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

Miren Azkarate-Askasua† Miguel Zerecero‡

October 18, 2024

Abstract

We propose a bootstrap method to correct limited mobility bias in the variance components of AKM models. Our method handles multiple corrections without increasing computational cost and works with any symmetric estimator of the error covariance matrix, including the one from Kline, Saggio, and Sølvsten (2020). Monte Carlo simulations show our method corrects the bias and is faster than alternative methods. Using French administrative data, we apply our method in four ways: (i) a full-sample variance decomposition of log wages under different assumptions on the error structure, (ii) correcting thousands of labor markets to study the relationship between market size and worker-firm or worker-coworker sorting, (iii) analyzing gender differences and (iv) life cycle patterns in wage components. In all cases, the corrections are important to interpret the results.

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, and Kirill Borusyak, Damon Clark, Fabrice Collard, Thomas Crossley, Yingying Dong, Patrick Fève, Matt Freedman, Simen Gaure, Silvia Goncalves, Cristina Gualdani, Koen Jochmans, Elia Lapenta, Tim Lee, Ying-Ying Lee, Jack Liebersohn, Thierry Magnac, Nour Meddahi, David Neumark, Matt Notowidigdo, Jean-Marc Robin, Uta Schönberg, and Mikkel Sølvsten for helpful comments. We thank Vsevolod Spirin for research assistance. We acknowledge initial financial support from TSE. This work is supported by a public grant by the French National Research Agency (ANR) as part of the 'Investissements d'avenir' program (reference: ANR-10-EQPX-17 – CASD). Zerecero acknowledges the support from CONACYT (reference: 329103/383874). Azkarate-Askasua acknowledges the support from the German Research Foundation (CRC-TR-224, B06). All errors are our own.

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

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

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:

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

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

A common exercise in labor economics is to do a variance decomposition of (1) plugging in the OLS estimator of the fixed effects. But even if the fixed effects estimator is unbiased, quadratic objects in the estimated parameters, such as the elements of a variance decomposition, suffer from small sample bias (Andrews, Gill, Schank, and Upward, 2008). In the AKM context, Abowd, Kramarz, Lengermann, and Pérez-Duarte (2004) dubbed the bias of these quadratic objects *"limited mobility bias"* as having few movers 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 that the limited mobility bias is systematically large, and it can change the economic interpretation of the results.

In previous work, Andrews et al. (2008) derive formulas for correcting the bias when the errors are homoscedastic, and 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 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 estimators of the covariance matrix of the errors, including the leave-

[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), Bartolucci, Devicienti, and Monzón (2018), 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.

one-out and leave-cluster-out estimators 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), by city (Dauth, Findeisen, Moretti, and Suedekum, 2022), or by occupation (Heath Milsom and Hou, 2024).[2] 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. 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.

Using French administrative data, we present four applications that highlight two key advantages of our bootstrap correction: (i) its ability to accommodate different assumptions about the error covariance matrix, and (ii) its scalability in handling corrections across different subgroups, while using the full sample to estimate the model.

First, we do a basic AKM variance decomposition under different assumptions for the covariance matrix of errors. We show that clustering at various levels—observation, match, or worker-occupation—produces similar results.

Second, we study sorting patterns between workers and firms in labor markets, defined as the intersection of occupation and commuting zone, resulting in over 5,000 markets in our sample. Our method corrects for each market while using the full sample for estimation. In contrast, previous studies estimate and correct for subgroups separately, losing information on workers who move between them.[3] We use these corrected estimates to revisit the idea in urban economics that larger labor markets lead to better worker-firm sorting. Our findings suggest

[2] Heath Milsom and Hou (2024) also do decompositions by city.

[3] For example, Dauth et al. (2022) estimate an AKM model for each city in Germany and do a correction for each city. Pérez, Meléndez, and Nuno-Ledesma (2023) do the same with data from Mexico. However, as Leknes et al. (2022) point out, these approaches lose valuable information on workers who move between cities. One exception is Heath Milsom and Hou (2024), who adapt the *pytwoway* function of Bonhomme et al. (2023) to implement the KSS correction. They estimate the model using the full sample and then loop over cities or occupations to apply the corrections. This increases the computational cost linearly with the number of corrections.

that sorting is indeed stronger in larger locations, as evidenced by a higher correlation between worker and firm fixed effects in larger labor markets. We find that this positive relationship is not driven by systematic bias differences across markets. In fact, after correcting for limited mobility bias, the positive relationship becomes even more pronounced, with the slope doubling or tripling in magnitude.

We also study the intensity of sorting in larger markets, measured by the correlation between worker fixed effects and the average fixed effects of coworkers (Lopes de Melo, 2018). Using uncorrected estimates, larger markets appear to have higher sorting intensity. However, after correction, the relationship reverses, suggesting workers are not as segregated by worker fixed effects in larger markets.

In the third application, we compare AKM decompositions by gender. Corrected estimates show a higher correlation between worker and firm fixed effects for women than for men, consistent across two periods: 2009-2014 and 2015-2019. Naive estimates, however, show a slightly higher correlation for men in the second period. This suggests that the assumption of a constant bias over different groups or time—like in Song et al. (2019)—may be inaccurate.

Lastly, we explore how AKM variance components evolve across the life cycle by correcting for each age group. We find that the log wage variance is increasing with age, mainly driven by an increase of the variance of worker fixed effects over the life cycle.

Literature. Our paper contributes to the small body of literature focused on accurately estimating quadratic forms. A key related work is KSS, which also presents an iterative method to correct the bias in quadratic forms. Their method includes clustering at the observation level or at the worker-firm match level. We extend this approach by allowing clustering at any level, as long as all model parameters remain identifiable when any single cluster is removed. This sample restriction is also required by KSS. We provide a detailed comparison between our bootstrap method and the KSS approach in Section 4, where we discuss the strengths of both methods.

Inspired by KSS, Babet, Godechot, and Palladino (2022) propose a modified split-sample correction method similar to the split-panel jackknife by Dhaene and Jochmans (2015). These methods are effective at reducing bias and are very fast, requiring only a few model estimations. In fact, Babet et al. (2022) show that their approach completely eliminates bias when the covariance matrix is diagonal. However, they do not explore the effects of their method

when the covariance matrix is not diagonal. Still, the results from Dhaene and Jochmans (2015) suggest that the split-sample method could also reduce bias in such cases.[4] Additionally, the methods from Babet et al. (2022) and Dhaene and Jochmans (2015) may become impractical when correcting for multiple subsamples simultaneously.

Since the bias arises from noisy estimates, Bonhomme, Lamadon, and Manresa (2019) suggest grouping worker and firm fixed effects to obtain more precise grouped estimates, which can then be used to calculate the quadratic forms. While this approach for estimating quadratic forms across the entire sample may be reasonable, it may become less effective if grouping fixed effects reduces heterogeneity within the subsamples of interest, especially when the number of subsamples far exceeds the number of fixed effects groups.

The paper is organized as follows. Section 1, derives the bias. Section 2 presents the bootstrap correction. Section 3 discusses the practical considerations when using unbiased estimators of the errors covariance matrix. Section 4 compares our method with the one developed by KSS. Section 5 presents the applications with the French data. Section 6 concludes.

1 The bias

Suppose we have 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:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2}$$

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 estimator of $\boldsymbol{\beta}$. We can now define an estimator of φ.

Definition 1 (Plug-in Estimator)**.** *The plug-in estimator of the quadratic form is:*

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

[4] Babet et al. (2022) do not explore in detail the properties of their split-sample estimator, as it is not the main focus of their paper. Extending their approach to non-diagonal covariance matrices and studying its properties could be a valuable direction for future research, given the practicality of the method.

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

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

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:*

$$\delta \equiv \mathbb{E}\left(\widehat{\boldsymbol{\beta}}^T\mathbf{A}\widehat{\boldsymbol{\beta}}\,\middle|\,\mathbf{X}\right) - \boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta} = \mathbf{tr}\left(\mathbf{S}_\mathbf{X}^T\mathbf{A}\mathbf{S}_\mathbf{X}\mathbb{V}\left(\boldsymbol{\varepsilon}\,|\,\mathbf{X}\right)\right). \tag{3}$$

Computing δ is infeasible as we do not know $\mathbb{V}\left(\boldsymbol{\varepsilon}\,|\,\mathbf{X}\right)$. Therefore, let $\widehat{\mathbf{V}}$ be an estimator of the covariance matrix $\mathbb{V}\left(\boldsymbol{\varepsilon}\,|\,\mathbf{X}\right)$. We can now define a bias correction and a bias corrected estimator of the quadratic form.

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

$$\widehat{\delta}_D \equiv \mathbf{tr}\left(\mathbf{S}_\mathbf{X}^T\mathbf{A}\mathbf{S}_\mathbf{X}\widehat{\mathbf{V}}\right). \tag{4}$$

Definition 4 (Bias corrected estimator)**.** *The direct bias corrected estimator of the quadratic form is*

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

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

Proposition 1 (Unbiasedness of $\widehat{\delta}_D$)**.** $\mathbb{E}\left(\widehat{\delta}_D\,\middle|\,\mathbf{X}\right) = \delta$ *if and only if* $\mathbb{E}\left(\widehat{\mathbb{V}}\,\middle|\,\mathbf{X}\right) = \mathbb{V}\left(\boldsymbol{\varepsilon}\,|\,\mathbf{X}\right)$.

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

Corollary 1 (Unbiasedness of $\widehat{\varphi}$)**.** $\mathbb{E}\left(\widehat{\varphi}\,|\,\mathbf{X}\right) = \varphi$ *if and only if* $\mathbb{E}\left(\widehat{\mathbb{V}}\,\middle|\,\mathbf{X}\right) = \mathbb{V}\left(\boldsymbol{\varepsilon}\,|\,\mathbf{X}\right)$.

For the case where the number of covariates grows with sample size, KSS show conditions for the consistency of $\widehat{\varphi}$ using a diagonal covariance matrix estimator.[5]

[5]See Assumption 1 and Lemma 3 in KSS.

2 Bootstrap correction

The direct bias correction $\widehat{\delta}_D$ is computationally infeasible in typical applications of AKM decompositions involving millions of fixed effects, as it requires inverting $\mathbf{X}^T\mathbf{X}$. To address this challenge, we propose a bootstrap-based estimation of $\widehat{\delta}_D$, which replicates the bias structure of the plug-in estimator, making the computation tractable.

To motivate the use of our bootstrap, first note that the 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 $\boldsymbol{v}^*$ be a random vector. Assume $\mathbb{E}\left(\boldsymbol{v}^* \mid \mathbf{X}, \boldsymbol{\varepsilon}\right) = \mathbf{0}$ and $\mathbb{V}\left(\boldsymbol{v}^* \mid \mathbf{X}, \boldsymbol{\varepsilon}\right) = \widehat{\mathbf{V}}$. Let $\widehat{\boldsymbol{\beta}^*}$ be the OLS estimator of the regression coefficients in regressing $\boldsymbol{v}^*$ 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:*

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

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

For this bootstrap to work, 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:

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

We simulate $\boldsymbol{r}$ from a 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 estimators of the covariance matrix, as Proposition 1 already imposes restrictions on $\widehat{\mathbf{V}}$ to get an unbiased estimator of δ: $\widehat{\mathbf{V}}$ should be an unbiased estimator of $\mathbb{V}(\boldsymbol{\varepsilon} \mid \mathbf{X})$. Recently, Kline et al. (2020), Anatolyev (2021), and Jochmans (2022) propose unbiased covariance matrix estimators that are robust to heteroscedasticity, 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:

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

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

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

where the ith diagonal terms of $\boldsymbol{\Lambda}_+$ and $\boldsymbol{\Lambda}_-$, 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:

$$\widehat{\mathbf{V}} = \mathbf{Q}\left(\boldsymbol{\Lambda}_+ - \boldsymbol{\Lambda}_-\right)\mathbf{Q}^T = \underbrace{\mathbf{Q}\boldsymbol{\Lambda}_+\mathbf{Q}^T}_{\widehat{\mathbf{V}}_+} - \underbrace{\mathbf{Q}\boldsymbol{\Lambda}_-\mathbf{Q}^T}_{\widehat{\mathbf{V}}_-}, \tag{5}$$

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 $\widehat{\delta}_D$)**.** *Let $\widehat{\boldsymbol{\beta}^*}_+$ and $\widehat{\boldsymbol{\beta}^*}_-$ be the OLS estimator of the regression coefficients in regressing $\boldsymbol{v}_+^*$ and $\boldsymbol{v}_-^*$ on $\mathbf{X}$. Then,*

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

The last proposition motivates the following bootstrap estimator for any covariance matrix estimator, 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 $j = 1 \ldots J$. Also, let $\widehat{\boldsymbol{\beta}^*}_+(j)$ and $\widehat{\boldsymbol{\beta}^*}_-(j)$ be the OLS estimator of the regression coefficients in regressing $\boldsymbol{v}_+^*(j)$ and $\boldsymbol{v}_-^*(j)$ on $\mathbf{X}$. Then, the bootstrap bias correction is defined as:*

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

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:*

1. *Unbiased:* $\mathbb{E}\left(\delta^* \,|\, \mathbf{X}, \boldsymbol{\varepsilon}\right) = \widehat{\delta}_D$.
2. *Consistent: Fix n and k, and let $J \longrightarrow \infty$, then $\delta^* \xrightarrow{a.s.} \widehat{\delta}_D$.*

The last proposition means that we can estimate the direct bias correction $\widehat{\delta}_D$ to arbitrary precision, and implies the following result.

