, has been very influential in the way labor economists think about wage determinants. The most basic version of the AKM model is:

$$\log w_{it} = \theta_i + \psi_{J(i,t)} + \varepsilon_{it} \tag{1}$$

where θ_i is worker i's fixed effect, $J(i,t)$ is a function that maps where worker i is employed in period t, $\psi_{J(i,t)}$ is the firm $J(i,t)$ fixed effect, and ε_{it} is a residual.

Suppose the model is well specified and the standard exclusion restriction holds. Researches have been interested in understanding how wage inequality is explained by firm-specific wage premiums. Even if the fixed effects are unbiased, quadratic objects in the estimated parameters such as the elements of a variance decomposition are biased (Andrews, Gill, Schank, and Upward, 2008). In the AKM context, Abowd, Kramarz, Lengermann, and Pérez-Duarte (2004) dubbed the bias of these quadratic objects as *limited mobility bias* as having few movers identifying the firm wage premiums leads to noisier estimates and to the bias of variance components. Using data for different countries, Bonhomme, Holzheu, Lamadon, Manresa, Mogstad, and Setzler (2023) show it has been shown that the limited mobility bias is systematically large, and it can change the economic interpretation of the results.

Andrews et al. (2008) derive formulas for correcting the bias when the errors are homoscedastic. Gaure (2014) provides formulas for more general variance structures. Unfortunately, the direct implementation of these corrections in high dimensional models is infeasible. The reason is that the corrections entail computing the inverse of an impractically large matrix, which has prevented the direct application of the correction formulas.[1]

In this paper, we propose a bootstrap method to correct for limited mobility bias that is computationally feasible. Compared to other methods in the literature that correct for this bias (Gaure, 2014; Kline, Saggio, and Sølvsten, 2020), the main advantage of our bootstrap method is

[1]Some examples of papers doing a variance decomposition of log wages into worker and firm fixed effects without correcting for limited mobility bias are: Sorkin (2018), Card, Cardoso, Heining, and Kline (2018), Alvarez, Benguria, Engbom, and Moser (2018) (who focus on changes over time and assume the bias is constant), Song, Price, Guvenen, Bloom, and Von Wachter (2019), Leknes, Rattsø, and Stokke (2022), Arellano-Bover and San (2023), and Helm, Kügler, and Schönberg (2023), among others.

that it allows the computation of many corrections without increasing the computational cost. Besides being scalable in the number of corrections, our method is easy to implement, fast, and it accommodates different estimates of the covariance matrix of the errors, including the leave-one-out and leave-cluster-out estimates used by Kline, Saggio, and Sølvsten (2020), KSS from now on.

To illustrate the advantages of our method, consider a researcher who is interested in understanding how much the different components of an AKM model explain the variance of log wages for different subgroups of the population. This can be done, for example, by estimating separate variance decompositions for workers by race and gender (Gerard, Lagos, Severnini, and Card, 2021), or by city (Dauth, Findeisen, Moretti, and Suedekum, 2022). The computational cost of correcting for the variance components with alternative methods scales linearly with the number of subgroups. The increasing cost has prevented researchers from analyzing variance components at increasingly finer partitions of the data.[2] Our method overcomes this limitation. The computational cost of doing an arbitrary number of corrections with our method is practically the same cost of doing one correction.

We apply our method to French administrative data where we study the sorting patterns of workers to firms in different labor markets. In this case, we define a labor market as the intersection of an occupation and a commuting zone. We then study a long hypothesized question in urban and labor economics: if larger labor markets have better sorting. Given this definition, we have over 8000 labor markets. Our bootstrap method allows us to do the corrections for each labor market while using the entire sample to estimate the model parameters. Previous studies instead make separate estimations and corrections for the different subgroups, so they lose the information of workers who move across subgroups.[3]

We find that sorting is stronger in larger locations as the correlation between worker and firm fixed effects is increasing in commuting zone population. We find that this positive slope is not driven by the downward bias of the correlation due to the limited mobility bias. Instead, we find that the positive slope is stable to the bias correction as it mainly increases all the

[2]One could think of conditioning for different occupations, industries, commuting zones, education groups etc. in the AKM model.

[3]For example, Dauth et al. (2022) estimate an AKM model for each city in Germany and do a correction for each city. By doing this they lose all the information of workers who move across cities, as Leknes et al. (2022) note.

correlations at the same time.

The paper is organized as follows. Section 2, derives the bias. Section 3 presents the bootstrap correction. Section 4 discusses the practical considerations when using unbiased estimates of the errors covariance matrix. Section 5 compares our method with the one developed by KSS. Section 6 presents the application with the French data. Finally, Section 7 concludes.

2 The bias

Suppose we have some data $(\mathbf{y}, \mathbf{X})$ where $\mathbf{y}$ is an $n \times 1$ vector and $\mathbf{X}$ is a matrix of covariates of dimension $n \times k$. Consider the linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbb{E}(\boldsymbol{\varepsilon} \,|\mathbf{X}) = \mathbf{0}$. We are interested in estimating the quadratic form $\varphi = \boldsymbol{\beta}^T \mathbf{A} \boldsymbol{\beta}$ for some known matrix $\mathbf{A}$ of dimensions $k \times k$, where $\mathbb{E}(\mathbf{A} \,|\, \mathbf{X}) = \mathbf{A}$.

Let $\widehat{\boldsymbol{\beta}}$ be the OLS estimate of $\boldsymbol{\beta}$. We can now define an estimate of φ.

Definition 1 (Plug-in Estimate)**.** *The plug-in estimate of the quadratic estimate is:*

$$\widehat{\varphi}_{PI} = \widehat{\boldsymbol{\beta}}^T \mathbf{A} \widehat{\boldsymbol{\beta}}.$$

Taking the conditional expectation over the plug-in estimate, we get:

$$\begin{aligned}
\mathbb{E}\left(\widehat{\boldsymbol{\beta}}^T \mathbf{A} \widehat{\boldsymbol{\beta}} \,\middle|\, \mathbf{X}\right) &= \mathbb{E}\left(\widehat{\boldsymbol{\beta}}^T \,\middle|\, \mathbf{X}\right) \mathbf{A} \mathbb{E}\left(\widehat{\boldsymbol{\beta}} \,\middle|\, \mathbf{X}\right) + \mathbf{tr}\left(\mathbf{A} \mathbb{V}\left(\widehat{\boldsymbol{\beta}} \,\middle|\, \mathbf{X}\right)\right) \\
&= \varphi + \mathbf{tr}\left(\mathbf{S}_{\mathbf{X}}^T \mathbf{A} \mathbf{S}_{\mathbf{X}} \mathbb{V}(\boldsymbol{\varepsilon} \,|\, \mathbf{X})\right),
\end{aligned}$$

where $\mathbf{S}_{\mathbf{X}} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T$.

Definition 2 (Bias)**.** *The bias of the quadratic form $\widehat{\boldsymbol{\beta}}^T \mathbf{A} \widehat{\boldsymbol{\beta}}$ is:*

$$\delta \equiv \mathbf{tr}\left(\mathbf{S}_{\mathbf{X}}^T \mathbf{A} \mathbf{S}_{\mathbf{X}} \mathbb{V}(\boldsymbol{\varepsilon} \,|\, \mathbf{X})\right). \tag{2}$$

Computing δ is infeasible as we do not know $\mathbb{V}(\boldsymbol{\varepsilon} \,|\, \mathbf{X})$. Therefore, let $\widehat{\mathbf{V}}$ be an estimate of the

covariance matrix $\mathbb{V}(\boldsymbol{\varepsilon} \mid \mathbf{X})$. We can now define a bias correction and a bias corrected estimate of the quadratic form.

Definition 3 (Direct bias correction)**.** *Using the covariance estimator* $\widehat{\mathbf{V}}$*, the direct bias correction of* $\widehat{\boldsymbol{\beta}}^T \mathbf{A} \widehat{\boldsymbol{\beta}}$ *is equal to:*

$$\widehat{\delta}_D \equiv \mathbf{tr}\left(\mathbf{S}_{\mathbf{X}}^T \mathbf{A} \mathbf{S}_{\mathbf{X}} \widehat{\mathbf{V}}\right). \tag{3}$$

Definition 4 (Bias corrected estimate)**.** *Given the direct bias correction* $\widehat{\delta}_D$*, then the bias corrected estimate of the quadratic form is*

$$\widehat{\varphi} = \widehat{\varphi}_{PI} - \widehat{\delta}_D.$$

Given the linearity of the trace and expectation operators, we get the next proposition.

Proposition 1 (Unbiasedness of $\widehat{\delta}_D$)**.** *The direct bias correction* $\widehat{\delta}_D$ *is an unbiased estimate of the bias if and only if* $\mathbb{E}\left(\widehat{\mathbf{V}} \,\middle|\, \mathbf{X}\right) = \mathbb{V}(\boldsymbol{\varepsilon} \mid \mathbf{X})$.

All proofs are in the Appendix. Given the previous proposition, the next result follows immediately.

Corollary 1 (Unbiasedness of $\widehat{\varphi}$)**.** *The direct bias correction* $\widehat{\varphi}$ *is an unbiased estimate of* φ *if and only if* $\mathbb{E}\left(\widehat{\mathbf{V}} \,\middle|\, \mathbf{X}\right) = \mathbb{V}(\boldsymbol{\varepsilon} \mid \mathbf{X})$.

For the case without conditioning on $\mathbf{X}$, KSS show conditions for the consistency of the bias corrected estimate $\widehat{\varphi}$ with diagonal covariance matrix estimates.[4]

3 Bootstrap correction

The computation of the direct bias correction $\widehat{\delta}_D$ is unfeasible in typical applications with millions of fixed effect, which prevents us from finding the inverse of $\mathbf{X}^T\mathbf{X}$. To overcome this limitation, we propose to estimate $\widehat{\delta}_D$ using a bootstrap where, by replicating the bias structure of the plug-in estimates, we can do it in a computationally feasible way.

[4]See Assumption 1 and Lemma 3 in their paper.

To motivate the use of our bootstrap, first note that bias δ is *flat*: it does not depend on the values of the true parameters $\boldsymbol{\beta}$. Thus, we can replicate the bias without paying attention to the value $\boldsymbol{\beta}$.

Let v^* be a random vector where $\mathbb{E}(v^* \mid \mathbf{X}) = \mathbf{0}$ and $\mathbb{V}(v^* \mid \mathbf{X}) = \widehat{\mathbf{V}}$. Let $\widehat{\boldsymbol{\beta}^*}$ be the OLS estimate of regressing v^* on $\mathbf{X}$. Then, the following proposition is the first step to motivate the bootstrap correction.

Proposition 2 (Equivalence to $\widehat{\delta}_D$)**.** *The conditional expectation on the quadratic form using $\widehat{\boldsymbol{\beta}^*}$ is equal to the direct bias correction:*

$$\mathbb{E}\left(\widehat{\boldsymbol{\beta}^*}^T \mathbf{A} \widehat{\boldsymbol{\beta}^*} \,\middle|\, \mathbf{X}\right) = \mathbf{tr}\left(\mathbf{S}_{\mathbf{X}}^T \mathbf{A} \mathbf{S}_{\mathbf{X}} \widehat{\mathbf{V}}\right) = \widehat{\delta}_D.$$

The previous proposition already suggests what to do: bootstrap v^* a number of times and get an estimate of $\mathbb{E}\left(\widehat{\boldsymbol{\beta}^*}^T \mathbf{A} \widehat{\boldsymbol{\beta}^*} \,\middle|\, \mathbf{X}\right)$ using a sample average.

We need to make sure that the covariance matrix of the bootstrapped errors is equal to $\widehat{\mathbf{V}}$. In practice, this means, first, to simulate a random vector $\boldsymbol{r}$ with independent entries with mean zero and unit variance, and find a matrix $\mathbf{B}$ such that:

$$\mathbb{V}(\mathbf{B}\boldsymbol{r} \mid \mathbf{X}) = \mathbf{B}\mathbb{V}(\boldsymbol{r})\mathbf{B}^T = \mathbf{B}\mathbf{B}^T = \widehat{\mathbf{V}}.$$

A popular choice to simulate vector $\boldsymbol{r}$ is to use the Rademacher distribution: each observation can be 1 or -1, each with probability $1/2$. With $\mathbf{B}$ in hand, the next step is to get the vector $\mathbf{B}\boldsymbol{r}$ a number of times, and for each time compute the quadratic form $\widehat{\boldsymbol{\beta}^*}^T \mathbf{A} \widehat{\boldsymbol{\beta}^*}$. Finally, we only need to take the sample average over the sequence of estimated quadratic forms to get an estimate of the direct bias correction $\widehat{\delta}_D$.

Choosing $\mathbf{B}$ is easy when $\widehat{\mathbf{V}}$ is positive semi-definite. For example, when $\widehat{\mathbf{V}}$ is diagonal with non-negative entries. Then, $\mathbf{B}$ is just a diagonal matrix with entries equal to the square root of the entries of $\widehat{\mathbf{V}}$. When $\widehat{\mathbf{V}}$ is not diagonal but still positive semi-definite, a common choice to find $\mathbf{B}$ is to use the Cholesky decomposition, popular in the VAR literature.

However, we do not want to restrict ourselves to positive semi-definite estimates of the covariance matrix. Proposition 1 already imposes restrictions on the covariance estimator $\widehat{\mathbf{V}}$ to get a good estimate of the bias: $\widehat{\mathbf{V}}$ should be an unbiased estimate of $\mathbb{V}(\boldsymbol{\varepsilon} \mid \mathbf{X})$.