Corollary 2. $\mathbb{E}(\delta^* \mid \mathbf{X}) = \delta$ *if and only if* $\mathbb{E}\left(\widehat{\mathbf{V}} \,\middle|\, \mathbf{X}\right) = \mathbb{V}(\boldsymbol{\varepsilon} \mid \mathbf{X})$.

The main computational cost of our method is to compute $\widehat{\boldsymbol{\beta}^*}_+$ and $\widehat{\boldsymbol{\beta}^*}_-$, not the number of quadratic forms to correct. In other words, if we would like to estimate 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 $\boldsymbol{\Lambda}$ such that $\widehat{\mathbf{V}} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T$.
3: Decompose $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}_+ - \boldsymbol{\Lambda}_-$, with $\boldsymbol{\Lambda}_+$ having the positive eigenvalues and $\boldsymbol{\Lambda}_-$ the absolute value of the negative eigenvalues.
4: Get $\mathbf{B}_+ = \mathbf{Q}(\boldsymbol{\Lambda}_+)^{1/2}$ and $\mathbf{B}_- = \mathbf{Q}(\boldsymbol{\Lambda}_-)^{1/2}$.
5: **for** $j = 1, \dots, J$ **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 $m = 1 \dots M$.
10: **end for**
11: Compute $\delta^*_m = \frac{1}{J}\sum_{j=1}^{J} \delta_m^{*(j)}$ for all $m = 1 \dots M$.

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.[6]

[6] 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.

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.
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*: it mostly 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 estimator.

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. 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, locations, etcetera.

The bootstrap correction is *flexible* compared to KSS: their 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) propose a similar bootstrap to correct for flat biases like the one considered here.[7] For simplicity, let us abstract about the decomposition of the variance estimator as the difference of two positive semi-definite matrices, but all arguments follow easily in that case. In other words, let us have that $\widehat{\mathbf{V}} = \widehat{\mathbf{V}}_+$. MacKinnon and Smith Jr (1998) 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, $\mathbb{V}\left(\delta^*_{MS} \middle| \mathbf{X}, \boldsymbol{\varepsilon}\right) \geq \mathbb{V}\left(\delta^* | \mathbf{X}, \boldsymbol{\varepsilon}\right)$.*

Given that we use independent Rademacher entries $\boldsymbol{r}$ to form $\boldsymbol{v}^* = \mathbf{B}\boldsymbol{r}$, then the conditions $\mathbb{E}(\boldsymbol{v}^* | \mathbf{X}, \boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathbb{E}\left((\boldsymbol{v}^*)^3 \middle| \mathbf{X}, \boldsymbol{\varepsilon}\right) = \mathbf{0}$ 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 estimator for the variance is unbiased, this means that our bootstrap estimator is more efficient than the more traditional one as proposed by MacKinnon and Smith Jr (1998).[8]

[7] Recall that a flat bias is one that does not depend on the parameters. The bias from the quadratic forms is flat because the trace term in (3) is independent of $\boldsymbol{\beta}$.

[8] The proposition only shows this for the case where the covariance matrix estimator is diagonal. We conjecture this is also the case with clustered errors.

2.1 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) the fact that the bias is a linear function of the covariance matrix. These two properties allow us to express the original bias, which is equal to a trace, as the difference of two traces.

To estimate these two traces, we first need to compute the matrices $\mathbf{B}_+$ and $\mathbf{B}_-$. Below, we present three different examples of unbiased estimators of $\mathbf{V}$ based on different assumptions about the error term, and we discuss the corresponding $\mathbf{B}_+$ and $\mathbf{B}_-$ matrices for each case.

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

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

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 estimator is unbiased with respect to the covariance matrix of the unobserved errors.

This covariance matrix estimator is positive semi-definite: 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 estimator: KSS use a diagonal covariance matrix estimator which is unbiased when the errors are heteroscedastic. The diagonal entries are:

$$\widehat{\mathbf{V}}_{ii} = \frac{\boldsymbol{y}_i\widehat{\boldsymbol{\varepsilon}}_i}{1-P_{ii}}, \tag{6}$$

where P_{ii} is the leverage of observation i, defined as the ith diagonal entry 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 estimator: This is a generalization of the leave-one-out covariance matrix estimator. 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 $\mathbf{M}_{gg} \equiv \mathbf{I}_{n_g} - \mathbf{P}_{gg}$, 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 estimator of variance for the gth diagonal block of $\widehat{\mathbf{V}}$ is:

$$\widehat{\mathbf{V}}_{gg} = \frac{1}{2} \left(\boldsymbol{y}_g \left(\widehat{\boldsymbol{\varepsilon}}_g^{LC} \right)^T + \widehat{\boldsymbol{\varepsilon}}_g^{LC} \boldsymbol{y}_g^T \right). \tag{7}$$

Clearly, the leave-cluster-out variance estimator is a generalization of the leave-one-out estimator 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 estimator of the gth diagonal block of the covariance matrix $\mathbb{V}(\boldsymbol{\varepsilon} \,|\, \mathbf{X})$.

Since the matrix $\widehat{\mathbf{V}}$ is block diagonal, we can do the spectral decomposition (5) for each diagonal block $\widehat{\mathbf{V}}_{gg}$, corresponding to the different clusters g, thereby simplifying the computation. This allows us to compute each block of $\mathbf{B}_+$ and $\mathbf{B}_-$ separately for each cluster g. Moreover, $\widehat{\mathbf{V}}_{gg}$ is determined by the outer product $\boldsymbol{y}_g \left(\widehat{\boldsymbol{\varepsilon}}_g^{LC} \right)^T$, and since these two vectors span a two-dimensional subspace, the matrix $\widehat{\mathbf{V}}_{gg}$ has at most two non-zero eigenvalues. Therefore, we

only need to compute two eigenvalues to do the spectral decomposition for each cluster. Also, as shown in the proposition below, there is a closed-form expression for $\mathbf{B}_+$ and $\mathbf{B}_-$ for each cluster, simplifying their computation further.

Proposition 6 (Decomposition of leave-cluster-out covariance matrix)**.** *Let $\mathbf{B}_{gg+}$ and $\mathbf{B}_{gg-}$ be the gth diagonal block of $\mathbf{B}_+$ and $\mathbf{B}_-$, respectively, each of dimensions $n_g \times n_g$. Also, let $(\boldsymbol{v} \cdot \boldsymbol{w})$ be the inner product of vectors $\boldsymbol{v}$ and $\boldsymbol{w}$. Then, given vectors $\boldsymbol{y}_g$ and $\widehat{\boldsymbol{\varepsilon}}_g^{LC}$, we have that:*

$$\mathbf{B}_{gg+} = \sqrt{\lambda_{g+}} \times \mathbf{Q}_{gg+}, \quad \mathbf{B}_{gg-} = \sqrt{|\lambda_{g-}|} \times \mathbf{Q}_{gg-},$$

where the non-zero eigenvalues of $\widehat{\mathbf{V}}_{gg}$ are

$$\lambda_{g+} = \frac{1}{2}\left((\widehat{\boldsymbol{\varepsilon}}_g^{LC} \cdot \boldsymbol{y}_g) + \sqrt{(\widehat{\boldsymbol{\varepsilon}}_g^{LC} \cdot \widehat{\boldsymbol{\varepsilon}}_g^{LC})(\boldsymbol{y}_g \cdot \boldsymbol{y}_g)}\right), \quad \lambda_{g-} = \frac{1}{2}\left((\widehat{\boldsymbol{\varepsilon}}_g^{LC} \cdot \boldsymbol{y}_g) - \sqrt{(\widehat{\boldsymbol{\varepsilon}}_g^{LC} \cdot \widehat{\boldsymbol{\varepsilon}}_g^{LC})(\boldsymbol{y}_g \cdot \boldsymbol{y}_g)}\right),$$

and the matrices $\mathbf{Q}_{gg+}$ and $\mathbf{Q}_{gg-}$ have all columns equal to zero except their first columns, which are the orthonormal eigenvectors of $\widehat{\mathbf{V}}_{gg}$ that correspond to the eigenvalues λ_{g+} and λ_{g-}, respectively. These orthonormal eigenvectors are equal to the following eigenvectors, but normalized:

$$\boldsymbol{q}_{g+} = a\boldsymbol{y}_g + \widehat{\boldsymbol{\varepsilon}}_g^{LC}, \quad \boldsymbol{q}_{g-} = -a\boldsymbol{y}_g + \widehat{\boldsymbol{\varepsilon}}_g^{LC},$$

where

$$a \equiv \sqrt{\frac{(\boldsymbol{y}_g \cdot \boldsymbol{y}_g)}{(\widehat{\boldsymbol{\varepsilon}}_g^{LC} \cdot \widehat{\boldsymbol{\varepsilon}}_g^{LC})}}.$$

3 Practical details when implementing the bootstrap

In this section, we discuss some practical details to implement the bootstrap method in three cases. Each of these cases uses a different unbiased estimator of the covariance matrix, suitable for different situations: the standard covariance estimator for homoscedastic errors; the leave-one-out estimator of Jochmans (2022) and KSS; and the leave-cluster-out estimator 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 do 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, taking the average across the estimated quadratic forms gives the bias estimate.

Case 2—Heteroscedastic errors: The leave-one-out covariance matrix estimator is a diagonal covariance matrix with entries described by equation (6) in Example 2 above. As some of these entries are negative, we 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: (i) estimating the leverage of each observation, and (ii) guaranteeing that the leverage of each observation is below 1 such that the variance estimator of observation i $\widehat{\mathbf{V}}_{ii}$ exists. We discuss their implementation below.

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

In the following we discuss the sample selection required for the existence of the leave-one-out and leave-cluster-out estimators and iterative procedures to estimate the leverages. We end the section explaining how to add extra covariates to the linear model while keeping the computational advantages of working with just two sets of fixed effects.

3.1 Existence of variance estimators: leave-one-out and leave-cluster-out

A key practical consideration in the leave-one-out and leave-cluster-out covariance estimation is ensuring that the estimators exist. This requires selecting a subsample from 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, as discussed by 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 as that observation completely pins down the worker fixed effect; and (ii) observations that upon removing them would leave some firms out of the connected set.[9]

Therefore, using the leave-one-out variance estimator 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.*[10] The leave-one-out connected set ensures that after removing a single observation all the parameters from the model are identified.

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. The idea is similar to above: the sample is leave-cluster-out estimable if upon removing one cluster from the sample all of the parameters are still identified. To gain intuition of the different types of sample restrictions needed in this case, we focus here on the clustering of errors at the match level. 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.

Sample selection algorithms. We provide algorithms to identify samples that are leave-cluster-out estimable for any type of cluster, generalizing the type of clustering considered by KSS. For cases where the clusters are specific to either firms or workers, we develop a fast sample selection algorithm using graph theory tools. This includes clustering at the observation or match

[9]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.

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

level, as well as at the firm-occupation or worker-occupation level. Details are discussed in Online Appendix A.

3.2 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 of $\mathbf{X}^T\mathbf{X}$ and get an estimator of the leverages.[11]

Proposition 7 (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 the sample average gives us an estimator of the leverages P_{ii} and $1 - P_{ii}$, as we define more precisely below.

Definition 6 (Estimators of P_{ii} and M_{ii})**.** *Let $\boldsymbol{r}(j)$ be a random vector that corresponds to the jth iteration and $\widehat{\boldsymbol{r}}(j)$ the fitted value of running a regression of $\boldsymbol{r}(j)$ on $\mathbf{X}$. Similarly, define $r_i(j)$ and $\widehat{r}_i(j)$ as the elements of $\boldsymbol{r}(j)$ and $\widehat{\boldsymbol{r}}(j)$ that correspond to observation i. Then, the estimators $\widehat{P}_{ii}$ and $\widehat{M}_{ii}$ are:*

$$\widehat{P}_{ii} \equiv \frac{1}{J}\sum_{j=1}^{J} \left(\widehat{r}_i(j)\right)^2 \quad \textit{and} \quad \widehat{M}_{ii} \equiv \frac{1}{J}\sum_{j=1}^{J} \left(r_i(j) - \widehat{r}_i(j)\right)^2.$$

While $\widehat{P}_{ii}$ and $\widehat{M}_{ii}$ are consistent estimators of P_{ii} and M_{ii}, they could still have values that do not lie between 0 and 1. To avoid that, we follow Kline, Saggio, and Sølvsten (2021) and define the following estimators:

[11] Kline et al. (2020) do the same procedure to estimate the leverages.

Definition 7. *The estimators $\overline{P}_{ii}$ and $\overline{M}_{ii}$ are:*

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

As it is clear from above, these estimators 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 $\overline{P}_{ii}$ and $\overline{M}_{ii}$ are always between 0 and 1.

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

3.3 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 8 (Approximation of Diagonal Blocks of **P** and **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 7, this result tell us we can define the following estimators of $\mathbf{P}_{gg}$ and $\mathbf{M}_{gg}$:

Definition 8 (Estimates block-diagonals of **P** and **M**)**.** *Let $\boldsymbol{r}(j)$ be a random vector that corresponds to the jth iteration 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 estimators $\widehat{\mathbf{P}}_{gg}$ and $\widehat{\mathbf{M}}_{gg}$ are defined as:*

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

The definitions for $\widehat{P}_{ii}$ and $\widehat{M}_{ii}$ are special cases of $\widehat{\mathbf{P}}_{gg}$ and $\widehat{\mathbf{M}}_{gg}$ when clusters have only one observation, i.e. $n_g = 1$ for all g.