Jochmans (2018), Kline et al. (2020), and Anatolyev (2021) propose unbiased covariance matrix estimators that are robust to heteroskedasticity, but are not positive semi-definite. A random vector with a non-positive semi-definite covariance matrix would contain complex numbers, which complicates the application of the bootstrap. However, we can bypass this complication by noting that we can decompose any real symmetric matrix as the difference of two real positive semi-definite matrices. To see this, assume $\widehat{\mathbf{V}}$ is symmetric but possibly not positive semi-definite. Using the spectral decomposition of a real symmetric matrix, we get:

$$\widehat{\mathbf{V}} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T,$$

where the matrix $\mathbf{\Lambda}$ is a diagonal matrix containing the eigenvalues of $\widehat{\mathbf{V}}$, with the ith diagonal term equal to λ_i. We can further decompose $\mathbf{\Lambda}$ as

$$\mathbf{\Lambda} = \mathbf{\Lambda}_+ - \mathbf{\Lambda}_-,$$

where the ith diagonal terms of $\mathbf{\Lambda}_+$ and $\mathbf{\Lambda}_-$, denoted $\lambda_{+,i}$ and $\lambda_{-,i}$, are equal to:

$$\lambda_{+,i} = \begin{cases} \lambda_i, & \text{if } \lambda_i \geq 0 \\ 0, & \text{otherwise,} \end{cases} \qquad \lambda_{-,i} = \begin{cases} |\lambda_i|, & \text{if } \lambda_i < 0 \\ 0, & \text{otherwise.} \end{cases}$$

This means that $\widehat{\mathbf{V}}$ is equal to:

$$\widehat{\mathbf{V}} = \mathbf{Q}\left(\mathbf{\Lambda}_+ - \mathbf{\Lambda}_-\right)\mathbf{Q}^T = \underbrace{\mathbf{Q}\mathbf{\Lambda}_+\mathbf{Q}^T}_{\widehat{\mathbf{V}}_+} - \underbrace{\mathbf{Q}\mathbf{\Lambda}_-\mathbf{Q}^T}_{\widehat{\mathbf{V}}_-}, \tag{4}$$

where $\widehat{\mathbf{V}}_+$ and $\widehat{\mathbf{V}}_-$ are positive semi-definite. The decomposition of $\widehat{\mathbf{V}}$ means that we can rewrite the direct bias correction as:

$$\widehat{\delta}_D = \mathbf{tr}\left(\mathbf{S}_\mathbf{X}^T\mathbf{A}\mathbf{S}_\mathbf{X}\widehat{\mathbf{V}}_+\right) - \mathbf{tr}\left(\mathbf{S}_\mathbf{X}^T\mathbf{A}\mathbf{S}_\mathbf{X}\widehat{\mathbf{V}}_-\right).$$

Each of these trace terms can be represented as the expectations of some quadratic form. To see

this, let us define the following two random vectors:

$$\boldsymbol{v}_+^* \equiv \underbrace{\mathbf{Q}\left(\boldsymbol{\Lambda}_+\right)^{1/2}}_{\mathbf{B}_+}\boldsymbol{r}, \text{ and } \boldsymbol{v}_-^* \equiv \underbrace{\mathbf{Q}\left(\boldsymbol{\Lambda}_-\right)^{1/2}}_{\mathbf{B}_-}\boldsymbol{r},$$

which leads to the next proposition.

Proposition 3 (Decomposition of δ_D)**.** *Let $\widehat{\boldsymbol{\beta}^*}_+$ and $\widehat{\boldsymbol{\beta}^*}_-$ be the OLS estimates of regressing $\boldsymbol{v}_+^*$ and $\boldsymbol{v}_-^*$ on $\mathbf{X}$. Then,*

$$\widehat{\delta}_D = \mathbb{E}\left(\widehat{\boldsymbol{\beta}^*}_+^T \mathbf{A}\widehat{\boldsymbol{\beta}^*}_+ \,\middle|\, \mathbf{X}\right) - \mathbb{E}\left(\widehat{\boldsymbol{\beta}^*}_-^T \mathbf{A}\widehat{\boldsymbol{\beta}^*}_- \,\middle|\, \mathbf{X}\right).$$

The last proposition motivates the following bootstrap estimator for any covariance matrix estimate, positive semi-definite or not.

Definition 5 (Bootstrap Bias Correction)**.** *Let $\boldsymbol{v}_+^*(j)$ and $\boldsymbol{v}_-^*(j)$ as the jth simulations of vectors $\boldsymbol{v}_+^*$ and $\boldsymbol{v}_-^*$, where $j = 1 \dots J$. Also, let $\widehat{\boldsymbol{\beta}^*}_+(j)$ and $\widehat{\boldsymbol{\beta}^*}_-(j)$ be the OLS estimates of regressing $\boldsymbol{v}_+^*(j)$ and $\boldsymbol{v}_-^*(j)$ on $\mathbf{X}$. Then, the bootstrap bias correction is defined as:*

$$\delta^* = \frac{1}{J}\sum_{j=1}^{J} \widehat{\boldsymbol{\beta}^*}_+(j)^T \mathbf{A}\widehat{\boldsymbol{\beta}^*}_+(j) - \frac{1}{J}\sum_{j=1}^{J} \widehat{\boldsymbol{\beta}^*}_-(j)^T \mathbf{A}\widehat{\boldsymbol{\beta}^*}_-(j).$$

The simple linear form of the bootstrap correction leads to the following result.

Proposition 4 (Unbiasedness and Consistency of δ^*)**.** *The bootstrap bias correction δ^* is a consistent and unbiased estimate of the direct bias correction $\widehat{\delta}_D$.*

The last proposition means that we can estimate the direct bias correction to arbitrary precision, and implies the following result

Corollary 2. *The bootstrap bias correction δ^* is an unbiased estimate of the bias δ if and only if $\mathbb{E}\left(\widehat{\mathbf{V}} \,\middle|\, \mathbf{X}\right) = \mathbb{V}\left(\boldsymbol{\varepsilon} \,|\, \mathbf{X}\right)$.*

The main computational cost of our method is to estimate $\widehat{\boldsymbol{\beta}^*}_+$ and $\widehat{\boldsymbol{\beta}^*}_-$, not the number of quadratic forms to correct. In other words, if we would like to estimate a bias corrections for the set of quadratic forms $\{\widehat{\boldsymbol{\beta}}^T \mathbf{A}_m \widehat{\boldsymbol{\beta}}\}$ for $m = 1 \dots M$, we just need to calculate the bootstrap analogous quadratic forms; a step with negligible computational cost.

To clarify this computational advantage and to summarize our bootstrap method, we present below a 'high-level' algorithm to do corrections for an arbitrary number of quadratic forms, provided we have a covariance matrix estimate $\widehat{\mathbf{V}}$.

Algorithm 1 Bootstrap Bias Correction

1: Let $\widehat{\mathbf{V}}$ be the covariance matrix estimate.
2: Using the spectral decomposition of $\widehat{\mathbf{V}}$ get $\mathbf{Q}$ and $\mathbf{\Lambda}$ such that $\widehat{\mathbf{V}} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$.
3: Decompose $\mathbf{\Lambda} = \mathbf{\Lambda}_+ - \mathbf{\Lambda}_-$, with $\mathbf{\Lambda}_+$ having the positive eigenvalues and $\mathbf{\Lambda}_-$ the absolute value of the negative eigenvalues.
4: Get $\mathbf{B}_+ = \mathbf{Q}\left(\mathbf{\Lambda}_+\right)^{1/2}$ and $\mathbf{B}_- = \mathbf{Q}\left(\mathbf{\Lambda}_-\right)^{1/2}$.
5: **for** $j = 1, \ldots, J$ **do**
6: Simulate a vector $\boldsymbol{r}$ of length n of independent Rademacher entries.
7: $\boldsymbol{v}_+^* = \mathbf{B}_+\boldsymbol{r}$, $\boldsymbol{v}_-^* = \mathbf{B}_-\boldsymbol{r}$.
8: Get $\widehat{\boldsymbol{\beta}^*}_+$ and $\widehat{\boldsymbol{\beta}^*}_-$ by solving:

$$\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}^*}_+ = \mathbf{X}^T\boldsymbol{v}_+^* \quad \text{and} \quad \mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}^*}_+ = \mathbf{X}^T\boldsymbol{v}_+^*.$$

9: Compute $\delta_m^{*(j)} = \left(\widehat{\boldsymbol{\beta}^*}_+^T\mathbf{A}_m\widehat{\boldsymbol{\beta}^*}_+\right) - \left(\widehat{\boldsymbol{\beta}^*}_-^T\mathbf{A}_m\widehat{\boldsymbol{\beta}^*}_-\right)$ for all $m = 1 \ldots M$.
10: **end for**
11: Compute $\delta_m^* = \frac{1}{J}\sum_{j=1}^J \delta_m^{*(j)}$ for all $m = 1 \ldots M$.

Step 8 of the algorithm shows the main computational cost of the algorithm: solving the normal equations. As mentioned before, this could be done by just running a regression of $\boldsymbol{v}_+^*$ and $\boldsymbol{v}_-^*$ on $\mathbf{X}$. This is a huge advantage of our method as it relies in common algorithms that estimate linear regressions with a large number of fixed effects. There are many of these algorithms in different software programs, so the implementation cost of our method is relatively low.[5]

Advantages of the bootstrap bias correction: We enumerate briefly the main advantages of our bootstrap estimator; we explain with more detail afterwards. In short, our bootstrap estimator is:

1. *General*: can use any real symmetric covariance matrix estimator.

[5]Some popular choices are the package *fixest* in R Bergé (2018), or *reghdfe* Correia (2017) in STATA. For our applications and simulations where we run AKM models, we follow KSS and use the preconditioned conjugate gradient method in Matlab with a preconditioner developed by Koutis, Miller, and Tolliver (2011) that is optimized for two-way fixed effects regressions.

2. *Scalable*: can compute corrections for different quadratic forms at the same time without increasing the computational cost.

3. *Flexible*: can do the correction of any quadratic form; no need to create complicated ad-hoc code for different corrections.

4. *Easy to implement*: only relies on the estimation of least square regressions.

The spectral decomposition argument above explains why the bootstrap correction is *general*: we can use it with any real symmetric covariance matrix estimate.

The bootstrap method is *scalable* to any number of corrections. Like we mentioned above, the main cost of our method is to solve for the normal equations for every iteration in the bootstrap. At the end of the iteration we need to compute the quadratic forms. In practice, the cost of computing an additional quadratic form is negligible compared to the cost of running the regression: once we pay the fixed cost of running the regression, computing more quadratic forms comes at almost not cost. In other words, there are increasing returns to the number of corrections. This opens the door to many more applications of interest that were prohibitively costly before. For example, in the AKM context, one could do corrections for different subsamples of the data, and explore how the moments change across different periods, occupations, genders, locations, etcetera.

The bootstrap correction is *flexible*: KSS's method requires the computation of an appropriate $\mathbf{A}_m$ matrix for each correction. Our method can compute the outcome of the quadratic forms without explicitly declaring $\mathbf{A}_m$. For example, besides correcting for the covariance of workers and firms fixed effects, one could correct for other moments that reflect labor market sorting, like the correlation between the worker fixed effect and the average fixed effect of the coworkers, as proposed by Lopes de Melo (2018).

The method is *easy to implement*: the bootstrap mostly relies on running least square regressions. Our method can take advantage from the continuous development of tools that increase the estimation speed of high dimensional linear models. Even more, it is easy to adapt the method to use Generalized Least Squares instead of OLS; for example, it is straightforward to adapt the bootstrap to use Weighted Least Squares.

Efficiency gains compared to alternative bootstraps: Using the bootstrap to correct for biases is ubiquitous in the literature. MacKinnon and Smith Jr (1998) (MS, henceforth) propose a similar bootstrap to correct for flat biases like the one considered here.[6] For simplicity, let us abstract about the decomposition of the variance estimate as the difference of two positive semi-definite matrices, but all arguments follow easily in that case. In other words, let us have that $\widehat{\mathbf{V}} = \widehat{\mathbf{V}}_+$. MS propose building the bootstrapped dependent variable by using the original estimate of $\boldsymbol{\beta}$, $\boldsymbol{y}^* = \mathbf{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{v}^*$, and use these new data $(\mathbf{X}, \boldsymbol{y}^*)$ to estimate $\widehat{\boldsymbol{\beta}}^*_{MS}$. Then, to compute the quadratic objects $\widehat{\boldsymbol{\beta}}^*_{MS}(j)^T \mathbf{A} \widehat{\boldsymbol{\beta}}^*_{MS}(j)$ for each bootstrap j and use them to calculate a bias correction of the form:

$$\delta^*_{MS} = \frac{1}{p}\sum_{j=1}^{p} \widehat{\boldsymbol{\beta}}^*_{MS}(j)^T \mathbf{A} \widehat{\boldsymbol{\beta}}^*_{MS}(j) - \widehat{\boldsymbol{\beta}}^T \mathbf{A} \widehat{\boldsymbol{\beta}}.$$

MS already note that one can estimate a flat bias correction by using any $\widehat{\boldsymbol{\beta}}$ to generate $\boldsymbol{y}^*$. In our bootstrap method we use $\widehat{\boldsymbol{\beta}} = \mathbf{0}$. As shown by the proposition below, this choice has some benefits in terms of the efficiency of the estimator.

Proposition 5 (Efficiency Gains)**.** *Let $\boldsymbol{v}^*$ be a vector of independent random variables with $\mathbb{E}(\boldsymbol{v}^* | \mathbf{X}, \boldsymbol{\varepsilon}) = \mathbf{0}$, $\mathbb{E}\left((\boldsymbol{v}^*)^2 \middle| \mathbf{X}, \boldsymbol{\varepsilon}\right) < \infty$, and $\mathbb{E}\left((\boldsymbol{v}^*)^3 \middle| \mathbf{X}, \boldsymbol{\varepsilon}\right) = \mathbf{0}$. Then, $\mathbb{V}\left(\delta^*_{MS} \middle| \mathbf{X}\right) \geq \mathbb{V}(\delta^* | \mathbf{X})$.*

Given that we use independent Rademacher entries $\boldsymbol{r}$ to form $\boldsymbol{v}^* = \mathbf{B}\boldsymbol{r}$, then the conditions $\mathbb{E}(\boldsymbol{v}^* | \mathbf{X}) = \mathbf{0}$ and $\mathbb{E}\left((\boldsymbol{v}^*)^3 \middle| \mathbf{X}\right) = \mathbf{0}$ are satisfied. The proposition tell us that choosing $\widehat{\boldsymbol{\beta}} = \mathbf{0}$ to form the bootstrapped dependent variable reduces the variance of the bias correction. Furthermore, if the estimate for the variance is unbiased, this means that our bootstrap estimate is more efficient than the more traditional one as proposed by MS.

3.1 Computation of $\mathbf{B}_+$ and $\mathbf{B}_-$: common examples

The bootstrap estimator owes its simple form to two properties: (i) the decomposition of the covariance matrix as the difference of two positive semi-definite matrices, and (ii) that the bias is a linear function of the covariance matrix. These two properties allow us to decompose the original bias, which is equal to a trace, as the difference of two traces.

[6] As stated before, flat bias is one that does not depend on the levels of the original estimates. The bias from the quadratic forms is flat because the trace term in (2) is independent of $\boldsymbol{\beta}$.