The requirement for the existence of the leave-cluster-out variance estimator is that $\mathbf{M}_{gg}$ is non-singular. As the proposition below shows, the estimator $\widehat{\mathbf{M}}_{gg}$ inherits the singularity of $\mathbf{M}_{gg}$. So we should select the sample to avoid this singularity cases before attempting to estimate $\mathbf{M}_{gg}$.

Proposition 9 (Singularity of $\widehat{\mathbf{M}}_{gg}$)**.** *If $\mathbf{M}_{gg}$ is singular, then $\widehat{\mathbf{M}}_{gg}$ is singular.*

Just as the leverages P_{ii} are trivially symmetric (the leverage is a scalar) and their values must be between 0 and 1, the block-diagonal components of $\mathbf{P}$ and $\mathbf{M}$ share similar properties: they are symmetric matrices with all their eigenvalues between 0 and 1. This is because 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.[12] 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 estimators of $\mathbf{P}_{gg}$ and $\mathbf{M}_{gg}$ are symmetric with eigenvalues between 0 and 1 by using the following estimators:

Definition 9. *The symmetric estimators of $\mathbf{P}_{gg}$ and $\mathbf{M}_{gg}$ are defined as:*

$$\overline{\mathbf{P}}^S_{gg} \equiv \mathbf{L}_{gg}^{-1}\widehat{\mathbf{P}}_{gg}\left(\mathbf{L}_{gg}^{-1}\right)^T, \quad \textit{and} \quad \overline{\mathbf{M}}^S_{gg} \equiv \mathbf{L}_{gg}^{-1}\widehat{\mathbf{M}}_{gg}\left(\mathbf{L}_{gg}^{-1}\right)^T,$$

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, $\overline{\mathbf{P}}^S_{gg} + \overline{\mathbf{M}}^S_{gg} = \mathbf{I}_{n_g}$, in line with the definition of $\mathbf{M}_{gg} \equiv \mathbf{I}_{n_g} - \mathbf{P}_{gg}$ and, as the proposition below shows, they have eigenvalues between 0 and 1.[13]

Proposition 10 (Eigenvalue properties of $\overline{\mathbf{P}}^S_{gg}$ and $\overline{\mathbf{M}}^S_{gg}$)**.** *Assume $\widehat{\mathbf{M}}_{gg}$ is non-singular. Then, the eigenvalues of $\overline{\mathbf{P}}^S_{gg}$ lie within $[0,1)$ and the eigenvalues of $\overline{\mathbf{M}}^S_{gg}$ lie within $(0,1]$.*

Using $\overline{\mathbf{M}}^S_{gg}$ we can compute the leave-cluster-out residuals $\widehat{\varepsilon}^{LC}_g$ and get $\widehat{\mathbf{V}}_{gg}$ as shown in (7). We show in Online Appendix C how to get the leave-cluster-out residuals without the explicit inversion of any matrix but by solving systems of linear equations.

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

[13] In small-scale simulations, we found that $\overline{\mathbf{M}}^S_{gg}$ is a more efficient estimator of $\mathbf{M}_{gg}$ than $\widehat{\mathbf{M}}_{gg}$.

3.4 Adding extra covariates

So far we have worked with the general linear model (2). However, the current code for the bootstrap procedure focuses on a basic specification with only two-way fixed effects and no additional covariates. This approach has significant computational advantages, which are detailed in Online Appendix E.[14]

If researchers wish to include additional covariates, they can first regress out these covariates in an initial estimation round, leaving a system that only contains the two-way fixed effects. This is the default method used by KSS. To illustrate, consider the linear model:

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

where $\mathbf{X}$ represents the matrix of dummies for the two fixed effects, and $\mathbf{W}$ contains the extra covariates. We can estimate this model using OLS, then define a new outcome variable $\widetilde{\mathbf{y}} \equiv \mathbf{y} - \mathbf{W}\widehat{\boldsymbol{\gamma}}$, and proceed with the transformed model $\widetilde{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} + \widetilde{\boldsymbol{\varepsilon}}$. This allows us to focus on the corrections related to the two fixed effects.

However, if the researcher is interested in quadratic forms involving the additional covariates, the procedure can be adjusted. For example, a labor economist might want to know how much of the total variance in log wages is explained by workers' education levels. If the dimension of $\boldsymbol{\gamma}$ is much smaller than $\boldsymbol{\beta}$, the OLS estimator of $\boldsymbol{\gamma}$ will usually be very precise in most applications. This means that any bias in the quadratic forms involving $\boldsymbol{\gamma}$ will likely be negligible and can be safely ignored.[15]

We can then compute all the quadratic objects involving $\boldsymbol{\gamma}$ from a first regression and work directly with them. In practice, all quadratic terms involving $\boldsymbol{\gamma}$ can be computed from the first regression. For the education example, this means calculating the total variance in log wages, the quadratic terms involving the education parameter $\boldsymbol{\gamma}$, and then only applying corrections to the quadratic forms involving $\boldsymbol{\beta}$ in the residualized system. Finally, these components can be combined to perform a full variance decomposition.

[14] Briefly, these advantages are: i) with only two-way fixed effects, the normal equations can be represented as a Laplacian system, for which fast algorithms exist (Koutis et al., 2011); and ii) leverages for workers who remain with the same firm have a closed-form solution.

[15] For example, the covariance of $\mathbf{X}\widehat{\boldsymbol{\beta}}$ and $\mathbf{W}\widehat{\boldsymbol{\gamma}}$ can be written as $\widehat{\boldsymbol{\beta}}^T\widetilde{\mathbf{A}}\widehat{\boldsymbol{\gamma}}$, where $\widetilde{\mathbf{A}}$ is a non-squared matrix. Since $\widehat{\boldsymbol{\gamma}}$ is typically very close (in probability) to $\boldsymbol{\gamma}$, we have that $\mathbb{E}\left(\widehat{\boldsymbol{\beta}}^T\widetilde{\mathbf{A}}\widehat{\boldsymbol{\gamma}} \,\middle|\, \mathbf{X}\right) \approx \mathbb{E}\left(\widehat{\boldsymbol{\beta}}^T\widetilde{\mathbf{A}}\boldsymbol{\gamma} \,\middle|\, \mathbf{X}\right) = \mathbb{E}\left(\widehat{\boldsymbol{\beta}}^T \,\middle|\, \mathbf{X}\right)\widetilde{\mathbf{A}}\boldsymbol{\gamma} = \boldsymbol{\beta}^T\widetilde{\mathbf{A}}\boldsymbol{\gamma}$.

4 Comparison with KSS

Both KSS and the bootstrap method rely on iterative procedures to estimate the bias, requiring the solution of multiple linear systems. The key difference between our methods lies in the type of linear systems being solved and the specific part of the trace term being approximated. KSS make approximations tailored to the matrix $\mathbf{A}$, which is specific to the moment being corrected. As a result, their method requires solving as many systems as there are moments to be corrected. In contrast, the bootstrap method estimates the leverages and two trace terms, limiting the number of systems to solve to *at most* three. This makes the bootstrap method particularly well-suited for applying multiple corrections to a given set of estimated fixed effects, such as subgroup-specific corrections. We explain with more detail below.

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 estimator $\widehat{\mathbf{V}}$, we can rewrite the direct bias correction (4) 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:

$$\mathbf{X}^T \mathbf{X} \boldsymbol{z} = \mathbf{A}_f \boldsymbol{r}. \tag{8}$$

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 (8) multiple times, which is analogous to solving the normal equations in our bootstrap method. However, the system in their approach differs from ours: it is a function of the specific quadratic form characterized by the matrix $\mathbf{A}$.

In some cases, such as performing a variance decomposition of an AKM model with worker and firm fixed effects, one can reuse the estimates from the correction of the variance of worker fixed effects and firm fixed effects to compute the correction of the covariance between worker and firm fixed effects. When making a single set of corrections, KSS requires solving three systems: one for the leverages, one for the worker fixed effects, and one for the firm fixed effects. If additional corrections are needed, for example for different subsamples, the number of systems to solve increases proportionally. Thus, KSS must solve *at least* three systems. By contrast, our method solves *at most* three systems regardless of the number of corrections: one for the leverages, and two for the matrices $\mathbf{V}_+$ and $\mathbf{V}_-$.

The key conceptual difference between the methods is that the bootstrap method approximates the entire trace term, whereas KSS approximates the diagonal terms $s_{ii}(\mathbf{A})$. This conceptual difference allows the bootstrap method to efficiently handle multiple corrections. However, it also reveals an advantage of KSS's approach: it can be easily adapted 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. In such cases, one could estimate $s_{ii}(\mathbf{A})$ once and compute an estimate of the direct bias correction for both wages and hours by adjusting the variance estimates $\widehat{\sigma}_i$.

In the end, the choice between KSS and the bootstrap method depends on the specific application. If there are more corrections than dependent variables, then it is better to use our bootstrap correction. If the opposite is true, then it is better to use KSS.

4.1 Speed and accuracy: KSS vs bootstrap

We compare our method to KSS in terms of both speed and accuracy. To do so, we simulate labor market data based on 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.

We apply the leave-one-out covariance matrix estimator for both methods. To increase comparability, we adopt the same data selection procedure as KSS (leaving out the worker) to ensure that all leverage values are less than one, i.e., $P_{ii} < 1 \ \forall i$. Although this sample-selection procedure is more restrictive than necessary, it ensures the sample meets this requirement. Online

Table 1: Monte Carlo simulations. Heteroscedastic errors

MSE×10,000		Observation		Match	
	Plug-in	Bootstrap	KSS	Bootstrap	KSS
var(Worker)	2,036.592	0.088	0.086	0.100	0.103
var(Firm)	605.942	0.312	0.312	0.317	0.324
cov(Worker, Firm)	528.039	0.107	0.106	0.116	0.122
Sample Selection	Worker	Worker	Worker	Worker	Worker
Clustering Level		Obs	Obs	Match	Match
Time (sec)		265	585	240	560

Notes: Simulated labor market with 5,266,714 observations. The table shows the Mean Squared Errors (MSE) of the different quadratic objects multiplied by 10,000. *Plug-in*: uncorrected; *Bootstrap*: bootstrap-corrected; *KSS*: corrected using KSS. *var(Worker)*: variance of worker fixed effects (θ), *var(Firm)*: variance of firm fixed effects (ψ), *cov(Worker, Firm)*: covariance between worker and firm fixed effects. At the bottom, *Sample Selection*: selection criterion for the leave-one-out connected set; *Clustering Level*: clustering level of the covariance matrix estimator of the error terms; *Time (sec)*: time in seconds. We do 300 iterations\bootstraps for the corrections.

Appendix A offers a detailed discussion of less restrictive data selection procedures that also satisfy the requirement.

In our simulations, the sample size is approximately 5 million observations per simulation, with an average of 3 movers per firm and 12 employees per firm. Both methods utilize the preconditioned conjugate gradient method in Matlab to solve the linear systems. We impose the same tolerance level for convergence and do 300 iterations for each method.[16]

Table 1 presents the results for different assumptions about the clustering level of the covariance matrix of the error terms. Columns under *Observation* assume heteroscedasticity with clustering at the observation level, while the columns under *Match* assume that the errors are clustered at the match level. We measure accuracy by reporting the Mean Squared Error (MSE) multiplied by $10,000$. As expected, both methods reduce the MSE with respect to the plug-in estimator. Comparing the *Bootstrap* and *KSS* estimators within clustering assumption, the MSEs of both methods are identical to six digits. The MSE of *Observation* columns are smaller as the clustering level of the true error terms is the observation.

The bootstrap method is overall faster than KSS as it takes less than half of the running time. This shows that even in the case where both methods need to solve three systems of equations

[16]Matlab codes and working examples are in https://github.com/mazkarate/bias_correction.

per iteration, our bootstrap method is faster than KSS.[17]

5 Applications using French employer-employee data

In this section, we show the advantages of our method through four applications. First, we do a standard AKM variance decomposition of log wages for the full sample, using different assumptions about how the errors are clustered. Second, we examine the relationship between sorting and the size of labor markets. To do this, we do corrections at the local labor market level. Specifically, we analyze two aspects of sorting: its *direction* and *intensity*. For direction, we use the correlation between worker and firm fixed effects. For intensity, we look at the correlation between worker fixed effects and the average fixed effects of their coworkers (Lopes de Melo, 2018). Third, we do a variance decomposition exercise by gender. Finally, we analyze how the components of an AKM variance decomposition change over the life cycle.

Our specification follows (1), with the addition of controls that we residualize before applying the corrections. Using the corrected estimates, we then do variance decompositions for each group g, where the groups vary depending on the application, as follows:

$$\text{var}\left(\log w_{it} \left| g\right.\right) = \text{var}\left(\theta_i \left| g\right.\right) + \text{var}\left(\psi_{\mathcal{J}(i,t)} \middle| g\right) + 2\text{cov}\left(\theta_i, \psi_{\mathcal{J}(i,t)} \middle| g\right) + \text{var}\left(\varepsilon_{it} \left| g\right.\right).$$

This equation decomposes the variance of log wages into components corresponding to worker fixed effects (θ_i), firm fixed effects ($\psi_{\mathcal{J}(i,t)}$), their covariance, and the residual variance.

Before presenting these applications, we briefly introduce the data below. Additional details on how the sample was constructed can be found in Online Appendix D.

5.1 Data

We follow Babet et al. (2022) to construct a panel using French administrative data from the *DADS Base Tous Salariés (BTS)*. Each year, the *BTS* provides a cross-sectional database with information on all jobs held by each worker in both the current and previous year. However, since

[17] We have optimized the correction for the simple case with only two fixed effects—which is also the default specification in KSS's code—leading to significant speed gains. Online Appendix E provides more details. In principle, these improvements could be incorporated into KSS's code, which we expect would also improve their running time.

worker identifiers change annually, linking the data across years is not straightforward. We apply the algorithm from Babet et al. (2022) to create a panel by matching identifiers based on observable characteristics across overlapping years.[18] Additionally, we group establishments within the same commuting zone to define firm identifiers.

Our analysis covers the years 2009 to 2019, divided into two periods: 2009-2014 and 2015-2019. The sample is restricted to private sector workers aged 20 to 60 in metropolitan France. We include only each worker's main job for the year, excluding cases with imputed or missing hours worked.

The main text focuses on the 2015-2019 sample for most exercises. The Supplemental Material provides the corresponding tables and figures for the 2009-2014 sample.

Labor markets are defined as combinations of commuting zones and 2-digit occupations. The dependent variable is the log of hourly wages, with quadratic and cubic age terms, as well as year fixed effects, included as controls.[19]

5.2 Basic AKM decomposition with different clustering

In this section, we conduct a variance decomposition of the AKM model, which is a standard exercise in the literature. The novelty here is applying corrections using different assumptions about error clustering.

Table 2 presents both the plug-in and corrected estimates using the bootstrap and KSS. We perform several bootstrap corrections with varying sample selection criteria to ensure the existence of the leave-one-out variance estimator for different levels of clustering. In column (1), we use the entire connected set with the assumption of homoscedastic errors. In column (2), we select the sample using a leave-observation-out strategy and assume heteroscedasticity for the error terms. In column (3), clustering is done at the match level and sample is selected using the leave-match-out strategy. In column (5), we apply the stricter leave-worker-out selection method used by KSS.[20] Column (4) reflects a case where the cluster is defined at the worker-occupation level, leading to a smaller sample compared to the other cases.[21]

[18] See Section 1 and Appendix C of their paper for further details.

[19] We remove the linear term of the polynomial to avoid collinearity with the year and worker fixed effects. Following Card et al. (2018), we the cubic polynomial is flat at age 40.

[20] Online Appendix A presents more details on the difference between these data selection methods.

[21] We use 4-digit occupation codes for the clustering.

Table 2: Application. Plug-in vs corrected estimates

		Bootstrap					
	Plug-in	(1)	(2)	(3)	(4)	(5)	KSS
var(y)	0.135	0.136	0.135	0.135	0.136	0.135	0.135
var(Worker)	0.095	0.091	0.088	0.085	0.085	0.086	0.088
var(Firm)	0.016	0.016	0.014	0.013	0.013	0.013	0.012
cov(Worker, Firm)	0.006	0.007	0.008	0.010	0.009	0.010	0.010
corr(Worker, Firm)	0.162	0.175	0.229	0.293	0.279	0.291	0.320
Sample Selection	Match	None	Obs	Match	Worker×Occ	Worker	Worker
Clustering Level			Obs	Match	Worker×Occ	Match	Match
Observations	53,340,591	56,713,306	53,984,000	53,340,591	51,456,580	52,350,422	52,350,422
Workers	15,028,603	16,130,300	15,155,610	15,028,603	14,522,368	14,685,248	14,685,248
Firms	842,333	1,384,988	932,785	842,333	702,639	810,403	810,403
Time (min)		70	151	133	149	133	303

Notes: Sample 2015-2019. *Plug-in*: uncorrected estimates; *Bootstrap*: bootstrap-corrected estimates; *KSS*: corrected estimates using KSS. *var(y)*: variance of residualized log hourly wages; *var(Worker)*: variance of worker fixed effects (θ); *var(Firm)*: variance of firm fixed effects (ψ); *cov(Worker, Firm)*: covariance between worker and firm fixed effects; *corr(Worker, Firm)*: correlation between worker and firm fixed effects. At the middle, *Sample Selection*: data selection procedure for the leave-one-out connected set: *None* takes the connected set, *Obs* leaves the observation out, *Match* leaves the worker-firm match out, *Worker*×*Occ* leaves the worker-occupation out, *Worker* leaves the worker out; *Clustering Level*: clustering level of the covariance matrix estimator of the error terms. At the bottom, *Observations*: person-year observations; *Workers*: number of workers; *Firms*: number of firms; and *Time (min)*: time in minutes. We do 300 iterations\bootstraps for the corrections.

To minimize the loss of observations during sample selection, we follow KSS and assume that observations that are not leave-cluster-out estimable but part of the connected firm set are clustered at the observation level. For example, if clustering is at the match level, workers who stay with the same firm are not leave-cluster-out estimable. In such cases, we treat their cluster as being at the observation level.

Focusing on the results assuming clustering at the match level, Table 2 shows that, as expected, the plug-in estimates underestimate covariance and overestimate variances. After correcting for the limited mobility bias, the correlation between worker and firm fixed effects increases significantly, from 0.162 to 0.293 (column (3)). The values are slightly lower when using clusters at the worker-occupation level. Comparing the bootstrap estimates in column (5) with the KSS estimates in column (6), we see a difference in the corrected variance estimates, leading to a slightly higher correlation estimate for KSS of 0.32 versus 0.29 for the bootstrap-corrected correlation. The results are similar for the 2009-2014 period which can be found on the Supple-

mental Material.

Comparing the computational time across exercises on the last row of Table 2, the bootstrap corrections are faster than KSS taking less than half the time. This time improvement is similar to what we found in Monte Carlo simulations.

The specification in column (1) shows the lowest correlation among the corrected estimates, followed by the estimate in column (2). The corrected correlation estimates remain quite similar when clustering is allowed at the match or worker-occupation level. In the following exercises, we do the corrections assuming clustering at the match level.

5.3 Sorting across French labor markets

In this section we study the relationship between sorting and labor market size. We follow a similar approach to Dauth et al. (2022), who use German data to test a well-known idea in urban and labor economics: larger markets lead to better matches. Dauth et al. (2022) find a positive relationship between the correlation of worker and firm fixed effects and city population. To address the limited mobility bias, they apply the KSS correction for each city, running an AKM model separately for each one. However, this approach loses information from workers moving across cities. Our method avoids this issue by estimating the model on the entire sample and then applying corrections for each labor market.[22]

Moreover, Dauth et al. (2022) suggest that a more precise definition of local labor markets is city-occupation pairs rather than entire cities, since *"workers looking for a job, and plants looking for an employee, are likely to search within specific occupations."*[23] We agree with this view, which is why we define labor markets as combinations of commuting zones and 2-digit occupations.[24]

[22] Pérez et al. (2023) do the same exercise as Dauth et al. (2022) but using Mexican data. They also find a positive gradient between the worker-firm fixed effect correlation and labor market size. Heath Milsom and Hou (2024) do a variance decomposition of an extended AKM using job (firm-occupation) fixed effects to the commuting zone population. In contrast with Dauth et al. (2022), Heath Milsom and Hou (2024) do the corrections using with the whole sample. Using similar French data like us, they find that the covariance between worker-job fixed effects is increasing in population.

[23] Page 1481 of their paper.

[24] Dauth et al. (2022) also examine the relationship between sorting and labor market size defined as city-occupation pairs (see Figure 12 of their paper). However, they do not correct for limited mobility bias in this case.

Sorting Direction. Figure 1 illustrates the relationship between the correlation of worker fixed effects (θ) and firm fixed effects (ψ) with local labor market size.[25] We use two measures of local labor market size: the number of workers and the number of firms. The plots on the left display the plug-in estimates of the correlation, suggesting that larger labor markets exhibit slightly better within-market sorting compared to smaller ones. While the correlation between worker and firm fixed effects increases with market size, the slope remains modest. In contrast, the plots on the right show estimates corrected for limited mobility bias using the leave-match-out variance estimator. After correcting for the bias, a clear positive gradient emerges, suggesting that larger labor markets offer better matching opportunities.[26]

Table 3 complements these results by presenting the OLS estimated slope coefficients of a regression between the worker-firm fixed effects correlation and the logarithm of labor market size. Again, we use the same two measures of labor market size: number of workers and number of firms. The *Plug-in* column presents the estimated slopes using the plug-in estimates, where the gradients are positive but modest across all labor market size measures. However, these gradients almost triple once we account for the bias as shown in the *Bootstrap* columns.

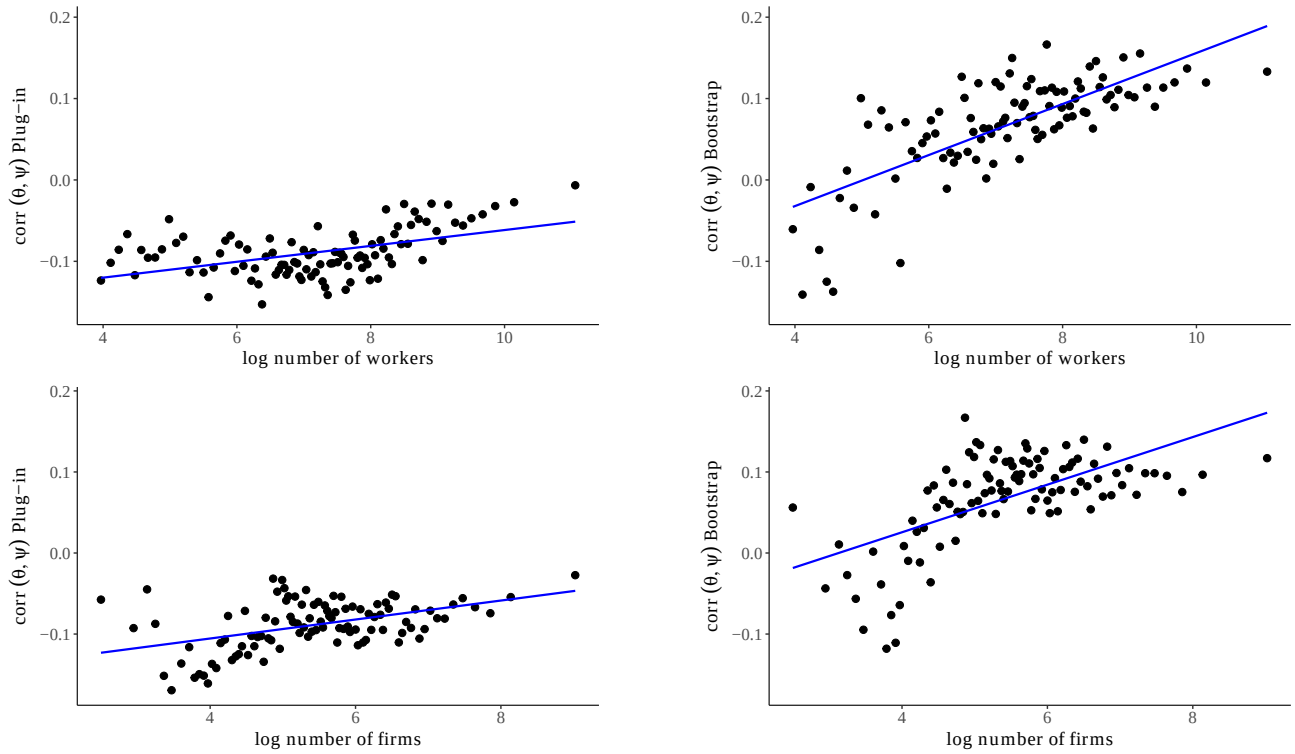
Online Appendix F shows that these patterns remain robust when defining local labor markets using combinations of commuting zones and 4-digit occupations, instead of 2-digit occupations. It also shows that, when labor markets are defined by commuting zones alone—as done in previous studies—the relationship between market size and sorting direction weakens when using corrected estimates. Further details are provided in the Online Appendix.

Time Comparison with KSS. We compare the computational time of the bootstrap and KSS corrections per groups in the application by estimating corrections per labor market. The entire bootstrap correction process was completed in 2.7 hours. In contrast, a researcher using KSS would need to run an AKM regression for each labor market. While running each regression individually would be faster due to fewer parameters in each model, this process would need

[25] We only use markets with more than 50 workers and five firms and excluded markets who gave corrected estimates for the correlation below -1 or above 1.

[26] While we show that failing to apply corrections leads to biased OLS estimates, another source of bias arises from reverse causality—where better sorting can lead to larger markets. Leknes et al. (2022) use historical mining locations in Norway as an instrument for market size, using the plug-in estimates for the correlations. Their IV estimates reveal a *downward* bias in the OLS estimates (see Table 2 in their paper), surprising as reverse causality suggests an *upward* bias. However, limited mobility bias may explain why the IV estimate exceeds the OLS estimate, similar to how our OLS estimate using corrected correlations is greater than when using plug-in correlations.