Below we present three different examples of unbiased estimates of $\mathbf{V}$ for different assumptions on the error term, and discuss their corresponding $\mathbf{B}_+$ and $\mathbf{B}_-$ matrices.

Example 1 —Homoscedastic Errors: Consider the following covariance matrix estimate:

$$\widehat{\mathbf{V}} = \widehat{\sigma}\mathbf{I}, \quad \widehat{\sigma} = \frac{1}{n-k}\sum_{i}^{n}\widehat{\boldsymbol{\varepsilon}}_i^2,$$

where $\widehat{\boldsymbol{\varepsilon}}_i = \boldsymbol{y}_i - \widehat{\boldsymbol{y}}_i$ is the OLS residual for the ith observation and $\mathbf{I}$ is the identity matrix. When the errors are homoscedastic, this covariance estimate is an unbiased estimate of the covariance matrix of the unobserved errors.

This covariance matrix estimate is a positive semi-definite matrix: it has only non-negative eigenvalues, meaning $\boldsymbol{\Lambda}_- = \mathbf{0}$. Also, it is a diagonal matrix, so following the decomposition above we then have that $\mathbf{Q} = \mathbf{I}$, and $\boldsymbol{\Lambda}_+ = \widehat{\mathbf{V}}$ leading to $\mathbf{B}_+ = (\boldsymbol{\Lambda}_+)^{1/2}$ and $\mathbf{B}_- = \mathbf{0}$.

Example 2 —Leave-one-out covariance estimate: KSS use a diagonal covariance matrix estimate which is unbiased when the errors are heteroskedastic. The diagonal entries are:

$$\widehat{\mathbf{V}}_{ii} = \frac{\boldsymbol{y}_i\widehat{\boldsymbol{\varepsilon}}_i}{1-P_{ii}}, \tag{5}$$

where P_{ii} is the leverage of observation i, defined as the ith diagonal of the projection matrix $\mathbf{P} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$.

$\widehat{\mathbf{V}}$ is a diagonal matrix but not necessarily positive semi-definite. According to the spectral decomposition we have that $\mathbf{Q} = \mathbf{I}$ so $\widehat{\mathbf{V}} = \boldsymbol{\Lambda}_+ - \boldsymbol{\Lambda}_-$, where $\boldsymbol{\Lambda}_+$ contains the positive diagonal entries of $\widehat{\mathbf{V}}$ and $\boldsymbol{\Lambda}_-$ the negative entries. We therefore have $\mathbf{B}_+ = (\boldsymbol{\Lambda}_+)^{1/2}$ and $\mathbf{B}_- = (\boldsymbol{\Lambda}_-)^{1/2}$.

Example 3 —Leave-cluster-out covariance estimate: This is a generalization of the leave-one-out covariance matrix estimate. It was also proposed by KSS and studied in more detail by Anatolyev (2021).

We introduce some notation to ease the exposition below. Assume we can divide the data $(\boldsymbol{y}, \mathbf{X})$ into G mutually exclusive clusters, where the gth cluster has n_g observations. This means that $n = \sum_{g=1}^{G} n_g$. Define as $\mathbf{X}_g$ a matrix of covariates for cluster g of dimension $n_g \times k$. Similarly, define $\boldsymbol{y}_g$ and $\boldsymbol{\varepsilon}_g$ as vectors of dimension n_g.

Define as $\mathbf{P}_{gg}$ the principal minor of the projection matrix $\mathbf{P}$ where we keep the observations that correspond to cluster g. If the data is rearranged such that all observations within a cluster are adjacent, then $\mathbf{P}_{gg}$ would be the gth diagonal block of $\mathbf{P}$. Without loss of generality, we will assume the data is ordered that way.

In a similar way, define $\mathbf{M}_{gg} = \mathbf{I}_{n_g} - \mathbf{P}_{gg}$, where $\mathbf{I}_{n_g}$ is the identity matrix of dimension $n_g \times n_g$. We can now define the analogous of the leave-one-out residuals $\widehat{\varepsilon}_i/(1-P_{ii})$ but for clusters instead of an observation. Following Anatolyev (2021), the leave-cluster-out residual is equal to:

$$\widehat{\boldsymbol{\varepsilon}}_g^{LC} = \mathbf{M}_{gg}^{-1}\widehat{\boldsymbol{\varepsilon}}_g.$$

Then, the leave-cluster-out symmetric estimate of variance for the gth diagonal block of $\widehat{\mathbf{V}}$ is:

$$\widehat{\mathbf{V}}_{gg} = \frac{1}{2}\left(\boldsymbol{y}_g\left(\widehat{\boldsymbol{\varepsilon}}_g^{LC}\right)^T + \widehat{\boldsymbol{\varepsilon}}_g^{LC}\boldsymbol{y}_g^T\right). \tag{6}$$

Clearly, the leave-cluster-out variance estimate is a generalization of the leave-one-out estimate of Example 2, which would correspond to all clusters having just a single observation. Anatolyev (2021) shows that $\widehat{\mathbf{V}}_{gg}$ is an unbiased estimate of the gth diagonal block of the covariance matrix $\mathbb{V}(\boldsymbol{\varepsilon}\,|\,\mathbf{X})$.

As the matrix $\widehat{\mathbf{V}}$ is block diagonal, we can do the spectral decomposition (4) for each diagonal block $\widehat{\mathbf{V}}_{gg}$ that correspond to the different clusters g, making the computation easier. We can therefore compute each block of $\mathbf{B}_+$ and $\mathbf{B}_-$ separately per cluster g.

4 Practical details when using unbiased covariance estimates

The previous section established the general bootstrap correction method. In this section, we discuss some practical details to implement the bootstrap method in three cases. Each of these cases uses a different unbiased estimate of the covariance matrix, suitable for different situations: the standard covariance estimate for homoscedastic errors; the leave-one-out estimator of Jochmans (2018) and KSS; and the leave-cluster-out estimate also introduced by KSS and developed in more detail by Anatolyev (2021).

Case 1 —Homoscedastic errors: This is the simplest case. The bias correction under this assumption on the errors was first proposed by Andrews et al. (2008). Gaure (2014) implements an iterative method to estimate the bias under the homoscedastic assumption, but it is not scalable like ours.

Having estimated the variance of the error terms $\widehat{\sigma}$ as explained in Example 1, we need to perform several bootstraps where we need to simulate the vector $\boldsymbol{r}$, get $\boldsymbol{v}^* = \sqrt{\widehat{\sigma}}\boldsymbol{r}$, run regressions of $\boldsymbol{v}^*$ on $\mathbf{X}$, and calculate the quadratic forms. Finally, the average across the estimated quadratic forms is the estimate of the bias.

Case 2 —Heteroscedastic errors: The leave-one-out covariance matrix estimate is a diagonal covariance matrix with entries described by equation (5) in Example 2 above. As some of these entries are negative, we need to do a decomposition and separate the negative entries from the non-negative ones to form the $\mathbf{B}_+$ and $\mathbf{B}_-$ matrices.

Using the leave-one-out covariance matrix estimator requires estimating the leverage of each observation and guaranteeing that the leverage of each observation is below 1 such that the variance estimate of observation i $\widehat{\mathbf{V}}_{ii}$ exists. We discuss their implementation below.

Case 3 —Clustered errors: The leave-cluster-out variance estimate shares the same two complications with the leave-one-out variance estimate: we need that the leave-cluster-out covariance estimate exists and we need to estimate the residual matrix $\mathbf{M} = \mathbf{I} - \mathbf{P}$.

In the following we discuss the sample selection required for the existence of the leave-one-out and leave-cluster-out estimates and iterative procedures to estimate leverages for each of the covariance matrix estimates.

4.1 Existence of variance estimates: leave-one-out and leave-cluster-out

One important practical consideration in the leave-one-out and leave-cluster out covariance estimation is that those exist. This requires to select a subsample of the connected set such that: (i) the leverages P_{ii} are below 1 for the leave-one-out estimator; and (ii) M_{gg} is non-singular Anatolyev (2021) for the leave-cluster-out estimator.

Leave-one-out connected set. In the AKM context, having leverages below 1 requires leaving out: (i) workers that appear only once in the sample: that observation completely pins down the worker fixed effect; and (ii) observations that upon removing them of the sample it would leave some firms unconnected.[7]

Therefore, using the leave-one-out variance estimate in the AKM context requires a stronger notion of connectivity than just using the connected set of firms: we need that each connected firm is not only connected by the movement of one worker observation. KSS denote this set of firms as the *leave-one-out connected set*.[8]

Leave-cluster-out connected set. To make sure that $\mathbf{M}_{gg}$ is non-singular for all the clusters, we need to compute a *leave-cluster-out* sample similar to the case above. We will focus on the clustering of errors at the match level. Firm and worker fixed effects will capture most of the correlation across unobserved components but we will allow the errors to be correlated within a worker-firm pair. This is the *leave-match-out* case considered by KSS which, again, requires a stronger notion of connectivity than the connected set.

To guarantee that $\mathbf{M}_{gg}$ is non-singular for all the matches we need to remove: (i) workers who only have one match; and (ii) matches whose removal from the sample would leave some firms unconnected.

We discuss more about how to efficiently find the problematic observations using graph theory tools in Appendix D.

4.2 Variance estimation: leave-one-out

Obtaining the leverages $P_{ii} = \mathbf{X}_i \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}_i^T$ suffer from the same computational cost as getting the direct bias correction. However, the proposition below shows that we can do another iterative procedure involving only linear regressions–akin to the bootstrap—to bypass the inversion

[7] Given the presence of both worker and firm fixed effects in the AKM model, only the difference between firm fixed effects is identified. Therefore, the identification of the relative difference of firm fixed effects for two firms requires having at least one worker moving between them. In practice, the largest connected set of firms is used when estimating AKM models.

[8] The leave-one-out connected set is a subset of the connected as it requires that more than one worker observation is connecting two firms.

of $\mathbf{X}^T\mathbf{X}$ and get an estimate of the leverages.

Proposition 6 (Leverage approximation)**.** *Let $\boldsymbol{r}$ be a random vector of dimension n with Rademacher entries. Also, let $\widehat{\boldsymbol{r}}$ be the fitted values after running a regression $\boldsymbol{r}$ on $\mathbf{X}$ with $\widehat{r}_i$ being the fitted value for the ith observation. Then,*

$$\mathbb{E}\left(\widehat{r}_i^2 \,\middle|\, \mathbf{X}\right) = P_{ii}, \quad \textit{and} \quad \mathbb{E}\left((r_i - \widehat{r}_i)^2 \,\middle|\, \mathbf{X}\right) = 1 - P_{ii}.$$

This means that we can simulate a random vector of Rademacher entries, calculate the square of the fitted values and of the residuals, and their sample averages give us an estimate of the leverages P_{ii} and $1 - P_{ii}$. We present the algorithm in the Appendix.

Let $\widehat{P}_{ii}$ be the estimate using the squared of the fitted values, and $\widehat{M}_{ii}$ be the estimate using the squared residuals—corresponding to the right expression on Proposition 6. To ensure our estimates are between 0 and 1, we follow Kline, Saggio, and Sølvsten (2021) and define the following estimates:

$$\overline{P}_{ii} \equiv \frac{\widehat{P}_{ii}}{\widehat{P}_{ii} + \widehat{M}_{ii}}, \quad \text{and} \quad \overline{M}_{ii} \equiv \frac{\widehat{M}_{ii}}{\widehat{P}_{ii} + \widehat{M}_{ii}}.$$

As it is clear from above, these estimates satisfy the constraint $\overline{P}_{ii} + \overline{M}_{ii} = 1$, and as both $\widehat{P}_{ii}$ and $\widehat{M}_{ii}$ are always non-negative, then both estimates are satisfied to be between 0 and 1.

With the leverages estimates in hand, we can compute the diagonal entries of the covariance matrix as $\widehat{\mathbf{V}}_{ii} = \boldsymbol{y}_i \left(\overline{M}_{ii}\right)^{-1} \widehat{\varepsilon}_i$. In the Appendix we discuss how to correct for the bias introduced by the non-linearity of $\left(\overline{M}_{ii}\right)^{-1}$.

4.3 Variance estimation: leave-cluster-out

When the number of covariates is large, we cannot explicitly compute $\mathbf{P}_{gg}$ or $\mathbf{M}_{gg}$ for the same reason as for the leave-one-out estimator. We can instead estimate them using linear regressions as suggested by the following proposition.

Proposition 7 (Approximation of Diagonal Blocks of $\mathbf{P}$ and $\mathbf{M}$)**.** *Let $\boldsymbol{r}$ be a random vector of dimension n with Rademacher entries. Also, let $\widehat{\boldsymbol{r}}$ be the fitted value after running a regression $\boldsymbol{r}$ on $\mathbf{X}$. Denote*

as $\boldsymbol{r}_g$ and $\widehat{\boldsymbol{r}}_g$ as the observations of vector $\boldsymbol{r}$ and $\widehat{\boldsymbol{r}}$, respectively, that correspond to cluster g. Then,

$$\mathbb{E}\left(\widehat{\boldsymbol{r}}_g\widehat{\boldsymbol{r}}_g^T \,\middle|\, \mathbf{X}\right) = \mathbf{P}_{gg}, \quad \textit{and} \quad \mathbb{E}\left(\left(\boldsymbol{r}_g - \widehat{\boldsymbol{r}}_g\right)\left(\boldsymbol{r}_g - \widehat{\boldsymbol{r}}_g\right)^T \,\middle|\, \mathbf{X}\right) = \mathbf{M}_{gg}.$$

Akin to Proposition 6, this tell us we can get estimates of $\mathbf{P}_{gg}$ and $\mathbf{M}_{gg}$ through an iterative procedure by taking sample averages over the matrices formed by the outer product of the fitted values and residuals of the gth cluster observations. We also describe the algorithm in the Appendix.

The requirement for the existence of the leave-cluster-out variance estimate is that $\mathbf{M}_{gg}$ is non-singular or equivalently that the largest eigenvalue of $\mathbf{P}_{gg}$ is strictly smaller than 1.

Proposition 8 (Singularity of $\widehat{\mathbf{M}}_{gg}$)**.** *If $\mathbf{M}_{gg}$ is singular then $\widehat{\mathbf{M}}_{gg}$ is singular.*

The sample selection in the leave-cluster-out leads to a non-singular $\mathbf{M}_{gg}$. In the following we describe an estimation procedure that guarantees that $\widehat{\mathbf{P}}_{gg}$ is symmetric with the largest eigenvalue strictly smaller than 1. The projection matrix $\mathbf{P}$ is idempotent, which means that the eigenvalues of $\mathbf{P}$ are either 0 or 1. As $\mathbf{P}$ is a real and symmetric matrix, its eigenvalues *interlace* the eigenvalues of its principal minor matrices.[9] This means that the eigenvalues of the block-diagonal matrix $\mathbf{P}_{gg}$ must be between 0 and 1, which implies the same for the matrix $\mathbf{M}_{gg}$.

We make sure that the estimates of $\mathbf{P}_{gg}$ and $\mathbf{M}_{gg}$ are symmetric with eigenvalues between 0 and 1 by using the following estimates:

$$\overline{\mathbf{P}}_{gg}^S = \mathbf{L}_{gg}^{-1}\widehat{\mathbf{P}}_{gg}\left(\mathbf{L}_{gg}^{-1}\right)^T, \quad \text{and} \quad \overline{\mathbf{M}}_{gg}^S = \mathbf{L}_{gg}^{-1}\widehat{\mathbf{M}}_{gg}\left(\mathbf{L}_{gg}^{-1}\right)^T,$$

where $\mathbf{L}_{gg}$ is the lower triangular Cholesky factor of $\widehat{\mathbf{P}}_{gg} + \widehat{\mathbf{M}}_{gg}$, i.e. $\mathbf{L}_{gg}\mathbf{L}_{gg}^T = \widehat{\mathbf{P}}_{gg} + \widehat{\mathbf{M}}_{gg}$. Clearly $\overline{\mathbf{P}}_{gg}^S + \overline{\mathbf{M}}_{gg}^S = \mathbf{I}_{n_g}$, and as the proposition below shows, they have eigenvalues within zero and one.

Proposition 9 (Eigenvalue properties of $\overline{\mathbf{P}}_{gg}^S$ and $\overline{\mathbf{M}}_{gg}^S$)**.** *Assume $\widehat{\mathbf{M}}_{gg}$ is non-singular. Then, the eigenvalues of $\overline{\mathbf{P}}_{gg}^S$ lie within $[0,1)$ and the eigenvalues of $\overline{\mathbf{M}}_{gg}^S$ lie within $(0,1]$.*

[9] This is known as the Eigenvalue Interlacing Theorem. A textbook treatment of the issue is found on p. 552 of Meyer (2000).

Using $\overline{\mathbf{M}}_{gg}^{S}$ we can compute the leave-cluster-out residuals $\widehat{\boldsymbol{\varepsilon}}_{g}^{LC}$, and estimate $\widehat{\mathbf{V}}_{gg}$ as shown in (6). We show in the Appendix how to get the leave-cluster-out residuals without the explicit inversion of any matrix.

5 Comparison with KSS

Both KSS and the bootstrap method rely on an iterative procedure to estimate the bias. Both our methods rely in solving linear systems several times. We both do the exactly same estimation of the leverage. The difference between our methods is what type of linear systems we are solving and what part of the trace term of the bias we are approximating.

Let $s_{ii}(\mathbf{A})$ be the ith diagonal element of matrix $\mathbf{S}_{\mathbf{X}}^{T}\mathbf{A}\mathbf{S}_{\mathbf{X}}$. With a diagonal covariance matrix estimate $\widehat{\mathbf{V}}$, we can rewrite the direct bias correction (3) as

$$\widehat{\delta}_{D} = \sum_{i} \widehat{\sigma}_{i} s_{ii}(\mathbf{A}),$$

where $\widehat{\sigma}_{i}$ is the ith diagonal element of $\widehat{\mathbf{V}}$. KSS estimate $s_{ii}(\mathbf{A})$ by using

$$\mathbb{E}\left(\left(\mathbf{X}_{i}\left(\mathbf{X}^{T}\mathbf{X}\right)^{-1}\mathbf{A}_{f}\boldsymbol{r}\right)^{2}\middle|\,\mathbf{X}\right) = s_{ii}(\mathbf{A}),$$

where $\mathbf{A}_{f}\mathbf{A}_{f}^{T} = \mathbf{A}$, and $\boldsymbol{r}$ is again an iid random vector where each entry has mean zero and unit variance. Then, they can simulate vectors $\boldsymbol{r}$ and solve the following linear system:

$$\mathbf{X}^{T}\mathbf{X}\boldsymbol{z} = \mathbf{A}_{f}\boldsymbol{r}. \tag{7}$$

With $\boldsymbol{z}$ in hand they just multiply it by $\mathbf{X}_{i}$ and square it. They do this a number of times and take the sample average to get an estimate of $s_{ii}(\mathbf{A})$.

The main computational burden of KSS's method is solving the system of equations (7) a number of times; analogous to solving the system of equations in our bootstrap. But this system is different than the one from our bootstrap: it is a function of the specific quadratic form, characterized by the matrix $\mathbf{A}$.

In some situations, like doing a variance decomposition of an AKM model with worker and

firm fixed effects, one can reuse the estimates corresponding to the correction of variance of workers fixed effects and variance of firm fixed effects to compute the correction of the covariance between worker and firm fixed effects. When doing one set of corrections, KSS needs to solve for three systems: one for the leverages, one for the workers fixed effects, and one for the firms fixed effects. If you want to do more corrections, for example for different subsamples, you would need to increase the number of systems to solve. This means KSS has to solve *at least* three systems. Our method has to solve *at most* three systems regardless of the number of corrections: one for the leverages, and two for $\mathbf{V}_+$ and $\mathbf{V}_-$.

The bootstrap method approximates the entire trace term, while KSS approximate the diagonal terms $s_{ii}(\mathbf{A})$. This is at the heart of the difference between our methods. This conceptual difference allows to scale the bootstrap method to any number of corrections. However, it also reveals an advantage of KSS over our method: it can be easily scaled to different dependent variables $\boldsymbol{y}$, as estimating $s_{ii}(\mathbf{A})$ only depends on $\mathbf{A}$ and $\mathbf{X}$. For example, Lachowska, Mas, Saggio, and Woodbury (2023) estimate AKM models with hours and wages as dependent variables. One can estimate $s_{ii}(\mathbf{A})$ once and compute the direct bias correction for the wages and hours by adjusting the variance estimates $\widehat{\sigma}_i$.

In the end, the suitability of KSS or the bootstrap method depends on the application. If there are more corrections than dependent variables then it is better to do the bootstrap. If the opposite is true, then it is better to do KSS.

5.1 Speed and accuracy across methods

We compare our method versus KSS in terms of speed and accuracy. We simulate labor market data according to the model specified in (1) and do a simple variance decomposition and their corrections. This exercise is the most beneficial for KSS as both their method and the bootstrap method have to solve only three systems of equations per iteration.[10]

We use the leave-one-out covariance matrix estimator for both methods.[11] To increase com-

[10]Recall that if we were to do more corrections for different subsamples KSS has to solve at least three systems of equations while the bootstrap method has to do at most three systems of equations. Naturally, in this case, the bootstrap method would be faster.

[11]The leave-match-out would be harder to compare as KSS take match averages of workers who move across firms, and assume that workers who stay in the same firm for the entire sample have no correlated errors within the match. This is because the parameters of those workers are not leave-match-out estimable, but they might be leave-one-out estimable. Averaging within each match would reduce the variance of the outcome variable by losing the

Table 1: Monte Carlo simulations. Heteroscedastic errors.

		Mean Squared Error (MSE$\times 10^3$)			
Model	Time	$\hat{\sigma}^2_\theta$	$\hat{\sigma}^2_\psi$	$\hat{\sigma}_{\theta,\psi}$	Average
Plug-in		20.366	6.059	5.280	10.569
Bootstrap	582.6	0.001	0.003	0.001	0.002
KSS	595.9	0.001	0.003	0.001	0.002

Notes: We simulate a labor market with a size of around 5 million observations. *Plug-in* is the plug-in estimator, *Bootstrap* implements our bootstrap method, and *KSS* is the Kline et al. (2020) method. True moments are computed at the leave-one-out connected set. In all the exercises the number of movers per firm is 3 and the average firm has 12 employees. *Time* is the computing time in seconds. $\hat{\sigma}^2_\theta$, $\hat{\sigma}^2_\psi$ and $\hat{\sigma}_{\theta,\psi}$ present respectively the mean squared errors of the corrected estimates of the variance of the worker fixed effects, variance of the firm fixed effects and the covariance between worker and firm effects. All the MSE are multiplied by 1000 due to high accuracy of the corrections. *Average* is the average MSE (also scaled).

parability, we use the same data selection method that KSS use to ensure the restriction $P_{ii} < 1$ is satisfied. Their method is more restrictive than necessary, but the resulting sample satisfies the restriction. In Appendix D we explain with more detail about less restrictive data selection procedures that satisfy the restriction. The resulting sample size is around 5 million observations per simulation, where the number of movers per firm is 3 and the average firm has 12 employees. Also, both methods use the preconditioned conjugate gradient method in Matlab to solve for the linear equations. We impose the same tolerance value for convergence and do 300 iterations for both methods.

Table 1 presents the results. As expected, both methods reduce the Mean Squared Error (MSE) compared to the plug-in estimates. Interestingly, the MSEs of both methods are identical to six digits. Also, the bootstrap method is faster by a small amount. This show that even in the case where both methods need to solve three systems of equations per iteration, our bootstrap method is as fast as KSS.

6 Sorting across French labor markets

We briefly introduce the data and present the results.

within-match variance components and would scale the importance of the remaining variance components.

6.1 Data

We follow Babet, Godechot, and Palladino (2022) to build a pseudo panel from French administrative data *DADS Base Tous Salariés - BTS*. Babet et al. (2022) provide a code that generates a panel version of the cross sectional data from *BTS* by matching the yearly identifiers across years.[12]

We focus on *BTS* from 2003 to 2019. We further restrict the sample to private sector workers that are working full-time. We only consider the main job in a year, working full time that appear at least twice in the sample and exclude observations where the number of hours was imputed or have missing information.[13] We define labor markets as commuting zone and 2-digit occupation combinations. We trim the log hourly wages yearly from the top at 0.05%. Our dependent variable is the log hourly wage residualized by age and age squared.

6.2 Results

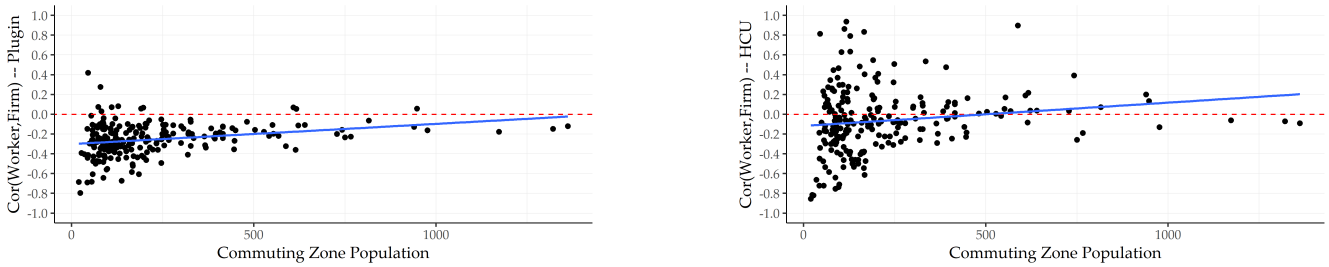
Figure 1 presents the relationship between the correlation of worker and firm fixed effects with the commuting zone population. Taking the plugin estimates on Panel (a), most of the commuting zone sorting estimates are negative and there seems to be a positive gradient between sorting and commuting zone size. Similarly, Panel (b) shows the estimates after correcting for limited mobility bias with the leave-one-out estimator of the variance of the error terms. We find that many correlations turn to be positive and the sorting-population gradient is stable.

7 Conclusion

In this paper, we propose a computationally feasible bootstrap method to correct for the small-sample bias found in all quadratic forms in the parameters of linear models with a very large number of covariates. We show using Monte Carlo simulations that the method is effective at reducing the bias. The application to French labor market data shows that the correction increases the correlation between firm and worker fixed effects. Depending on the sample and on the specification, our bias correction method changes the sign of that correlation and in all

[12] We refer the reader to Section 1 and Appendix C of their paper for additional details.
[13] We provide additional details on the sample construction in Appendix A.

Figure 1: Application. Sorting Across French Labor Markets.



Notes: These figures present on the x axis commuting zone population. On the y axis, Panel (a) shows on the estimated correlation between worker and firm fixed effects with the plug-in estimates and Panel (b) shows the corrected correlation estimated with the leave-one-out covariance matrix estimator. We remove the commuting zone with the highest population for readability of the figure.

cases it changes the relative importance of the different components in explaining the variance of log wages.

The only requirements to implement our correction is to have a bootstrap procedure that is consistent with the assumption on the variance-covariance matrix of the error term and to estimate the model several times. The correction can thus be applied easily to any study running an AKM type regression or two-way fixed effects regressions. Our method is similar in time to Kline et al. (2020) and as accurate in the simulations. The main advantage of our approach is that it allows to increase the number of moments to correct without increasing the computational costs.

References

ABOWD, J., F. KRAMARZ, P. LENGERMANN, AND S. PÉREZ-DUARTE (2004): "Are good workers employed by good firms? A test of a simple assortative matching model for France and the United States," *Unpublished Manuscript*.

ABOWD, J. M., R. H. CREECY, AND F. KRAMARZ (2002): "Computing person and firm effects using linked longitudinal employer-employee data," Tech. rep., Center for Economic Studies, US Census Bureau.

ABOWD, J. M., F. KRAMARZ, AND D. N. MARGOLIS (1999): "High wage workers and high wage firms," *Econometrica*, 67, 251–333.

ALVAREZ, J., F. BENGURIA, N. ENGBOM, AND C. MOSER (2018): "Firms and the decline in earnings inequality in Brazil," *American Economic Journal: Macroeconomics*, 10, 149–189.

ANATOLYEV, S. (2021): "Leave-cluster-out and variance estimation," Tech. rep.

ANDREWS, M. J., L. GILL, T. SCHANK, AND R. UPWARD (2008): "High wage workers and low wage firms: negative assortative matching or limited mobility bias?" *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171, 673–697.

ARELLANO-BOVER, J. AND S. SAN (2023): "The Role of Firms and Job Mobility in the Assimilation of Immigrants: Former Soviet Union Jews in Israel 1990-2019," .