Figure 1: Sorting direction and labor market size: CZ × 2-digit occupations



Notes: Binned scatter plots between sorting direction—the correlation between worker (θ) and firm (ψ) fixed effects—and labor market (combination of commuting zone and 2-digit occupations) size. x-axis: two different measures of size by the logarithm of the (i) number of workers for the top figures, and (ii) number of firms for the bottom figures. y-axis: on the left, plug-in estimates, on the right, bias-corrected estimates.

Table 3: Gradient of sorting on labor market size: CZ × 2-digit occupations

	Sorting Direction				Sorting Intensity			
	Plug-in		Bootstrap		Plug-in		Bootstrap	
log No. Workers	0.0098		0.0315		-0.0009		-0.0125	
	(0.0015)		(0.0025)		(0.0012)		(0.0019)	
log No. Firms		0.0118		0.0292		0.0039		-0.0018
		(0.0018)		(0.0031)		(0.0015)		(0.0023)
Number of Markets	5,688				5,687			

Notes: Slope coefficients of an OLS regression of sorting direction—worker-firm correlation—, and sorting intensity—worker-coworker correlation—with different measures of labor market (combination of commuting zone and 2-digit occupations) size. Standard errors in parenthesis. *Plug-in*: slope estimate using plug-in estimates. *Bootstrap*: slope estimate using bootstrap-corrected estimates with the leave-match out covariance matrix estimator.

to be repeated thousands of times. To estimate how long this would take, we calculated how many markets could be corrected using KSS within the same time frame as the bootstrap. We found that KSS could correct 690 markets, which is 12.13% of the total. Extrapolating from this, it would take more than 22 hours to complete the same exercise using KSS.

Moreover, there would be significant information loss from excluding observations of workers who were employed outside their local labor market in other periods.

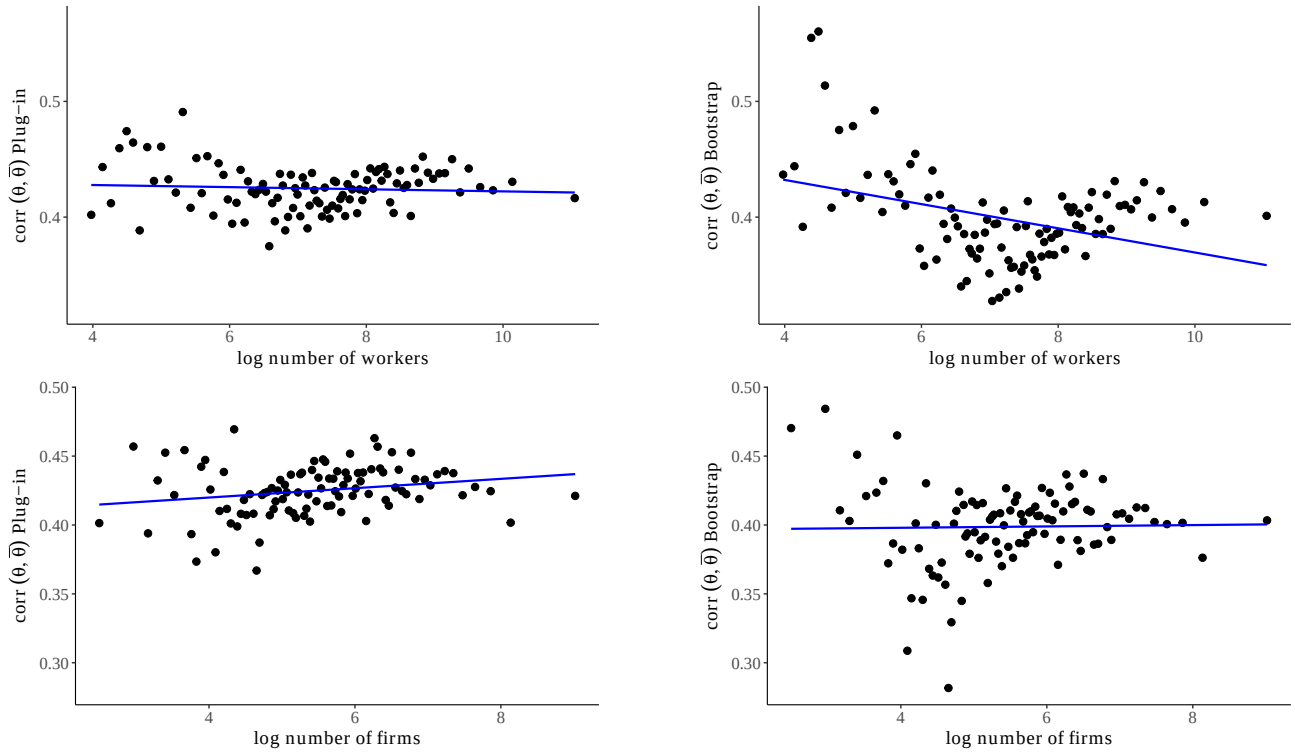
Sorting Intensity. The use of the AKM model to measure sorting has been debated. In a simple matching model, Eeckhout and Kircher (2011) show that wage data alone cannot identify the *sign of sorting*, but it can be used to measure the *intensity of sorting*. They also show that the AKM correlation between worker and firm fixed effects is not necessarily informative about sorting intensity. However, Lopes de Melo (2018) argues that the correlation between worker fixed effects (θ) and coworkers fixed effects ($\bar{\theta}$) from the AKM model is informative regarding the strength of sorting. We therefore examine the relationship between this correlation and labor market size.[27]

Figure 2 and the last four columns of Table 3 present the results. While Figure 1 and the *Sorting Direction* columns of Table 3 show that sorting is more positive in larger markets, Figure 2 and the last four columns reveal that sorting intensity is lower in these markets, or at best, not significant.

Using corrected estimates reduces the gradient between sorting intensity and market size, even reversing its sign when using the number of firms to define market size. Online Appendix F shows that these patterns on sorting intensity are robust when defining labor markets using combinations of commuting zones and 4-digit occupations. However, when labor markets are defined only by commuting zones, the result is reversed: the relationship between sorting intensity and market size is greater when using corrected estimates. This is in line with Dauth

[27] In the models of Eeckhout and Kircher (2011) and Lopes de Melo (2018), wages are not necessarily monotonic with respect to firm productivity. This implies that the estimated fixed effects may not reflect the true productivity types of firms, which could be problematic if we aim to identify production complementarities between firms and workers. However, if we are only interested in understanding sorting patterns based on wages, the AKM model remains informative. As noted by Bartolucci et al. (2018): "The correlation from Abowd, Kramarz, and Margolis's (1999) methodology is informative on the extent to which high-wage workers sort into high-paying firms. Whenever worker and firm fixed effects are increasing in their unobservable productive characteristics, this correlation is also informative about sorting by latent productivity."

Figure 2: Sorting intensity and labor market size: CZ × 2-digit occupations



Notes: Binned scatter plots between sorting intensity—the correlation between worker fixed effects (θ) and the average of coworkers ($\bar{\theta}$)—and labor market (combination of commuting zone and 2-digit occupations) size. x-axis: two different measures of size by the logarithm of the (i) number of workers for the top figures, and (ii) number of firms for the bottom figures. y-axis: on the left, plug-in estimates, on the right, bias-corrected estimates.

et al. (2022) that report similar findings with uncorrected correlations.[28]

From these two exercises on sorting direction and intensity, we can conclude that, using a more granular definition of labor markets, larger markets may exhibit better matches—where high-wage workers are employed by high-wage firms—but the link between market size and sorting intensity is weak or even negative, indicating reduced worker segregation in larger markets.

[28] See Table B.1 of their Online Appendix.

Table 4: AKM decomposition: gender differences over time

	2009-2014				2015-2019			
	Plug-in		Bootstrap		Plug-in		Bootstrap	
	Women	Men	Women	Men	Women	Men	Women	Men
var(y)	0.123	0.148	0.123	0.148	0.119	0.144	0.119	0.144
var(Worker)	0.081	0.106	0.074	0.097	0.081	0.104	0.072	0.093
var(Firm)	0.017	0.018	0.014	0.014	0.016	0.017	0.012	0.013
cov(Worker, Firm)	0.007	0.007	0.010	0.010	0.006	0.007	0.009	0.010
corr(Worker, Firm)	0.180	0.152	0.321	0.258	0.153	0.157	0.300	0.279
Sample Selection	Match	Match	Match	Match	Match	Match	Match	Match
Clustering Level			Match	Match			Match	Match
Observations	29,335,213	39,469,810	29,335,213	39,469,810	24,255,509	29,085,082	24,255,509	29,085,082
Workers	6,870,591	8,908,312	6,870,591	8,908,312	6,753,534	8,284,846	6,753,534	8,284,846
Firms	816,979	831,326	816,979	831,326	723,826	733,848	723,826	733,848

Notes: Sample 2009-2014 and 2015-2019. *Plug-in*: uncorrected estimates; *Bootstrap*: corrected estimates. *var(y)*: variance of residualized log hourly wages; *var(Worker)*: variance of worker fixed effects (θ); *var(Firm)*: variance of firm fixed effects (ψ); *cov(Worker, Firm)*: covariance between worker and firm fixed effects; *corr(Worker, Firm)*: correlation between worker and firm fixed effects. At the middle, *Sample Selection*: data selection procedure for the leave-one-out connected set leaves the match out; *Clustering Level*: clustering level of the covariance matrix estimator of the error terms. At the bottom, *Observations*: person-year observations; *Workers*: number of workers; *Firms*: number of firms; and *Time (min)*: time in minutes. We do 300 bootstraps for the correction.

5.4 Gender differences in AKM decomposition

We conduct a basic AKM decomposition conditional on gender to explore differences across men and women within and across time periods. This analysis also serves to evaluate the validity of the common assumption that limited mobility bias remains constant across groups and/or over time.

Table 4 presents the results for 2009-2014 and 2015-2019. Focusing on the results for 2009-2014, the plug-in estimates of the correlation between worker and firm fixed effects suggest that both high-wage men and women sort similarly into high-wage firms, with women showing slightly higher sorting. However, the corrected estimates in Table 4 reveal a larger difference than indicated by the plug-in estimates, with women having a higher correlation than men. This pattern continues in 2015-2019, where the gap between plug-in and corrected estimates is even more pronounced. While the plug-in estimates suggest that men have a slightly higher correlation than women, the corrected estimates show that women actually have the higher correlation.

Looking now at the evolution of the correlation over time, the bias-corrected estimates of

Table 4 show that the correlation has decreased from 2009-2014 to 2015-2019 for women passing from 0.321 to 0.3 while it increased for men from 0.258 to 0.279. The reasons for this shift are unclear and exploring them in detail is beyond the scope of this paper.

It might be tempting to compare the relative magnitudes of the plug-in correlation estimates and assume that the bias affects both estimates similarly within and across periods. However, our analysis shows that this assumption is not always valid, as the correction can change the relative estimates for men and women within periods and across time.

5.5 Wage decomposition over the life cycle

In this final application, we study life cycle patterns on worker-firm sorting and on the importance of the different elements of a variance decomposition. We do this by correcting estimates for each age group. This exercise highlights the advantage of estimating the model using the entire sample. If we were to split the sample by age and perform the corrections separately, it would be impossible to identify the worker fixed effects, since there is only one observation per worker per year. By using the full sample, we overcome this limitation and can still make age-specific corrections.

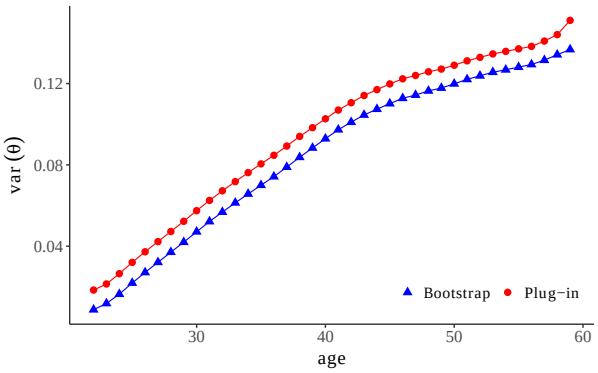
Figure 3 shows the results. Panels (a)-(c) present the plug-in and bootstrap-corrected estimates of the variance decomposition. Panels (a) and (b) show that the variance of worker and firm fixed effects increase with age, though the plug-in estimates are overestimated, with the gaps staying constant over the life cycle. Panel (c) shows the worker-firm covariance also rises with age, but the gap between plug-in and corrected estimates widens early and again later in life. Panel (d) shows that most of the sorting improvement occurs before age 30, after which the worker-firm correlation remains fairly flat, with slight differences in trends between the plug-in and corrected estimates. This contrasts with Borovičková and Shimer (2017), who find that the correlation between worker and job type increases consistently with age.[29]

Panels (e) and (f) present variance decompositions by age for the bootstrap-corrected estimates.[30] Panel (e) shows the decomposition in levels where it is clear that the variance of

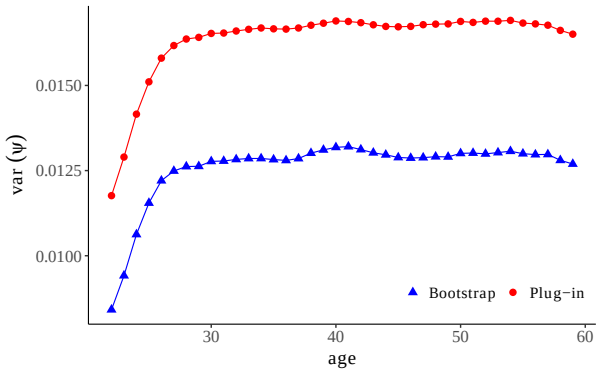
[29] See Figure 3 in their paper.

[30] We restrict the sample to individuals aged 22 to 59. This is because corrected correlations were negative for ages below 22, and there is a large increase in the variance of log wages at age 60, which is also observed in the 2009-2014 sample.

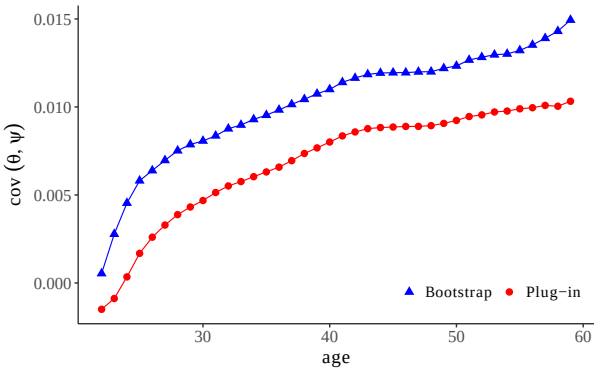
Figure 3: Life cycle patterns



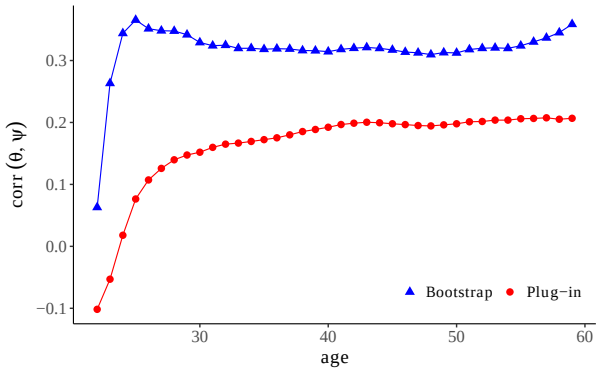
(a) Variance of worker fixed effects



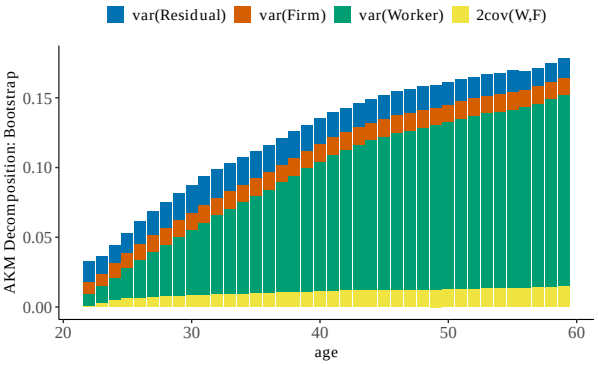
(b) Variance of firm fixed effects



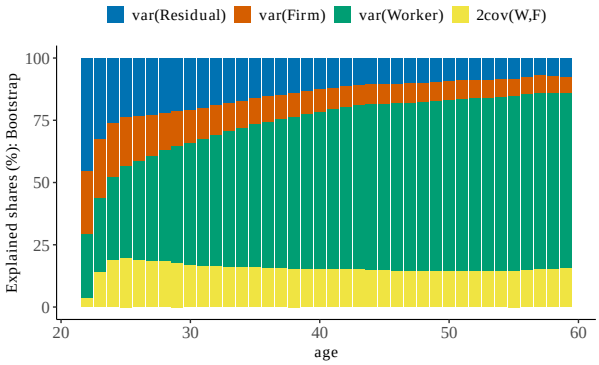
(c) Covariance of worker-firm fixed effects



(d) Correlation of worker-firm fixed effects



(e) Decomposition in levels



(f) Decomposition in explained shares

Notes: Sample 2015-2019. x-axis: age. Panels (a)-(d) show plug-in and bootstrap-corrected estimates. Panels show: (a) variance of worker effects; (b) variance of firm effects; (c) covariance of worker-firm effects; (d) correlation of worker-firm effects; bootstrap-corrected variance decompositions (e) in levels, (f) as explained shares.

residualized log hourly wages is increasing over the life cycle. This is consistent with findings from Lemieux (2006) and Heathcote, Storesletten, and Violante (2005) who, among others, document an increasing wage inequality over the life cycle related to experience or age. Panel (f) presents the decomposition as explained shares, showing that worker fixed effects are the primary driver of wage inequality over the life cycle. Initially, they explain only 21% of wage inequality, but by the end of the working age, they account for 71%. The share explained by the worker-firm covariance closely follows the corrected correlation, rising before age 25 and remaining steady with a slight U-shape over the life cycle. In contrast, the variance of firm effects and the residual variance decline in importance along the life cycle.

In summary, differences in worker fixed effects account for most of the variation in (residual) wages across the life cycle. To our knowledge, no other paper has demonstrated this result. This finding also suggests that using very long panels in AKM regressions—while keeping one fixed effect per worker for the entire sample—may be a wrong assumption.

6 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, such as in typical AKM applications. We show using Monte Carlo simulations that the method corrects the bias and is faster than KSS in simple AKM decompositions.

The main advantage of our approach is that it allows to increase the number of moments to correct without increasing significantly the computational costs, and allows for different assumptions on the error term.

The application using French labor market data shows that the bootstrap correction is useful to evaluate the components of variance decompositions per subgroups, where previous methods would impose a high time burden to do it.

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," .

BARTOLUCCI, C., F. DEVICIENTI, AND I. MONZÓN (2018): "Identifying sorting in practice," *American Economic Journal: Applied Economics*, 10, 408–438.

BONHOMME, S. (2020): "Econometric analysis of bipartite networks," in *The econometric analysis of network data*, Elsevier, 83–121.

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.

BONHOMME, S., T. LAMADON, AND E. MANRESA (2019): "A distributional framework for matched employer employee data," *Econometrica*, 87, 699–739.

BOROVIČKOVÁ, K. AND R. SHIMER (2017): "High wage workers work for high wage firms," Tech. rep., National Bureau of Economic Research.

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.

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

DHAENE, G. AND K. JOCHMANS (2015): "Split-panel jackknife estimation of fixed-effect models," *The Review of Economic Studies*, 82, 991–1030.

EECKHOUT, J. AND P. KIRCHER (2011): "Identifying sorting—in theory," *The Review of Economic Studies*, 78, 872–906.

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.

HEATH MILSOM, L. AND S. HOU (2024): "The Role of Firms and Occupations in Wage Inequality," .

HEATHCOTE, J., K. STORESLETTEN, AND G. L. VIOLANTE (2005): "Two views of inequality over the life cycle," *Journal of the European Economic Association*, 3, 765–775.

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

JOCHMANS, K. (2022): "Heteroscedasticity-robust inference in linear regression models with many covariates," *Journal of the American Statistical Association*, 117, 887–896.

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.

LEMIEUX, T. (2006): "Increasing residual wage inequality: Composition effects, noisy data, or rising demand for skill?" *American economic review*, 96, 461–498.

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.

PÉREZ, J. P., J. MELÉNDEZ, AND J. G. NUNO-LEDESMA (2023): "Matching and City Size Wage Gaps under the Shadow of Informality: Evidence from Mexico," .

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

Proofs

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

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

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 estimator is $\widehat{\boldsymbol{\beta}^*} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{v}^*$. As $\mathbb{E}\left(\boldsymbol{v}^* \,|\, \mathbf{X}, \boldsymbol{\varepsilon}\right) = 0$, then we have that $\mathbb{E}\left(\widehat{\boldsymbol{\beta}^*} \,\middle|\, \mathbf{X}, \boldsymbol{\varepsilon}\right) = 0$. Then, using the formula for the expectation of quadratic forms we get:

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

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

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

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

Proof of Proposition 4: *Unbiased.* Conditional on $\mathbf{X}$ and $\boldsymbol{\varepsilon}$, the expectations of δ^* is:

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

Consistent. Fix n and k. Let $J \longrightarrow \infty$. Using the two components of the difference of averages from the definition of δ^*, we have that:

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

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

Proof of Corollary 2: Using the law of iterated expectations we have:

$$\mathbb{E}\left(\delta^* \,|\, \mathbf{X}\right) = \mathbb{E}\left(\mathbb{E}\left(\delta^* \,|\, \mathbf{X}, \boldsymbol{\varepsilon}\right) \,|\, \mathbf{X}\right) = \mathbb{E}\left(\widehat{\delta}_D \,\middle|\, \mathbf{X}\right) = \delta,$$

where we used Proposition 4 in the second equality and the assumption $\mathbb{E}\left(\widehat{\mathbf{V}} \,\middle|\, \mathbf{X}\right) = \mathbb{V}\left(\boldsymbol{\varepsilon} \,|\, \mathbf{X}\right)$ in the last equality. □

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,

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

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

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

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:

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

The expression above can be rewritten as:

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

□

Before proving Proposition 6, let us introduce the following auxiliary Lemma.

Lemma 1. *Let $\boldsymbol{v}$ and $\boldsymbol{w}$ be two vectors. Let $(\boldsymbol{v}\cdot\boldsymbol{w})$ be the inner product of $\boldsymbol{v}$ and $\boldsymbol{w}$. Denote the matrix $\mathbf{A} \equiv \boldsymbol{v}\boldsymbol{w}^T + \boldsymbol{w}\boldsymbol{v}^T$. Then, the non-zero eigenvalues of A are equal to*

$$\lambda = (\boldsymbol{v}\cdot\boldsymbol{w}) \pm \sqrt{(\boldsymbol{v}\cdot\boldsymbol{v})(\boldsymbol{w}\cdot\boldsymbol{w})},$$

with corresponding eigenvectors

$$\boldsymbol{u} = \pm a\boldsymbol{v} + \boldsymbol{w}, \quad \textit{and } a = \sqrt{\frac{(\boldsymbol{w}\cdot\boldsymbol{w})}{(\boldsymbol{v}\cdot\boldsymbol{v})}}.$$

Proof. Let $\boldsymbol{u}$ be an eigenvector of $\mathbf{A}$ associated to a non-zero eigenvalue λ. Then $\boldsymbol{u}$ is in the subspace of $\boldsymbol{v}$ and $\boldsymbol{w}$. This means that $\boldsymbol{u} = a\boldsymbol{v} + b\boldsymbol{w}$ for some scalars a and b. Then we have that:

$$\mathbf{A}\boldsymbol{u} = \mathbf{A}\left(a\boldsymbol{v} + b\boldsymbol{w}\right) = a\mathbf{A}\boldsymbol{v} + b\mathbf{A}\boldsymbol{w}.$$

Developing $\mathbf{A}\boldsymbol{v}$ and $\mathbf{A}\boldsymbol{w}$ and using the fact that $\mathbf{A}\boldsymbol{u} = \lambda\boldsymbol{u}$ we get

$$\begin{aligned}\mathbf{A}\boldsymbol{u} &= a\left[(\boldsymbol{v}\cdot\boldsymbol{w})\boldsymbol{v} + (\boldsymbol{v}\cdot\boldsymbol{v})\boldsymbol{w}\right] + b\left[(\boldsymbol{w}\cdot\boldsymbol{w})\boldsymbol{v} + (\boldsymbol{v}\cdot\boldsymbol{w})\boldsymbol{w}\right] \\ &= \left[a(\boldsymbol{v}\cdot\boldsymbol{w}) + b(\boldsymbol{w}\cdot\boldsymbol{w})\right]\boldsymbol{v} + \left[a(\boldsymbol{v}\cdot\boldsymbol{v}) + b(\boldsymbol{v}\cdot\boldsymbol{w})\right]\boldsymbol{w} \\ &= \lambda\left[a\boldsymbol{v} + b\boldsymbol{w}\right].\end{aligned}$$

The last equality implies a system of two equations with two unknowns. This system can be represented in matrix form as:

$$\underbrace{\begin{pmatrix}(\boldsymbol{v}\cdot\boldsymbol{w}) & (\boldsymbol{w}\cdot\boldsymbol{w}) \\ (\boldsymbol{v}\cdot\boldsymbol{v}) & (\boldsymbol{v}\cdot\boldsymbol{w})\end{pmatrix}}_{\mathbf{B}}\begin{pmatrix}a\\b\end{pmatrix} = \lambda\begin{pmatrix}a\\b\end{pmatrix}.$$

This is an eigensystem and shows that the non-zero eigenvalues of $\mathbf{A}$ are the eigenvalues of the matrix $\mathbf{B}$. Also, the eigenvectors of $\mathbf{B}$ determine the values of the scalars a and b. Finding the roots of the characteristic polynomial of $\mathbf{B}$ gives us $\lambda = (\boldsymbol{v}\cdot\boldsymbol{w}) \pm \sqrt{(\boldsymbol{v}\cdot\boldsymbol{v})(\boldsymbol{w}\cdot\boldsymbol{w})}$. Finally, we can normalize $b = 1$, and to find a we solve the following equation for each eigenvalue:

$$(\boldsymbol{v}\cdot\boldsymbol{w})a + (\boldsymbol{w}\cdot\boldsymbol{w}) = \lambda a.$$

The solution is $a = \frac{(\boldsymbol{w}\cdot\boldsymbol{w})}{\lambda - (\boldsymbol{v}\cdot\boldsymbol{w})}$. Substituting for each value of the two eigenvalues and simplifying gives the final result. □

Proof of Proposition 6: Let $\boldsymbol{v} \equiv \widehat{\boldsymbol{\varepsilon}}_g^{LC}$ and $\boldsymbol{w} \equiv \frac{1}{2}\boldsymbol{y}_g$. Then by Lemma 1 we get the result. □

Proof of Proposition 7: 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:

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

Proof of Proposition 8: 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 an auxiliary Lemma that will prove helpful for proving Proposition 10.