BABET, D., O. GODECHOT, AND M. G. PALLADINO (2022): "In the land of AKM: Explaining the dynamics of wage inequality in France," .

BERGÉ, L. (2018): "Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm," Tech. rep., Department of Economics at the University of Luxembourg.

BONHOMME, S., K. HOLZHEU, T. LAMADON, E. MANRESA, M. MOGSTAD, AND B. SETZLER (2023): "How Much Should we Trust Estimates of Firm Effects and Worker Sorting?" *Journal of Labor Economics*, 41.

CARD, D., A. R. CARDOSO, J. HEINING, AND P. KLINE (2018): "Firms and labor market inequality: Evidence and some theory," *Journal of Labor Economics*, 36, S13–S70.

CARD, D., J. HEINING, AND P. KLINE (2013): "Workplace heterogeneity and the rise of West German wage inequality," *The Quarterly Journal of Economics*, 128, 967–1015.

CORREIA, S. (2017): "Linear models with high-dimensional fixed effects: An efficient and feasible estimator," *Unpublished manuscript, http://scorreia. com/research/hdfe. pdf (last accessed 25 October 2019)*, 4.

DAUTH, W., S. FINDEISEN, E. MORETTI, AND J. SUEDEKUM (2022): "Matching in cities," *Journal of the European Economic Association*, 20, 1478–1521.

GAURE, S. (2014): "Correlation bias correction in two-way fixed-effects linear regression," *Stat*, 3, 379–390.

GERARD, F., L. LAGOS, E. SEVERNINI, AND D. CARD (2021): "Assortative matching or exclusionary hiring? the impact of employment and pay policies on racial wage differences in brazil," *American Economic Review*, 111, 3418–3457.

HELM, I., A. KÜGLER, AND U. SCHÖNBERG (2023): "Displacement Effects in Manufacturing and Structural Change," .

JOCHMANS, K. (2018): "Heteroskedasticity-robust inference in linear regression models," *arXiv preprint arXiv:1809.06136*.

JOCHMANS, K. AND M. WEIDNER (2019): "Fixed-Effect Regressions on Network Data," *Econometrica*, 87, 1543–1560.

KLINE, P., R. SAGGIO, AND M. SØLVSTEN (2020): "Leave-out estimation of variance components," *Econometrica*, 88, 1859–1898.

——— (2021): "Improved stochastic approximation of regression leverages for bias correction of variance components," Tech. rep.

KOUTIS, I., G. L. MILLER, AND D. TOLLIVER (2011): "Combinatorial preconditioners and multilevel solvers for problems in computer vision and image processing," *Computer Vision and Image Understanding*, 115, 1638–1646.

LACHOWSKA, M., A. MAS, R. SAGGIO, AND S. A. WOODBURY (2023): "Work hours mismatch," Tech. rep., National Bureau of Economic Research.

LEKNES, S., J. RATTSØ, AND H. E. STOKKE (2022): "Assortative labor matching, city size, and the education level of workers," *Regional Science and Urban Economics*, 96, 103806.

LOPES DE MELO, R. (2018): "Firm wage differentials and labor market sorting: Reconciling theory and evidence," *Journal of Political Economy*, 126, 000–000.

MACKINNON, J. G. AND A. A. SMITH JR (1998): "Approximate bias correction in econometrics," *Journal of Econometrics*, 85, 205–230.

MARCUS, M. AND W. R. GORDON (1971): "An extension of the Minkowski determinant theorem," *Proceedings of the Edinburgh Mathematical Society*, 17, 321–324.

MEYER, C. D. (2000): *Matrix analysis and applied linear algebra*, Society for Industrial and Applied Mathematics.

MOHAMMADI, M. (2016): "On the bounds for diagonal and off-diagonal elements of the hat matrix in the linear regression model," *REVSTAT Statistical Journal*, 14, 75–87.

SONG, J., D. J. PRICE, F. GUVENEN, N. BLOOM, AND T. VON WACHTER (2019): "Firming up inequality," *The Quarterly Journal of Economics*, 134, 1–50.

SORKIN, I. (2018): "Ranking firms using revealed preference," *The Quarterly Journal of Economics*, 133, 1331–1393.

APPENDIX

A Proofs

Proof of Proposition 1: By the linearity of the trace and expectation operators we have that

$$\mathbb{E}\left(\widehat{\delta}_D \,\middle|\, \mathbf{X}\right) = \mathbb{E}\left(\mathbf{tr}\left(\mathbf{S}_{\mathbf{X}}^T \mathbf{A} \mathbf{S}_{\mathbf{X}} \widehat{\mathbf{V}} \,\middle|\, \mathbf{X}\right)\right) = \mathbf{tr}\left(\mathbf{S}_{\mathbf{X}}^T \mathbf{A} \mathbf{S}_{\mathbf{X}} \mathbb{E}\left(\widehat{\mathbf{V}} \,\middle|\, \mathbf{X}\right)\right) = \mathbf{tr}\left(\mathbf{S}_{\mathbf{X}}^T \mathbf{A} \mathbf{S}_{\mathbf{X}} \mathbb{V}\left(\varepsilon \,|\, \mathbf{X}\right)\right) = \delta. \quad \square$$

Proof of Corollary 1:

$$\mathbb{E}\left(\widehat{\varphi} \,|\, \mathbf{X}\right) = \varphi - \mathbb{E}\left(\widehat{\delta}_D \,\middle|\, \mathbf{X}\right) + \delta = \varphi - \delta + \delta = \varphi. \quad \square$$

Proof of Proposition 2: The OLS estimate of running a regression of v^* on $\mathbf{X}$ is $\widehat{\boldsymbol{\beta}^*} = \left(\mathbf{X}^T X\right)^{-1} \mathbf{X}^T v^*$. As $\mathbb{E}\left(v^* \,|\, \mathbf{X}\right) = 0$, then we have that $\mathbb{E}\left(\widehat{\boldsymbol{\beta}^*} \,\middle|\, \mathbf{X}\right) = 0$. Then, using the formula for the expectation of quadratic forms we get:

$$\mathbb{E}\left(\widehat{\boldsymbol{\beta}^*}^T \mathbf{A} \widehat{\boldsymbol{\beta}^*} \,\middle|\, \mathbf{X}\right) = \mathbf{tr}\left(\mathbf{A} \mathbb{V}\left(\widehat{\boldsymbol{\beta}^*} \,\middle|\, \mathbf{X}\right)\right) = \mathbf{tr}\left(\mathbf{S}_{\mathbf{X}}^T \mathbf{A} \mathbf{S}_{\mathbf{X}} \mathbb{V}\left(v^* \,|\, \mathbf{X}\right)\right) = \mathbf{tr}\left(\mathbf{S}_{\mathbf{X}}^T \mathbf{A} \mathbf{S}_{\mathbf{X}} \widehat{\mathbf{V}}\right) = \widehat{\delta}_D,$$

where the second equality we use $\mathbb{V}\left(\widehat{\boldsymbol{\beta}^*} \,\middle|\, \mathbf{X}\right) = \mathbf{S}_{\mathbf{X}} \mathbb{V}\left(v^* \,|\, \mathbf{X}\right) \mathbf{S}_{\mathbf{X}}^T$ and the cyclical property of the trace. The third equality follows by the definition of v^* where $\mathbb{V}\left(v^* \,|\, \mathbf{X}\right) = \widehat{\mathbf{V}}$. $\square$

Proof of Proposition 3: First, given the decomposition of $\widehat{\mathbf{V}} = \widehat{\mathbf{V}}_+ - \widehat{\mathbf{V}}_-$ and the linearity of the trace operator, we have that

$$\widehat{\delta}_D = \mathbf{tr}\left(\mathbf{S}_{\mathbf{X}}^T \mathbf{A} \mathbf{S}_{\mathbf{X}} \widehat{\mathbf{V}}\right) = \mathbf{tr}\left(\mathbf{S}_{\mathbf{X}}^T \mathbf{A} \mathbf{S}_{\mathbf{X}} \widehat{\mathbf{V}}_+\right) - \mathbf{tr}\left(\mathbf{S}_{\mathbf{X}}^T \mathbf{A} \mathbf{S}_{\mathbf{X}} \widehat{\mathbf{V}}_-\right).$$

As $\mathbb{E}\left(v_+^* \,|\, \mathbf{X}\right) = 0$ and $\mathbb{E}\left(v_-^* \,|\, \mathbf{X}\right) = 0$, then we have that $\mathbb{E}\left(\widehat{\boldsymbol{\beta}^*}_+ \,\middle|\, \mathbf{X}\right) = 0$ and $\mathbb{E}\left(\widehat{\boldsymbol{\beta}^*}_- \,\middle|\, \mathbf{X}\right) = 0$. Then, as with Proposition 2 we have:

$$\mathbb{E}\left(\widehat{\boldsymbol{\beta}^*}_+^T \mathbf{A} \widehat{\boldsymbol{\beta}^*}_+ \,\middle|\, \mathbf{X}\right) = \mathbf{tr}\left(\mathbf{S}_{\mathbf{X}}^T \mathbf{A} \mathbf{S}_{\mathbf{X}} \widehat{\mathbf{V}}_+\right), \quad \text{and} \quad \mathbb{E}\left(\widehat{\boldsymbol{\beta}^*}_-^T \mathbf{A} \widehat{\boldsymbol{\beta}^*}_- \,\middle|\, \mathbf{X}\right) = \mathbf{tr}\left(\mathbf{S}_{\mathbf{X}}^T \mathbf{A} \mathbf{S}_{\mathbf{X}} \widehat{\mathbf{V}}_-\right). \quad \square$$

Proof of Proposition 4: Under the bootstrap, i.e. conditional on X, the only source of randomness are v_+^* and v_-^*. So we prove that the bootstrap correction is an unbiased and consistent estimate of the directed bias correction under the bootstrap.

Unbiased. Taking expectations over δ^*, we have

$$\begin{aligned}\mathbb{E}\left(\delta^* \mid \mathbf{X}\right) &= \frac{1}{J}\sum_{j=1}^{J}\mathbb{E}\left(\widehat{\boldsymbol{\beta}^*}_+(j)^T\mathbf{A}\widehat{\boldsymbol{\beta}^*}_+(j)\,\middle|\,\mathbf{X}\right) - \frac{1}{J}\sum_{j=1}^{J}\mathbb{E}\left(\widehat{\boldsymbol{\beta}^*}_-(j)^T\mathbf{A}\widehat{\boldsymbol{\beta}^*}_-(j)\,\middle|\,\mathbf{X}\right)\\ &= \frac{1}{J}\sum_{j=1}^{J}\left[\mathbb{E}\left(\widehat{\boldsymbol{\beta}^*}_+(j)^T\mathbf{A}\widehat{\boldsymbol{\beta}^*}_+(j)\,\middle|\,\mathbf{X}\right) - \mathbb{E}\left(\widehat{\boldsymbol{\beta}^*}_-(j)^T\mathbf{A}\widehat{\boldsymbol{\beta}^*}_-(j)\,\middle|\,\mathbf{X}\right)\right]\\ &= \frac{1}{J}\sum_{j=1}^{J}\widehat{\delta}_D = \widehat{\delta}_D.\end{aligned}$$

Consistent. Using the two components of the difference of averages from the definition of δ^*, we have that:

$$\begin{aligned}&\frac{1}{J}\sum_{j=1}^{J}\widehat{\boldsymbol{\beta}^*}_+(j)^T\mathbf{A}\widehat{\boldsymbol{\beta}^*}_+(j) \xrightarrow{a.s.} \mathbb{E}\left(\widehat{\boldsymbol{\beta}^*}_+(j)^T\mathbf{A}\widehat{\boldsymbol{\beta}^*}_+(j)\,\middle|\,\mathbf{X}\right), \text{ and}\\ &\frac{1}{J}\sum_{j=1}^{J}\widehat{\boldsymbol{\beta}^*}_-(j)^T\mathbf{A}\widehat{\boldsymbol{\beta}^*}_-(j) \xrightarrow{a.s.} \mathbb{E}\left(\widehat{\boldsymbol{\beta}^*}_-(j)^T\mathbf{A}\widehat{\boldsymbol{\beta}^*}_-(j)\,\middle|\,\mathbf{X}\right),\end{aligned}$$

as each quadratic form is iid with defined expectation. Then, $\delta^* \xrightarrow{a.s.} \widehat{\delta}_D$. □

Proof of Proposition 5: We have that for bootstrap j,

$$\widehat{\boldsymbol{\beta}}^*_{MS}(j)^T\mathbf{A}\widehat{\boldsymbol{\beta}}^*_{MS}(j) = \widehat{\boldsymbol{\beta}}^T\mathbf{A}\widehat{\boldsymbol{\beta}} + \boldsymbol{v}^*(j)^T\mathbf{S}_{\mathbf{X}}^T\mathbf{A}\mathbf{S}_{\mathbf{X}}\boldsymbol{v}^*(j) + 2\boldsymbol{v}^*(j)^T\mathbf{S}_{\mathbf{X}}^T\mathbf{A}\widehat{\boldsymbol{\beta}}.$$

We have that

$$\mathbb{V}\left(\delta^*_{MS}\middle|\,\mathbf{X},\boldsymbol{\varepsilon}\right) = \frac{1}{J}\mathbb{V}\left(\widehat{\boldsymbol{\beta}}^*_{MS}(j)^T\mathbf{A}\widehat{\boldsymbol{\beta}}^*_{MS}(j)\middle|\,\mathbf{X},\boldsymbol{\varepsilon}\right).$$

Let the matrix $\mathbf{S}_{\mathbf{X}}^T\mathbf{A}\mathbf{S}_{\mathbf{X}} \equiv \mathbf{Z}$, with elements (i,j) equal to $z_{i,j}$. Also, let the vector $\mathbf{S}_{\mathbf{X}}^T\mathbf{A}\widehat{\boldsymbol{\beta}} \equiv \boldsymbol{w}$ with element k equal to w_k. We will ignore the index j for clarity. Then,

$$cov\left(\boldsymbol{v}^{*T}\mathbf{Z}\boldsymbol{v}^*,\quad 2\boldsymbol{v}^{*T}\boldsymbol{w}\middle|\,\mathbf{X},\boldsymbol{\varepsilon}\right) = \mathbb{E}\left(\left(\sum_{i=1}^{n}\sum_{j=1}^{n}z_{i,j}v_i^*v_j^*\right)\left(\sum_{k=1}^{n}w_kv_k^*\right)\,\middle|\,\mathbf{X},\boldsymbol{\varepsilon}\right),$$

where we use the fact that $\mathbb{E}\left(v_i^*\mid\mathbf{X},\boldsymbol{\varepsilon}\right) = 0$. Then,

$$\mathbb{E}\left(\left(\sum_{i=1}^{n}\sum_{j=1}^{n}z_{i,j}v_i^*v_j^*\right)\left(\sum_{k=1}^{n}w_kv_k^*\right)\,\middle|\,\mathbf{X},\boldsymbol{\varepsilon}\right) = \left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}z_{i,j}w_k\mathbb{E}\left(v_i^*v_j^*v_k^*\mid\mathbf{X},\boldsymbol{\varepsilon}\right)\right) = 0,$$

where we use that the bootstrap errors are independent across observations and the fact that $\mathbb{E}\left((v_i^*)^3 \mid \mathbf{X}, \boldsymbol{\varepsilon}\right) = 0$.