Lemma 2. *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 for all vectors $\mathbf{B}\boldsymbol{v}$:

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

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

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

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

$\square$

Lemma 3 (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 (PSD) as they are averages of matrices formed by outer products of vectors. By assumption, $\widehat{\mathbf{M}}_{gg}$ is non-singular. Together with PSD, then $\widehat{\mathbf{M}}_{gg}$ has strictly positive eigenvalues and is positive definite. Then, $\widehat{\mathbf{P}}_{gg} + \widehat{\mathbf{M}}_{gg}$ is also positive definite (PD), so its inverse exist and is also PD. Now, using Lemma 2 we can show that $\overline{\mathbf{P}}_{gg}$ has non-negative eigenvalues and $\overline{\mathbf{M}}_{gg}$ 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 10: 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 3. 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 $\widehat{\mathbf{P}}_{gg} + \widehat{\mathbf{M}}_{gg} = \mathbf{L}_{gg}\mathbf{L}_{gg}^T$ and $\mathbf{L}_{gg}$ is non-singular. Then, $\overline{\mathbf{P}}_{gg} = \left(\mathbf{L}_{gg}\mathbf{L}_{gg}^T\right)^{-1}\widehat{\mathbf{P}} = \left(\mathbf{L}_{gg}^T\right)^{-1}\mathbf{L}_{gg}^{-1}\widehat{\mathbf{P}}$. Pre-multiply $\overline{\mathbf{P}}_{gg}$ by $\mathbf{L}_{gg}^T$ and post-multiply it by $\left(\mathbf{L}_{gg}^T\right)^{-1}$ and we get

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

Then, $\overline{\mathbf{P}}_{gg}$ and $\overline{\mathbf{P}}^S_{gg}$ are similar matrices, which means they have the same eigenvalues. By Lemma 3 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 9: 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 $J = KJ^*$, $K \in \mathbb{N}$. 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. □

ONLINE APPENDIX

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

Miren Azkarate-Askasua and Miguel Zerecero

A Sample selection for leave-cluster-out variance estimators

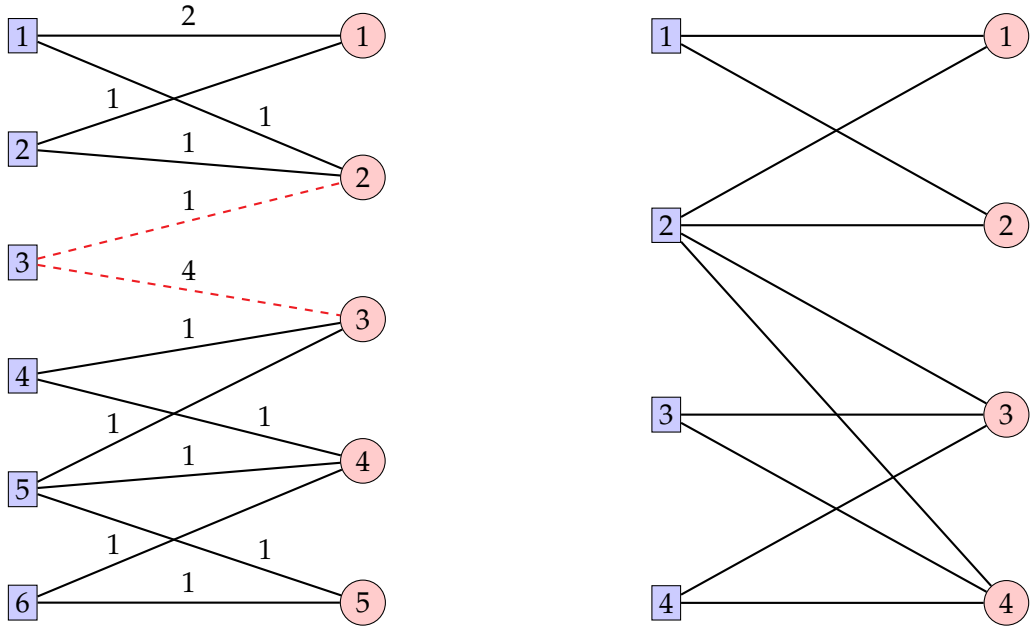
We explain how to apply standard graph theory tools to select the sample for leave-cluster-out variance estimators when clusters are either firm- or worker-specific. This approach includes both the leave-one-out and leave-match-out cases, corresponding to clusters at the observation and match level. Using the leave-one-out variance estimator requires that $P_{ii} \neq 1$, or equivalently $M_{ii} \neq 0$. For the leave-cluster-out variance estimator, M_{gg} must not be singular.

To clarify the following points, we first represent the worker-firm data as a bipartite graph (Bonhomme 2020). Figure A1 shows an example with six workers and five firms. The vertices on the left represent workers, and the vertices on the right represent firms. The edges between these vertices represent worker-firm matches, and each edge can have a weight corresponding to the number of observations for that match. For example, the edge between worker 3 and firm 3 has a weight of 4, meaning we have 4 observations of worker 3 employed by firm 3.

Connected set. First, we restrict our analysis to a connected component of the sample. This ensures that the rank condition is satisfied and that all fixed effects can be compared (Abowd, Creecy, and Kramarz 2002; Card, Heining, and Kline 2013; Jochmans and Weidner 2019). In practice, we keep the largest connected set. The graph in Figure A1 is already connected.

Leave-one-out sample. The leave-one-out requirement is stronger. It requires that all parameters can still be identified after removing any single observation. If an observation has a leverage of 1, then identifying a parameter depends entirely on that observation. In terms of the worker-firm network, this means that removing the edge corresponding to that observation would disconnect the network, and the edge's weight is 1.

Figure A1: 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.

An edge that disconnects the network when removed is known as a *bridge* or a *cut-edge*. In Figure A1a, 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 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 check if the remaining network has no bridges with unit weights. In this case, it does not, so this subsample would be suitable for the leave-one-out variance estimator.

We have developed an algorithm that leverages this idea to efficiently select the sample to use the leave-one-out variance estimator.

Leave-match-out. Similarly, to use the leave-cluster-out variance estimator 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 A1a, 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.

Just like the leave-one-out case, we also have a highly efficient algorithm specifically for the leave-match-out case.

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*. As we explain below, the leave-worker-out approach is more restrictive than the leave-match-out and leave-one-out approaches explained before.

Worker 3 in Figure A1a 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. In this example, leave-worker-out leads to the same sample selection as leave-match-out. However, this is not always the case, and in general, using leave-worker-out is a stronger requirement than the leave-match-out approach. Consider Figure A1b. The graph has no bridges, therefore it is leave-match-out estimable. However, worker 2 is an articulation point. If we were to follow KSS's leave-worker-out procedure we

would remove *all* observations corresponding to worker 2. We would then 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 which was already leave-match-out estimable.

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. However, in our applications we use the leave-one-out or leave-match-out procedures.

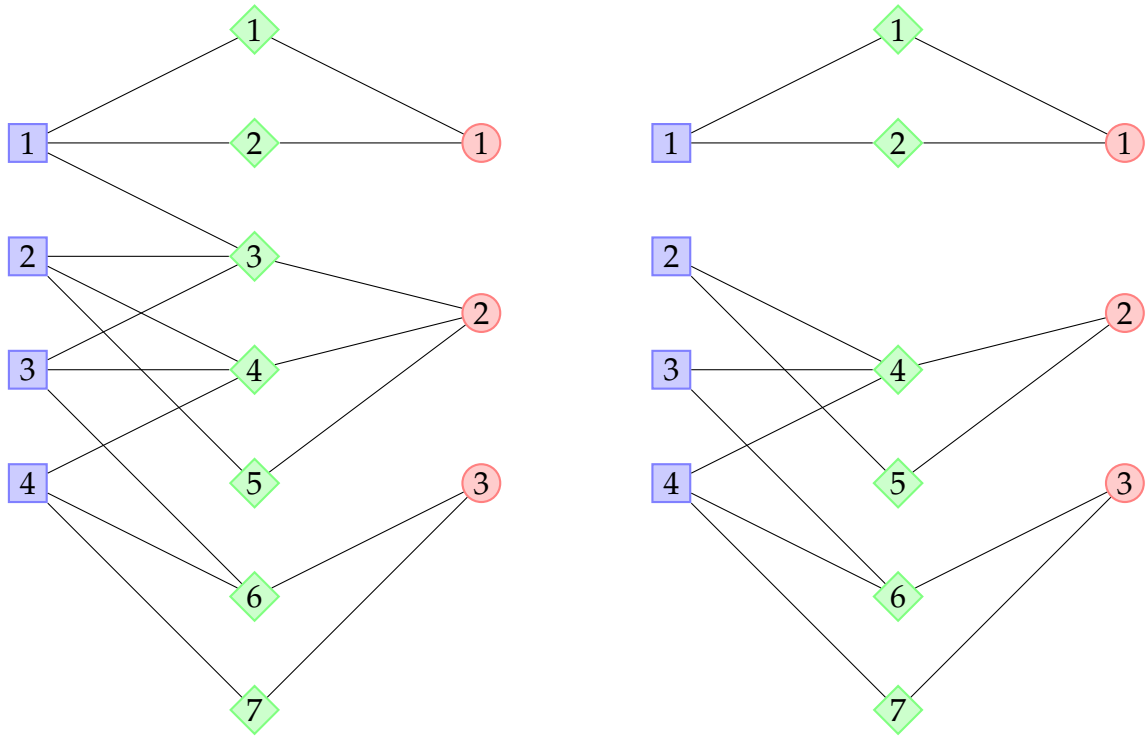
Leave-cluster-out: worker-specific or firm-specific cluster. Whenever the clusters are firm-specific or worker-specific we can also use graph theory tools to device an algorithm that selects a leave-cluster-out estimable sample. Of course, we cannot have clusters that contain an entire firm or an entire worker's spell: this clusters contain all the information to identify a firm or worker fixed effect. But we could have, for example, clustering at the firm-occupation or worker-occupation level.

Clusters contain collections of observations, which may include parts of or entire firm-worker matches. As a result, the bipartite network representation of the data is not that useful. However, we can envision the labor market as a *tripartite* graph. This graph would contain three sets of vertices: workers, firms, and clusters. The clusters vertices work as intermediaries between the firms and the workers.

To see how the tripartite graph representation of the data is helpful consider Figure A2a. Here, the vertices to the left represent workers, vertices in the middle are clusters, and vertices to the right, firms. This would be an example where the clusters are firm specific: each cluster belongs to only one firm. The advantage of representing the data like this is that the cluster vertices group all the edges connecting workers and firms. We can then find the clusters that are articulation points. This means, clusters that upon their removal will leave the graph disconnected. In the example on Figure A2a, cluster 3 is an articulation point. If we would remove it we would have two disconnected subgraphs as shown in Figure A2b. The algorithm would detect cluster 3 quickly and remove it. Then it would pick the largest connected set. In this case the set formed by workers 2, 3, and 4, firms 2 and 3, and clusters 4 to 7.

Leave-cluster-out: clusters with multiple workers and firms. When clusters contain multiple firms and workers, the idea of the tripartite graph becomes impractical. It's possible to

Figure A2: Tripartite graphs of workers, clusters, and firms



Notes: Workers are represented by square vertices to the left of each graph, clusters by diamond vertices at the middle, and firms are represented by circle vertices in the right of each graph.

find networks where workers and firms seem connected only through shared clusters, even if they're actually disconnected. In these situations, we can use a "brute force" approach. First, compute the largest connected set of workers and firms. Second, remove all workers and firms belonging to a single cluster. Then, remove each cluster at a time and check if the resulting graph has the same number of vertices (workers and firms) and is still connected. If the graph becomes disconnected or has fewer vertices, discard that cluster. Repeat this process until no more problematic clusters remain.

The brute force method is much slower than graph theory-based algorithms and can be memory-intensive, especially for clusters with many observations.

The cluster algorithms align with the specialized algorithms for leave-one-out and leave-match-out cases. If we define the cluster at the observation or match level, the leave-cluster-

out algorithm will select the same sample as the leave-one-out or leave-match-out algorithms, though it may run slightly slower. The brute force algorithm, however, is much slower.

B Additional details on leverage estimation

The estimators $\widehat{P}_{ii}$ 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 3, the leave-one-out residual for observation i is equal to $\widehat{\varepsilon}_i/M_{ii}$. As we use the estimator $\overline{M}_{ii}$ rather than the true value M_{ii}, we introduce some non-linearity bias. We correct it up to a second order.