This means that:

$$\mathbb{V}\left(\delta_{MS}^* \middle| \mathbf{X}, \boldsymbol{\varepsilon}\right) = \frac{1}{J}\mathbb{V}\left(\boldsymbol{v}^{*T}\mathbf{S}_{\mathbf{X}}^T\mathbf{A}\mathbf{S}_{\mathbf{X}}\boldsymbol{v}^* \middle| \mathbf{X}, \boldsymbol{\varepsilon}\right) + \frac{4}{J}\mathbb{V}\left(\boldsymbol{v}^{*T}\mathbf{S}_{\mathbf{X}}^T\mathbf{A}\widehat{\boldsymbol{\beta}} \middle| \mathbf{X}, \boldsymbol{\varepsilon}\right).$$

The expression above can be rewritten as:

$$\mathbb{V}\left(\delta_{MS}^* \middle| \mathbf{X}, \boldsymbol{\varepsilon}\right) = \mathbb{V}\left(\delta^* \middle| \mathbf{X}, \boldsymbol{\varepsilon}\right) + \frac{4}{J}\mathbb{V}\left(\boldsymbol{v}^{*T}\mathbf{S}_{\mathbf{X}}^T\mathbf{A}\widehat{\boldsymbol{\beta}} \middle| \mathbf{X}, \boldsymbol{\varepsilon}\right) \geq \mathbb{V}\left(\delta^* \middle| \mathbf{X}\right).$$

□

Proof of Proposition 6: First note that the fitted value for observation i after running a regression of $\boldsymbol{r}$ on $\mathbf{X}$ is $\widehat{r}_i = \mathbf{X}_i\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{r}$, where $\mathbf{X}_i$ correspond to the ith row of $\mathbf{X}$. Then,

$$\mathbb{E}\left(\widehat{r}_i^2 \middle| \mathbf{X}\right) = \mathbf{X}_i\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbb{E}\left(\boldsymbol{r}\boldsymbol{r}^T\right)\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_i^T = \mathbf{X}_i\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_i^T = P_{ii},$$

where we used the fact that $\mathbb{E}\left(\boldsymbol{r}\boldsymbol{r}^T\right) = \mathbf{I}$.

Now, let $\mathbf{1}_i$ be a vector of length n of zeros everywhere except for the ith observation. Then, we do something similar for the squared residuals:

$$\begin{aligned}
\mathbb{E}\left((r_i - \widehat{r}_i)^2 \middle| \mathbf{X}\right) &= \mathbb{E}\left(r_i^2\right) - 2\mathbb{E}\left(\widehat{r}_i r_i \mid \mathbf{X}\right) + \mathbb{E}\left(\widehat{r}_i^2 \middle| \mathbf{X}\right) \\
&= 1 - 2\mathbf{X}_i\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbb{E}\left(\boldsymbol{r}r_i\right) + P_{ii} \\
&= 1 - 2\mathbf{X}_i\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{1}_i + P_{ii} \\
&= 1 - 2\mathbf{X}_i\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_i^T + P_{ii} \\
&= 1 - 2P_{ii} + P_{ii} = 1 - P_{ii}. \quad \square
\end{aligned}$$

Proof of Proposition 7: The fitted value vector for observations belonging to cluster g after running a regression of $\boldsymbol{r}$ on $\mathbf{X}$ is $\widehat{\boldsymbol{r}}_g = \mathbf{X}_g\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{r}$, where $\mathbf{X}_g$ correspond to the rows of the observations belonging to cluster g. Then,

$$\mathbb{E}\left(\widehat{\boldsymbol{r}}_g\widehat{\boldsymbol{r}}_g^T \middle| \mathbf{X}\right) = \mathbf{X}_g\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbb{E}\left(\boldsymbol{r}\boldsymbol{r}^T\right)\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_g^T = \mathbf{X}_g\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_g^T = \mathbf{P}_{gg}.$$

Let $\mathbf{O}_g$ be a row selection matrix of dimensions $n_g \times n$ that when multiplied to a matrix it selects the rows corresponding to the observations of cluster g. Then,

$$\begin{aligned}
\mathbb{E}\left(\left(\boldsymbol{r}_g - \widehat{\boldsymbol{r}}_g\right)\left(\boldsymbol{r}_g - \widehat{\boldsymbol{r}}_g\right)^T \middle| \mathbf{X}\right) &= \mathbb{E}\left(\boldsymbol{r}_g \boldsymbol{r}_g^T\right) - \mathbb{E}\left(\boldsymbol{r}_g \widehat{\boldsymbol{r}}_g^T \middle| \mathbf{X}\right) - \mathbb{E}\left(\widehat{\boldsymbol{r}}_g \boldsymbol{r}_g^T \middle| \mathbf{X}\right) + \mathbb{E}\left(\widehat{\boldsymbol{r}}_g \widehat{\boldsymbol{r}}_g^T \middle| \mathbf{X}\right) \\
&= \mathbf{I}_{n_g} - \mathbb{E}\left(\boldsymbol{r}_g \boldsymbol{r}^T\right) \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}_g^T - \mathbf{X}_g \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T \mathbb{E}\left(\boldsymbol{r}\boldsymbol{r}_g^T\right) + \mathbf{P}_{gg} \\
&= \mathbf{I}_{n_g} - \mathbf{O}_g \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}_g^T - \mathbf{X}_g \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{O}_g^T + \mathbf{P}_{gg} \\
&= \mathbf{I}_{n_g} - \mathbf{X}_g \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}_g^T - \mathbf{X}_g \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}_g^T + \mathbf{P}_{gg} \\
&= \mathbf{I}_{n_g} - 2\mathbf{X}_g \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}_g^T + \mathbf{P}_{gg} = \mathbf{I}_{n_g} - \mathbf{P}_{gg} = \mathbf{M}_{gg}. \quad \square
\end{aligned}$$

Let us introduce the definitions of $\widehat{\mathbf{P}}_{gg}$ and $\widehat{\mathbf{M}}_{gg}$ and an auxiliary Lemma that will prove helpful for proving Proposition 2. The definitions for $\widehat{P}_{ii}$ and $\widehat{M}_{ii}$ follow from collapsing the clusters to have only one observation, i.e. $n_g = 1$ for all g.

Definition 6 (Estimates block-diagonals of **P** and **M**). *Let $\boldsymbol{r}(j)$ be a random vector that corresponds to the jth realization and $\widehat{\boldsymbol{r}}(j)$ the fitted value of running a regression of $\boldsymbol{r}(j)$ on $\mathbf{X}$. In a similar way, define $\boldsymbol{r}_g(j)$ and $\widehat{\boldsymbol{r}}_g(j)$ as the vectors containing the observations of $\boldsymbol{r}(j)$ and $\widehat{\boldsymbol{r}}(j)$ that correspond to cluster g. Then, The estimates $\widehat{\mathbf{P}}_{gg}$ and $\widehat{\mathbf{M}}_{gg}$ are defined as:*

$$\widehat{\mathbf{P}}_{gg} = \frac{1}{J}\sum_{j=1}^{J} \widehat{\boldsymbol{r}}_g(j)\widehat{\boldsymbol{r}}_g(j)^T, \quad \textit{and} \quad \widehat{\mathbf{M}}_{gg} = \frac{1}{J}\sum_{j=1}^{J} \left(\boldsymbol{r}_g(j) - \widehat{\boldsymbol{r}}_g(j)\right)\left(\boldsymbol{r}_g(j) - \widehat{\boldsymbol{r}}_g(j)\right)^T.$$

Lemma 1. *Let $\mathbf{A}$ be a positive definite matrix and matrix $\mathbf{B}$ be positive semi-definite. Then $\mathbf{AB}$ has only non-negative eigenvalues. If $\mathbf{B}$ is positive definite, then $\mathbf{AB}$ has only positive eigenvalues.*

Proof. Let $\boldsymbol{v}$ be an eigenvector of $\mathbf{AB}$ with associated eigenvalue λ, i.e. $\mathbf{AB}\boldsymbol{v} = \lambda\boldsymbol{v}$. As $\mathbf{A}$ is positive definite we have that:

$$(\mathbf{B}\boldsymbol{v})^T \mathbf{A}\,(\mathbf{B}\boldsymbol{v}) = \lambda \boldsymbol{v}^T \mathbf{B}^T \boldsymbol{v} \geq 0.$$

The expression above can be equal to zero if $\lambda = 0$. $\mathbf{B}$ is positive semi-definite so $\boldsymbol{v}^T\mathbf{B}^T\boldsymbol{v} \geq 0$, which means $\lambda \geq 0$.

For the case where $\mathbf{B}$ is positive definite, we have that for any non-zero vector $\boldsymbol{v}^T\mathbf{B}^T\boldsymbol{v} > 0$, which means that $\mathbf{B}\boldsymbol{v} > 0$. Similarly as $\mathbf{A}$ is positive definite we have that:

$$(\mathbf{B}\boldsymbol{v})^T \mathbf{A}\,(\mathbf{B}\boldsymbol{v}) = \lambda \boldsymbol{v}^T \mathbf{B}^T \boldsymbol{v} > 0, \quad \Longrightarrow \quad \lambda > 0.$$

$\square$

It is easy to see that the estimates above satisfy the constraint $\overline{\mathbf{P}}_{gg} + \overline{\mathbf{M}}_{gg} = \mathbf{I}_{n_g}$. The proposition below shows that $\overline{\mathbf{P}}_{gg}$ and $\overline{\mathbf{M}}_{gg}$ the estimates also satisfy the constraints on the eigenvalues.

Lemma 2 (Eigenvalue properties of $\overline{\mathbf{P}}_{gg}$ and $\overline{\mathbf{M}}_{gg}$)**.** *Define the following matrices:*

$$\overline{\mathbf{P}}_{gg} = \left(\widehat{\mathbf{P}}_{gg} + \widehat{\mathbf{M}}_{gg}\right)^{-1} \widehat{\mathbf{P}}_{gg}, \quad \textit{and} \quad \overline{\mathbf{M}}_{gg} = \left(\widehat{\mathbf{P}}_{gg} + \widehat{\mathbf{M}}_{gg}\right)^{-1} \widehat{\mathbf{M}}_{gg}.$$

Assume $\widehat{\mathbf{M}}_{gg}$ is non-singular. Then, the eigenvalues of $\overline{\mathbf{P}}_{gg}$ lie within $[0,1)$ and the eigenvalues of $\overline{\mathbf{M}}_{gg}$ lie within $(0,1]$.

Proof. First, both $\widehat{\mathbf{P}}_{gg}$ and $\widehat{\mathbf{M}}_{gg}$ are positive semi-definite as they are averages of matrices formed by outer products of vectors. By assumption, we have that $\widehat{\mathbf{M}}_{gg}$ is non-singular. Together with the positive semi-definite property, this implies that $\widehat{\mathbf{M}}_{gg}$ has strictly positive eigenvalues. As it is symmetric, then $\widehat{\mathbf{M}}_{gg}$ is positive definite. Then, $\widehat{\mathbf{P}}_{gg} + \widehat{\mathbf{M}}_{gg}$ is positive definite as well, which means its inverse exist and is also positive definite. Now, using Lemma 1 we can show that $\overline{\mathbf{P}}_{gg} = \left(\widehat{\mathbf{P}}_{gg} + \widehat{\mathbf{M}}_{gg}\right)^{-1} \widehat{\mathbf{P}}_{gg}$ has non-negative eigenvalues and $\overline{\mathbf{M}}_{gg} = \left(\widehat{\mathbf{P}}_{gg} + \widehat{\mathbf{M}}_{gg}\right)^{-1} \widehat{\mathbf{M}}_{gg}$ has only positive eigenvalues. Let λ be an eigenvalue of $\overline{\mathbf{P}}_{gg}$. Then, we have that as $\overline{\mathbf{M}}_{gg} = \mathbf{I}_{n_g} - \overline{\mathbf{P}}_{gg}$, then $1 - \lambda$ is an eigenvalue of $\overline{\mathbf{M}}_{gg}$. We can conclude then that all eigenvalues of $\overline{\mathbf{P}}_{gg}$ are in $[0,1)$ and the eigenvalues of $\overline{\mathbf{M}}_{gg}$ are in $(0,1]$. □

Proof of Proposition 9: First, we will show that $\overline{\mathbf{P}}^S_{gg}$ and $\overline{\mathbf{M}}^S_{gg}$ are similar matrices to $\overline{\mathbf{P}}_{gg}$ and $\overline{\mathbf{M}}_{gg}$, defined in Lemma 2. As $\widehat{\mathbf{M}}_{gg}$ is non-singular then $\widehat{\mathbf{P}}_{gg} + \widehat{\mathbf{M}}_{gg}$ is positive definite and there exists a unique Cholesky decomposition where $\widehat{\mathbf{P}}_{gg} + \widehat{\mathbf{M}}_{gg} = \mathbf{L}_{gg}\mathbf{L}^T_{gg}$ and $\mathbf{L}_{gg}$ is non-singular. Then, $\overline{\mathbf{P}}_{gg} = \left(\mathbf{L}_{gg}\mathbf{L}^T_{gg}\right)^{-1} \widehat{\mathbf{P}} = \left(\mathbf{L}^T_{gg}\right)^{-1} \mathbf{L}^{-1}_{gg}\widehat{\mathbf{P}}$. Pre-multiply $\overline{\mathbf{P}}_{gg}$ by $\mathbf{L}^T_{gg}$ and post-multiply it by $\left(\mathbf{L}^T_{gg}\right)^{-1}$ and we get

$$\mathbf{L}^T_{gg}\overline{\mathbf{P}}_{gg}\left(\mathbf{L}^T_{gg}\right)^{-1} = \mathbf{L}^T_{gg}\left(\mathbf{L}^T_{gg}\right)^{-1}\mathbf{L}^{-1}_{gg}\widehat{\mathbf{P}}_{gg}\left(\mathbf{L}^T_{gg}\right)^{-1} = \mathbf{L}^{-1}_{gg}\widehat{\mathbf{P}}_{gg}\left(\mathbf{L}^{-1}_{gg}\right)^T = \overline{\mathbf{P}}^S_{gg}.$$

Then, $\overline{\mathbf{P}}_{gg}$ and $\overline{\mathbf{P}}^S_{gg}$ are similar matrices, which means they have the same eigenvalues. By Lemma 2 we have then that the eigenvalues of $\overline{\mathbf{P}}^S_{gg}$ lie within $[0,1)$. Similar argument to show that the eigenvalues of $\overline{\mathbf{M}}^S_{gg}$ lie within $(0,1]$. □

Proof of Proposition 8: Let $\widehat{\boldsymbol{m}}_{gg}(j) \equiv \left(\boldsymbol{r}_g(j) - \widehat{\boldsymbol{r}}_g(j)\right)\left(\boldsymbol{r}_g(j) - \widehat{\boldsymbol{r}}_g(j)\right)^T$. Then, denote $\widehat{\mathbf{M}}_{gg}(J)$ as the average over J realizations of $\widehat{\boldsymbol{m}}_{gg}(j)$:

$$\widehat{\mathbf{M}}_{gg}(J) = \frac{1}{J}\sum_{j=1}^{J} \widehat{\boldsymbol{m}}_{gg}(j).$$

We have that $\widehat{\mathbf{M}}_{gg}(J) \xrightarrow{a.s.} \mathbf{M}_{gg}$. By the continuous mapping theorem we have that

$$\det\left(\widehat{\mathbf{M}}_{gg}(J)\right) \xrightarrow{a.s.} \det\left(\mathbf{M}_{gg}\right),$$

where $\det\left(\mathbf{M}_{gg}\right) = 0$ by assumption that $\mathbf{M}_{gg}$ is singular.

As $\widehat{\boldsymbol{m}}_{gg}(j)$ is an outer product it is positive semi-definite and singular. This means that $\det\left(\widehat{\mathbf{M}}_{gg}(1)\right) = 0$. Also, the Minkowski determinant theorem (see Marcus and Gordon, 1971) implies that the determinant of the sum of two positive semi-definite matrices is greater or equal to the sum of the determinants of each matrix. All of this implies that $\det\left(\widehat{\mathbf{M}}_{gg}(J)\right) \geq 0$.

We proceed by contradiction. Suppose there exists a $J^* > 1$ such that with positive probability

$$\det\left(\widehat{\mathbf{M}}_{gg}(J^*)\right) > 0.$$

Now fix J^* and let $J = KJ^*$. Then, we can rewrite $\widehat{\mathbf{M}}_{gg}(J)$ as:

$$\widehat{\mathbf{M}}_{gg}(J) = \frac{1}{J}\sum_{k=1}^{K} J^* \times \widehat{\mathbf{M}}_{gg}^{(k)}(J^*),$$

where $\widehat{\mathbf{M}}_{gg}^{(k)}(J^*)$ denotes the kth realization of $\widehat{\mathbf{M}}_{gg}(J^*)$. Then,

$$\det\left(\widehat{\mathbf{M}}_{gg}(J)\right) = \det\left(\sum_{k=1}^{K}\frac{J^*}{J} \times \widehat{\mathbf{M}}_{gg}^{(k)}(J^*)\right) \geq \sum_{k=1}^{K}\frac{J^*}{J} \times \det\left(\widehat{\mathbf{M}}_{gg}^{(k)}(J^*)\right) = \frac{1}{K}\sum_{k=1}^{K}\det\left(\widehat{\mathbf{M}}_{gg}^{(k)}(J^*)\right).$$

Denote the last expression in the right as $\overline{D}(K)$. As $K \to \infty$, then $\overline{D}(K) \xrightarrow{a.s.} \mathbb{E}\left(\det\left(\widehat{\mathbf{M}}_{gg}(J^*)\right)\right)$. As the determinant of $\widehat{\mathbf{M}}_{gg}(J^*)$ is always non-negative, and with positive probability it can be strictly positive, then $\mathbb{E}\left(\det\left(\widehat{\mathbf{M}}_{gg}(J^*)\right)\right) > 0$. But as $K \to \infty$, then $J \to \infty$ which means that $\det\left(\widehat{\mathbf{M}}_{gg}(J)\right) \xrightarrow{a.s.} 0$. This leads to a contradiction. □

B Additional details on leverage estimation

The estimates $\widehat{P}_{ii}$ and $\widehat{M}_{ii}$ are:

$$\widehat{P}_{ii} = \frac{1}{J_M}\sum_{j=1}^{J_M}(\widehat{r}_i(j))^2 \quad \text{and} \quad \widehat{M}_{ii} = \frac{1}{J_M}\sum_{j=1}^{J_M}(r_i(j) - \widehat{r}_i(j))^2,$$

where $r_i(j)$ is the jth realization of the ith entry of Rademacher random vector; $\widehat{r}_i(j)$ is the ith fitted value of running the regression of the jth realization of the random vector on $\mathbf{X}$.

As covered in Case 2 of Section 4, the leave-one-out residual for observation i is equal to $\widehat{\varepsilon}_i/M_{ii}$. As we use the estimate $\overline{M}_{ii}$ rather than the actual value M_{ii}, we introduce some non-linearity bias. We correct it up to a second order.

To do this, note that the expected value of the second-order approximation of $1/\overline{M}_{ii}$ is

$$\mathbb{E}\left(\frac{1}{\overline{M}_{ii}}\right) \approx \frac{1}{M_{ii}} + \frac{P_{ii}}{M_{ii}^3}\mathbb{E}\left(\widehat{M}_{ii} - M_{ii}\right)^2 - \frac{1}{M_{ii}^2}\left(\mathbb{E}\left((\widehat{P}_{ii} - P_{ii})(\widehat{M}_{ii} - M_{ii})\right)\right).$$

The feasible bias corrected estimate of $1/M_{ii}$ would be

$$\frac{1}{\overline{M}_{ii}}\left(1 - \frac{\overline{P}_{ii}}{\overline{M}_{ii}^2}\widehat{\text{var}}(\widehat{M}_{ii}) + \frac{1}{\overline{M}_{ii}}\widehat{\text{cov}}\left(\widehat{P}_{ii}, \widehat{M}_{ii}\right)\right),$$

where $\widehat{\text{var}}$ and $\widehat{\text{cov}}$ are sample variance and covariance estimates.[14]

Direct computation. Alternatively, an exact computation of the leverage is possible by using the definition of fitted values $\widehat{\boldsymbol{y}} = \mathbf{P}\boldsymbol{y}$ and a regression-intensive procedure. We have that the leverage of observation i is equal to

$$P_{ii} = \frac{\partial \widehat{y}_i}{\partial y_i},$$

where y_i and $\widehat{y}_i$ are the ith elements of $\widehat{\boldsymbol{y}}$ and $\boldsymbol{y}$.

The following remark shows how to compute these leverages without computing the projection matrix $\mathbf{P}$ using only linear regressions.

[14] The sample variance of $\widehat{M}_{ii}$ is $\frac{1}{J_M}\left(\left[\frac{1}{J_M-1}\sum_{j=1}^{J_M}(r_i(j) - \widehat{r}_i(j))^4\right] - \frac{J_M}{J_M-1}\widehat{M}_{ii}^2\right)$. The sample covariance is $\frac{1}{J_M}\left(\left[\frac{1}{J_M-1}\sum_{j=1}^{J_M}(r_i(j) - \widehat{r}_i(j))^2\widehat{r}_i(j)^2\right] - \widehat{M}_{ii}\widehat{P}_{ii}\right)$.

Proposition 10. *Let $\tilde{\boldsymbol{y}}(i)$ be a vector of length n where every entry is equal to zero, except the ith entry that is equal to one. The leverage of observation i is equal to the fitted value $\widehat{y}_i$ of a linear regression of $\tilde{\boldsymbol{y}}(i)$ on $\mathbf{X}$.*

Proof. Let $\mathbf{P}_i$ be the ith row of the projection matrix $\mathbf{P}$. Then, for any vector $\boldsymbol{y}$ we have that the ith fitted value $\widehat{y}_i$ is equal to $\widehat{y}_i = \mathbf{P}_i \boldsymbol{y} = \sum_j P_{ij} y_j$. Let $\boldsymbol{y} = \tilde{\boldsymbol{y}}(i)$. Then $\widehat{y}_i = P_{ii}$. □

When the data set is large, the direct computation of the leverages is not feasible. We leave the exact computation for the problematic cases identified by the following diagnostic.

Diagnostic and adjustment. Although using $\overline{M}_{ii}$ as the estimate of M_{ii} rules out nonsensical estimates outside the $[0, 1]$ interval, the estimates for $1/M_{ii}$, could still violate some theoretical bounds. We detect problematic estimations of $1/M_{ii}$ by checking that they are consistent with the theoretical bounds for the leverages $P_{ii} \in [1/n, 1]$. These bounds are derived from the following proposition, which might be well known for some readers.

Proposition 11. *Let $\mathbf{X}$ be a full rank matrix of dimensions $n \times k$, where a vector of ones can be obtained through column operations. Let $\mathbf{P} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$, with ith diagonal element P_{ii}. Then $1/n \leq h_{ii} \leq 1$ for all i.*

Proof. As $\mathbf{P}$ is idempotent then $P_{ii} = P_{ii}^2 + \sum_{j \neq i} P_{ij}^2$. Then $P_{ii} \leq P_{ii}^2 \Longrightarrow P_{ii} \leq 1$. Now, let $\tilde{\mathbf{X}}$ be the full rank matrix of dimensions $n \times k$ that contains a vector of ones after doing column operations on $\mathbf{X}$. Then define $\tilde{\mathbf{P}} = \tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'$ with diagonal elements $\tilde{P}_{ii}$. It is well known that $1/n \leq \tilde{P}_{ii}$ (see for example Lemma 2.2 in Mohammadi (2016)). As $\mathbf{X}$ and $\tilde{\mathbf{X}}$ have the same column space, then $\mathbf{P} = \tilde{\mathbf{P}}$. Thus, $1/n \leq P_{ii}$. □

The corollary of the proposition above is that $1/M_{ii} \geq n/(n-1)$. Thus, we check if our estimates of $1/M_{ii}$ satisfy this bound. We directly compute leverages corresponding to the estimates of $1/Mii$ that fall outside those bounds by using the result of Proposition 10.

The following algorithm takes as inputs the covariates $\mathbf{X}$ and gives output a combination of actual and estimates for $1/M_{ii}$ that will be used for the computation of the leave-one-out residuals.

Steps 1 to 8 of the algorithm estimate $\widehat{P}_{ii}$ and $\widehat{M}_{ii}$. Steps 9 and 10 compute the necessary objects to compute the bias correction coming from the non-linearity of $1/M_{ii}$. Steps 12 to 19 perform the diagnostic and, if necessary, the computation of the actual leverage P_{ii}.

Algorithm 2 Estimate leverages, diagnosis and compute those out of bounds

$\boldsymbol{z}_P^{(0)} = \mathbf{0}$, $\boldsymbol{z}_M^{(0)} = \mathbf{0}$, $\boldsymbol{z}_2^{(0)} = \mathbf{0}$, and $\boldsymbol{z}_{PM}^{(0)} = \mathbf{0}$ are vectors of length n.
for $j = 1, \ldots, J_M$ **do**
    Simulate a vector $\boldsymbol{r}$ of length n of mutually independent Rademacher entries.
    Compute fitted values $\widehat{\boldsymbol{r}}$ from a regression of $\boldsymbol{r}$ on $\mathbf{X}$.
    Compute $\boldsymbol{z}_P^{(j)} = \boldsymbol{z}_P^{(j-1)} + (\widehat{\boldsymbol{r}})^2$ and $\boldsymbol{z}_M^{(j)} = \boldsymbol{z}_M^{(j-1)} + (\boldsymbol{r} - \widehat{\boldsymbol{r}})^2$.
    Compute $\boldsymbol{z}_2^{(j)} = \boldsymbol{z}_2^{(j-1)} + (\boldsymbol{r} - \widehat{\boldsymbol{r}})^4$ and $\boldsymbol{z}_{PM}^{(j)} = \boldsymbol{z}_{PM}^{(j-1)} + (\boldsymbol{r} - \widehat{\boldsymbol{r}})^2 (\widehat{\boldsymbol{r}})^2$
end for
Compute $\widehat{P}_{ii} = z_{P,i}^{(J_M)} / J_M$ and $\widehat{M}_{ii} = z_{M,i}^{(J_M)} / J_M$ for all $i \in \{1, \ldots, n\}$.
Compute $\widehat{\text{var}}(\widehat{M}_{ii}) = \frac{1}{J_M} \left(\frac{z_{2,i}^{(J_M)}}{J_M - 1} - \frac{J_M}{J_M - 1} \widehat{M}_{ii}^2 \right)$ for all $i \in \{1, \ldots, n\}$.
Compute $\widehat{\text{cov}}(\widehat{P}_{ii}, \widehat{M}_{ii}) = \frac{1}{J_M} \left(\frac{z_{PM,i}^{(J_M)}}{J_M - 1} - \frac{J_M}{J_M - 1} \widehat{P}_{ii} \widehat{M}_{ii} \right)$ for all $i \in \{1, \ldots, n\}$.
Compute $\overline{M}_{ii} = \frac{\widehat{M}_{ii}}{\widehat{P}_{ii} + \widehat{M}_{ii}}$ for all $i \in \{1, \ldots, n\}$.
for $i = 1, \ldots, n$ **do**
    if $\frac{1}{\overline{M}_{ii}} \left(1 - \frac{\overline{P}_{ii}}{\overline{M}_{ii}^2} \widehat{\text{var}}(\widehat{M}_{ii}) + \frac{1}{\overline{M}_{ii}} \widehat{\text{cov}} \left(\widehat{P}_{ii}, \widehat{M}_{ii} \right) \right) \leq \frac{n}{n-1}$ **then**
        Generate $\tilde{\boldsymbol{y}}(i) \in \mathbb{R}^n$, where $\tilde{\boldsymbol{y}}(i)_{i' \neq i} = 0$, $\tilde{\boldsymbol{y}}(i)_{i'=i} = 1$.
        Compute the fitted values $\widehat{\tilde{\boldsymbol{y}}}(i)$ of a regression of $\tilde{\boldsymbol{y}}(i)$ on X.
        Get leverage $P_{ii} = \widehat{\tilde{\boldsymbol{y}}}(i)_{i'=i}$.
        Get $1/M_{ii} = 1/(1 - P_{ii})$.
    end if
end for

C Additional details on computation of leave-cluster-out variance estimate

The goal after estimating $\overline{\mathbf{M}}^S_{gg}$ is to get the leave-cluster-out residuals. Here we show how to avoid doing unnecessary matrix inversions after doing the Cholesky decomposition.

The leave-cluster-out residuals for cluster g are:

$$\widehat{\boldsymbol{\varepsilon}}^{LC}_g = \left(\overline{\mathbf{M}}^S_{gg}\right)^{-1}\widehat{\boldsymbol{\varepsilon}}_g = \mathbf{L}^T_{gg}\left(\widehat{\mathbf{M}}_{gg}\right)^{-1}\mathbf{L}_{gg}\widehat{\boldsymbol{\varepsilon}}_g \quad \Longleftrightarrow \quad \widehat{\mathbf{M}}_{gg}\left(\mathbf{L}^T_{gg}\right)^{-1}\widehat{\boldsymbol{\varepsilon}}^{LC}_g = \mathbf{L}_{gg}\widehat{\boldsymbol{\varepsilon}}_g.$$

We can then find $\boldsymbol{z} \equiv \left(\mathbf{L}^T_{gg}\right)^{-1}\widehat{\boldsymbol{\varepsilon}}^{LC}_g$ that solve $\widehat{\mathbf{M}}_{gg}\boldsymbol{z} = \mathbf{L}_{gg}\widehat{\boldsymbol{\varepsilon}}_g$. This is much more efficient than inverting $\widehat{\mathbf{M}}_{gg}$ directly. Finally, we get $\widehat{\boldsymbol{\varepsilon}}^{LC}_g = \mathbf{L}^T_{gg}\boldsymbol{z}$.

The following algorithm presents the steps to estimate the different diagonal blocks $\widehat{\mathbf{V}}_{gg}$ as well as the matrices to compute the bootstrap residuals for each cluster, which we denote $\mathbf{B}_{gg+}$ and $\mathbf{B}_{gg-}$.

Algorithm 3 Estimate $\widehat{\mathbf{M}}^S_{gg}$. Compute $\widehat{\mathbf{V}}_{gg}$, $\mathbf{B}_{gg+}$, and $\mathbf{B}_{gg-}$

1: For all $g = 1 \dots G$, $\widehat{\boldsymbol{\varepsilon}}_g$ and $\boldsymbol{y}_g$ are the observations of $\widehat{\boldsymbol{\varepsilon}}$ and $\boldsymbol{y}$ corresponding to cluster g.
2: For all $g = 1 \dots G$, $\boldsymbol{z}^{(0)}_{P,g} = \mathbf{0}$, $\boldsymbol{z}^{(0)}_{M,g} = \mathbf{0}$ are matrices of dimensions $n_g \times n_g$.
3: **for** $j = 1, \dots, J_M$ **do**
4: Simulate a vector $\boldsymbol{r}$ of length n of mutually independent Rademacher entries.
5: Compute fitted values $\widehat{\boldsymbol{r}}$ from a regression of $\boldsymbol{r}$ on $\mathbf{X}$.
6: For all $g = 1 \dots G$, compute $\boldsymbol{z}^{(j)}_{P,g} = \boldsymbol{z}^{(j-1)}_{P,g} + \widehat{\boldsymbol{r}}_g\widehat{\boldsymbol{r}}^T_g$.
7: For all $g = 1 \dots G$, compute $\boldsymbol{z}^{(j)}_{M,g} = \boldsymbol{z}^{(j-1)}_{M,g} + \left(\boldsymbol{r}_g - \widehat{\boldsymbol{r}}_g\right)\left(\boldsymbol{r}_g - \widehat{\boldsymbol{r}}_g\right)^T$.
8: **end for**
9: **for** $g = 1, \dots, G$ **do**
10: Compute $\widehat{\mathbf{P}}_{gg} = \boldsymbol{z}^{(J_M)}_{P,g}/J_M$ and $\widehat{\mathbf{M}}_{gg} = \boldsymbol{z}^{(J_M)}_{M,g}/J_M$.
11: Get $\mathbf{L}_{gg}$ via Cholesky decomposition such that $\mathbf{L}_{gg}\mathbf{L}^T_{gg} = \widehat{\mathbf{P}}_{gg} + \widehat{\mathbf{M}}_{gg}$.
12: Get $\boldsymbol{z}$ such that $\widehat{\mathbf{M}}_{gg}\boldsymbol{z} = \mathbf{L}_{gg}\widehat{\boldsymbol{\varepsilon}}_g$.
13: $\widehat{\boldsymbol{\varepsilon}}^{LC}_g = \mathbf{L}^T_{gg}\boldsymbol{z}$.
14: Compute $\widehat{\mathbf{V}}_{gg} = \frac{1}{2}\left(\boldsymbol{y}_g\left(\widehat{\boldsymbol{\varepsilon}}^{LC}_g\right)^T + \widehat{\boldsymbol{\varepsilon}}^{LC}_g\boldsymbol{y}^T_g\right)$.
15: Get $\mathbf{B}_{gg+}$, and $\mathbf{B}_{gg-}$ using Steps 2 to 4 of Algorithm 1.
16: **end for**

D Sample selection for leave-one-out and leave-match-out cases

To use the leave-one-out variance estimate requires that $P_{ii} < 1$, or equivalently $M_{ii} > 0$. To use the leave-match-out variance estimate, which in the AKM case corresponds to allowing unrestricted correlation of errors within a worker-firm match, requires that M_{gg} is not singular. We explain here how to use standard tools from graph theory to select the sample such that the above restrictions are satisfied for each case.

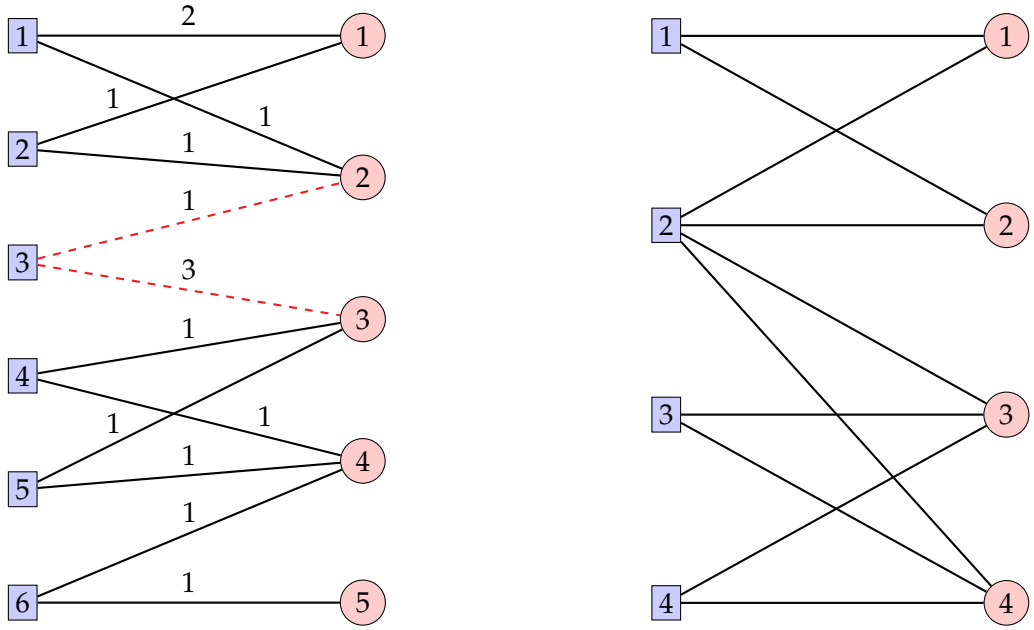
To clarify the arguments below, let us first represent the worker-firm data as a bipartite graph. Figure 2a represents an example with six workers and five firms. The vertices to the left represent workers, and the vertices to the right, firms. The links between vertices, denoted edges, represent matches between workers and firms. Importantly, each of these edges can have different weights corresponding to the number of observations that we have for each worker-firm match. For example, the edge connecting worker 3 with firm 3 has a weight of 3, meaning we have 3 observations (or periods) where we observe worker 3 employed in firm 3.

Connected set. The first thing to do is to restrict our analysis to a connected component of the sample. This is a well known requirement (Abowd, Creecy, and Kramarz 2002; Card, Heining, and Kline 2013; Jochmans and Weidner 2019) to ensure the rank condition is satisfied and that we can compare all of our fixed effects estimates. In practice, we keep the largest connected set. The examples depicted in Figure 2a are already connected graphs.

Leave-one-out sample. The leave-one-out requirement is stronger. It requires that we can identify all of the parameters after removing any one observation. If an observation has a leverage of one it means that the identification of a parameter in the model depends entirely on that observation. In terms of the worker-firm network this means that: first, deleting an edge corresponding to that observation would disconnect the network; and second, the weight of that edge is equal to one.

An edge whose deletion would disconnect the network is known as a *bridge* or a cut-edge. In Figure 2a, the dashed lines representing the edges connecting worker 3 with firms 2 and 3 are bridges: removing one of them would disconnect the network. However, only the observation corresponding to the edge connecting worker 3 and firm 2 would have a leverage of 1. If we

Figure 2: Bipartite graphs of workers and firms



Notes: Workers are represented by square vertices to the left of each graph and firms are represented by circle vertices in the right of each graph. If appropriate, weights are represented by numbers above edges.

were to remove one observation corresponding to the match between worker 3 and firm 3, we would still be able to identify the parameters.

Removing the bridge with weight of 1 connecting worker 3 with firm 2 would disconnect the graph. We could then work with the largest connected subgraph, in this case the graph formed by workers 3 to 6 and firms 3 to 5. We can then check again if the remaining network has no bridges with unit weights. For this particular subgraph is easy to see that is the case. Then, this subsample would be suitable for the leave-one-out variance estimate.

Leave-match-out. Similarly, to use the leave-cluster-out variance estimate we need to restrict the sample such that the deletion of all observations corresponding to the cluster, or in this case, the match, would still allow us to identify all the parameters in the model. Naturally, all workers that were only employed by one firm are not leave-match-out estimable: all the information for the worker fixed effect is contained in the observations corresponding to that worker's unique match.

To ensure the sample allows to use the leave-match-out estimator we should restrict the sample such that the bipartite graph has no bridges. As each edge of the graph corresponds to a realized match, removing one edge corresponds to removing one match. Thus, we should remove the matches that correspond to edges that are bridges. The weight of the edge does not matter in this case as it only represents the number of observations of a given match. In the example of Figure 2a, we would remove the observations corresponding to both bridges connecting worker 3 with firms 2 and 3. The resulting sample would have data on workers 4 to 6 and firms 3 to 5. This sample is more restricted than the leave-one-out sample, which is an expected result as the leave-match-out sample restriction is stronger.

Leave-worker-out. KSS propose an algorithm that makes the sample suitable to use the leave-match-out estimator, and therefore also suitable for the leave-one-out estimator. KSS remove from the sample all those workers that are cut vertices or *articulation points*. This means workers whose deletion would disconnect the graph. We name this procedure as *leave-worker-out*.

Worker 3 in Figure 2a constitutes an articulation point. Their removal would disconnect the graph and lead to the same leave-match-out subsample that we would obtain with the bridge deleting procedure explained above. However, using leave-worker-out is a stronger requirement than the leave-match-out procedure. Consider Figure 2b. The graph has no bridges, therefore it is leave-match-out estimable. However, worker 2 constitutes an articulation point. If we were to follow the leave-worker-out procedure of KSS we would remove *all* observations corresponding to worker 2. We would probably work with the data corresponding to workers 3 and 4, and firms 3 and 4 as well. This is a much smaller subsample compared to the original sample.

When we compare the performance of our method with respect to KSS we use the leave-worker-out procedure to use the same samples and to make the methods as comparable as possible.

Online Appendix

A Sample construction

The data source *BTS* is a repeated cross section with the universe of jobs per year. The data records with a yearly worker identifier all the jobs of a worker in a given year and in the previous year. That is, one cannot directly create a panel of workers as the worker identifiers change every year. The data has information on age, a firm and establishment identifiers, main job, occupation, gender and the municipality of the establishment. Babet et al. (2022) overcome the yearly changing identifiers by leveraging that each identifier has t and $t-1$ information. They proposed a way to match on additional information across years and generously made the code public. We directly use their code to generate the yearly sample with pseudo identifiers that are generated to create a panel version.

We make additional sample restrictions. We focus on main jobs of workers working full time at the private sector with positive hourly wages, with occupation, location and age information. We yearly trim the 0.05% of the hourly wages from above. The source variables we use are:

- *SIREN*: is the firm identifier.
- *NIC*: combined with *SIREN* gives the establishment identifier.
- *SBRUT*: gross yearly earnings in the job. We keep observations with positive earning information.
- *NBHEUR*: total yearly hours in the job. Hourly wages in the job are defined as *SBRUT*/*NBHEUR*. We keep observations with positive hourly wages.
- *PPS*: indicator of main job or *poste principal* by keeping observations with *PPS* equal to 1.
- *IRNBHEUR*: we drop observations with imputed hours by keeping only IRNBHEUR equal to D.
- *PCS* and *CS*: *PCS* is a 4-digit occupation classification that is well maintained starting in 2009. In 2008 and before, the 2-digit classification *CS* was well maintained. We therefore take *CS* before 2009 and take the first 2 digits of *PCS* from 2009 onward.

- *COMT*: is the municipality identifier. We match the municipality codes to the commuting zone classification in 2020 *ZEMP2020*.
- *AGE*: age of the worker in year t. We keep observations with age information.
- *DOMEMP*: is a variable that can be used to restrict to workers in the private sector. In 2008 and before, private workers are those with *DOMEMP* equal to 1, 6 or 9. From 2009 on one needs to keep the codes 6, 7, 8 and 9.
- *CPFD*: is a variable used to keep full time employees when *CPFD* equals C.