Let $1/\overline{M}_{ii} \equiv f\left(\widehat{P}_{ii}, \widehat{M}_{ii}\right)$, which shows it is a function of both $\widehat{P}_{ii}$ and $\widehat{M}_{ii}$. The expected value of the second-order approximation of $f\left(\widehat{P}_{ii}, \widehat{M}_{ii}\right)$ around P_{ii} and M_{ii} is:

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

The feasible bias corrected estimator 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 estimators.[31]

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

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

[31] 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 $\frac{1}{J_M}\left(\left[\frac{1}{J_M-1}\sum_{j=1}^{J_M}(r_i(j) - \widehat{r}_i(j))^2\widehat{r}_i(j)^2\right] - \widehat{M}_{ii}\widehat{P}_{ii}\right)$.

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.

Proposition 11. *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 estimator 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 12. *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 $P_{ii} \leq P_{ii}^2 \implies P_{ii} \leq 1$. 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 11.

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 do 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 estimator

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 solves $\widehat{\mathbf{M}}_{gg}\boldsymbol{z} = \mathbf{L}_{gg}\widehat{\boldsymbol{\varepsilon}}_g$. This is much more efficient than inverting $\widehat{\mathbf{M}}_{gg}$ directly. Finally, we get $\widehat{\boldsymbol{\varepsilon}}^{LC}_g = \mathbf{L}^T_{gg}\boldsymbol{z}$.

Algorithm 3 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**

C.1 Simulation of bootstrap errors with leave-cluster-out variance estimate

Here we show how to simulate $\mathbf{B}_{gg+}\boldsymbol{r}$ and $\mathbf{B}_{gg-}\boldsymbol{r}$ using only two vectors instead of two matrices, which improves memory efficiency.

Focusing on $\mathbf{B}_{gg+}$ (the same applies for $\mathbf{B}_{gg-}$), Proposition 6 shows that $\mathbf{B}_{gg+}$ is a matrix where only the first column contains non-zero values, i.e., $\mathbf{B}_{gg+} = \left[\boldsymbol{b}_{gg+}, \mathbf{0}, \ldots, \mathbf{0}\right]$. Let $\boldsymbol{r} = \left[r_1, \ldots, r_{n_g}\right]'$ be a vector of independent Rademacher entries. Then, $\mathbf{B}_{gg+}\boldsymbol{r} = r_1\boldsymbol{b}_{gg+}$, where r_1 is a random scalar. This means we only need to simulate one Rademacher random variable per cluster, r_1, and work with the vector $\boldsymbol{b}_{gg+}$.

When simulating for all the clusters, we can stack all the vectors $\{\boldsymbol{b}_{gg+}\}$ and multiply each one by its corresponding simulated Rademacher value. This approach resembles the Wild Block bootstrap.

D Sample construction

The data source *BTS* is a repeated cross section with the universe of jobs per year. Worker identifiers change yearly but the data records all the jobs of a worker in a given year and in the previous one. That is, we do not have a panel of workers by matching identifiers, as the worker identifiers change every year, but one can construct the panel by matching other observable characteristics. 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) provide a code to match workers across datasets using this information. The code generates an individual identifier that tracks workers across years, which we use to generate the panel.

After creating the panel, we make additional sample restrictions. We focus on main jobs of workers at the private sector working in metropolitan France with positive hourly wages, with occupation, location, age and gender information. We focus on prime aged workers who are between 20 and 60 years old. To avoid noisy hourly wages, we only keep observations with at least 90 days and 100 hours worked. We also exclude spells that have unfeasible number of hours beyond 24 hours per day worked. We exclude occupations that are related to farming or public employment.[32] Finally, we eliminate spells that have hourly wages below 80% or above

[32]In particular, after using a data provided filter for public workers, we also exclude the following 4-digit occupations according to their PCS-ESE classification: 331A, 333A, 333B, 333C, 333D, 334A, 335A, 451A, 451B,

1,000 times the hourly minimum wage per year and worker identifiers that do not have single yearly observations at some point of our samples 2009-2014 or 2015-2019. After the minimum wage filters, we transform nominal hourly wages to real ones by using the CPI.

The source variables we use are:

- *AGE*: age of the worker in a given year. As Babet et al. (2022) note in their Appendix C.1, there are age discrepancies across years. To overcome that, we impute the worker age such that it is consistent with the age at first appearance on the sample. We restrict to prime age workers aged between 20 and 60.
- *COMT*: is the municipality identifier. We match the municipality codes to the commuting zone classification in 2020 *ZEMP2020*.
- *DOMEMP*: is a variable that can be used to restrict to workers whose workplace is the private sector. Private workers are those with *DOMEMP* equal to 6, 7, 8 and 9.
- *DOMEMP_EMPL*: is a variable that can be used to restrict to workers whose employer is a private firm. Private workers are those with *DOMEMP_EMPL* equal to 6, 7, 8 and 9.
- *IR_NBHEUR*: we drop observations with imputed hours by keeping only IR_NBHEUR equal to D.
- *NBHEUR*: total yearly hours in the job. Hourly wages in the job are defined as *S_BRUT*/*NBHEUR*. We keep observations with positive hourly wages.
- *NIC*: the concatenation of *SIREN* and *NIC* gives the establishment identifier. Our definition of establishment is the aggregation of all the establishments of a firm within commuting zone (*ZEMP2020*).
- *PCS*: is a 4-digit occupation classification that is well maintained starting in 2009. In particular we use the *Nomenclatures des professions et catégories socioprofessionnelles des emplois salariés des employeurs privés et publics (PCS-ESE).* We use 2-digit occupations by taking the first 2 digits of *PCS*. We remove workers with farming occupations (2-digit occupation

451C, 452A, 452B, 521A, 521B, 522A, 523A, 523B, 523C, 523D, 524A, 524B, 524C, 524D, 525C, 531A, 531B, 531C, 532A, 532B, 532C, 533A.

equal to 10), public occupations and unassigned codes (missing information or 2 digits equal to 99).

- *PPS*: indicator of main job or *poste principal*. We keep observations with *PPS* equal to 1.
- *REGT*: denotes the region of the establishment. We restrict to metropolitan France by keeping observations with *REGT* higher than 6 and dropping unclassified regions (*REGT* equal to 99).
- *S_BRUT*: gross yearly earnings in the job. We keep observations with positive earning information.
- *SEXE*: gender of the worker. Man if *SEXE* equal to 1 and woman if equal to 2.
- *SIREN*: is the firm identifier.
- *TYP_EMPLOI*: is a variable that can be used to restrict to workers with ordinary jobs by keeping *TYP_EMPLOI* equal to "O".

E Efficient AKM regression

Here we describe an efficient method for estimating the basic two-way fixed effect model:

$$y_{it} = \theta_i + \psi_{\mathcal{J}(i,t)} + \varepsilon_{it}.$$

This approach improves the performance of bias correction by making each bootstrap regression more efficient. This method is applicable only when the model includes two sets of fixed effects. Additional covariates should be residualized beforehand.

Assume all firms in our sample are connected through worker mobility. Within each firm, there are two types of workers: stayers (those who remain in the same firm throughout the sample) and movers (those who switch firms). The key idea is that the firm fixed effects estimated using only movers are *numerically identical* to those estimated using the entire sample. Therefore, we can run the regression with only movers, simplifying the process by reducing

the number of covariates, as we exclude the fixed effects of the stayers. In our application with French data, these fixed effects represent the largest share of covariates.

To clarify, consider the first order condition for the worker fixed effect θ_i for a worker i who has worked for the same firm throughout the sample:

$$\sum_t \left(y_{it} - \widehat{\theta}_i - \widehat{\psi}_{\mathcal{J}(i,t)} \right) = 0. \tag{E1}$$

The OLS estimator of θ_i is simply the average of y_{it} over the entire sample, minus the firm fixed effect estimator $\widehat{\psi}_{\mathcal{J}(i,t)}$.

Now, consider the first order condition for the firm $\boldsymbol{J} = \mathcal{J}(i,t)$ where the stayer worker i was employed:

$$\sum_{j \in \boldsymbol{J}} \sum_t \left(y_{jt} - \widehat{\theta}_j - \widehat{\psi}_{\boldsymbol{J}} \right) = 0.$$

This sum can be split into two parts: one for movers and one for stayers:

$$\sum_{j \in \{\boldsymbol{J} \cap \text{Movers}\}} \sum_t \left(y_{jt} - \widehat{\theta}_j - \widehat{\psi}_{\boldsymbol{J}} \right) + \sum_{j \in \{\boldsymbol{J} \cap \text{Stayers}\}} \sum_t \left(y_{jt} - \widehat{\theta}_j - \widehat{\psi}_{\boldsymbol{J}} \right) = 0.$$

The second term, which groups the stayers, equals zero because for all stayers $\sum_t \left(y_{jt} - \widehat{\theta}_j - \widehat{\psi}_{\boldsymbol{J}} \right) = 0$, as shown by the stayer's first order condition (E1) above. This means that stayers do not provide additional information for identifying the firm fixed effects, allowing us to exclude them from the regression without changing the result.

After running the regression with only movers, we can estimate the stayers' fixed effects using:

$$\widehat{\theta}_i = \frac{1}{T_i} \sum_t \left(y_{it} - \widehat{\psi}_{\mathcal{J}(i,t)} \right), \tag{E2}$$

where T_i is just the length of the time interval of worker i in the sample.

This sequential estimation method avoids combining all observations and covariates, reducing the size of the system of normal equations to be solved. It is memory efficient because the matrices needed for the normal equations with only movers are much smaller. This is important in applications like ours, with limited memory space and millions of fixed effects.

Advantages for leverage estimation. We can use the same logic to simplify the estimation of the leverages for each observation. As we explained in the main text, we estimate leverages by running regressions of Rademacher random variables on $\mathbf{X}$ and using either the fitted values or the residuals. The fitted value in the simple two-way fixed effect regression is

$$\widehat{y}_{it} = \widehat{\theta}_i + \widehat{\psi}_{\mathcal{J}(i,t)}.$$

Again, as stayers have no impact on the estimators of the firm fixed effects, we can then run the regressions to estimate the leverages using only the observations for the movers.

For the stayers, is even more simple. As explained in Online Appendix B, the leverage of observation i is just $\frac{\partial \widehat{y}_{it}}{\partial y_{it}}$. From equation (E2) above we have that, for stayers, their leverages are:

$$\frac{\partial \widehat{y}_{it}}{\partial y_{it}} = \frac{\partial \widehat{\theta}_{it}}{\partial y_{it}} + \frac{\partial \widehat{\psi}_{\mathcal{J}(i,t)}}{\partial y_{it}} = \frac{\partial \widehat{\theta}_{it}}{\partial y_{it}} + 0 = \frac{1}{T_i},$$

where we use $\frac{\partial \widehat{\psi}_{\mathcal{J}(i,t)}}{\partial y_{it}} = 0$ as, again, stayers do not affect the firm fixed effect estimator.

When workers within the same firm belong to multiple clusters, they are still estimable using the leave-cluster-out method. The projection matrix element for the row corresponding to observation it and the column corresponding to observation $i't'$ can be computed as $\frac{\partial \widehat{y}_{i't'}}{\partial y_{it}}$. In fact, the leverage is just a special case where $it = i't'$. By a similar argument than before, we have $\frac{\partial \widehat{y}_{i't'}}{\partial y_{it}} = 1/T_i$ if $i = i'$, and $\frac{\partial \widehat{y}_{i't'}}{\partial y_{it}} = 0$ if $i \neq i'$.

This implies that within a cluster, the projection matrix $\mathbf{P}$ and the residual matrix $\mathbf{M}$ are block diagonal. One block contains all the movers within the cluster, while the other diagonal blocks correspond to stayers. Consequently, we can treat these stayers as if they belong to a different cluster, separating them from the original cluster for analysis.

One simulation per stayer. To do the bootstrap we could simulate, for each observation, v^*_{it}. As explained above, we just run the regressions using only the movers to identify the firm fixed effects. A quick inspection of equation (E2), tell us that we do not need to simulate the entire vector for each observation of each stayer, but rather it suffices to simulate the *average*. In other words, if a stayer is T_i periods in the sample, there is no need to simulate T_i times a random variable v^*_{it} for each period. Instead we can simulate $\frac{1}{T_i}\sum_t v^*_{it}$. The only requirement is to make sure that the variance of the average is consistent with the variances of the individual v^*_{it}'s.

This means we only need to do a simulation per stayer. This increases the computation speed by reducing the number of simulations to do for each bootstrap. Also, it is much more memory efficient as it reduces the size of vectors of outcome variables to simulate during the bootstrap. The gains can be significant in typical applications where the majority of workers in the sample are stayers.

Solving the normal equations. The normal equations of the two-way fixed effect model support a Laplacian representation. Let $\mathbf{D}$ be the matrix of worker dummies and $\mathbf{F}$ the matrix of firm dummies,with one firm's fixed effect removed for normalization. Define the matrix of covariates as $\mathbf{X} = [\mathbf{D}, -\mathbf{F}]$. By changing the sign of the matrix $\mathbf{F}$, $\mathbf{X}^T\mathbf{X}$ becomes a Laplacian matrix. This is very helpful as systems of equations involving Laplacian matrices can be solved in a very efficient way. We solve for the normal equations in Matlab using the preconditioned conjugate gradient method, with the preconditioner from Koutis, Miller, and Tolliver (2011), optimized for this type of Laplacian systems.

F Sorting with alternative labor market definitions

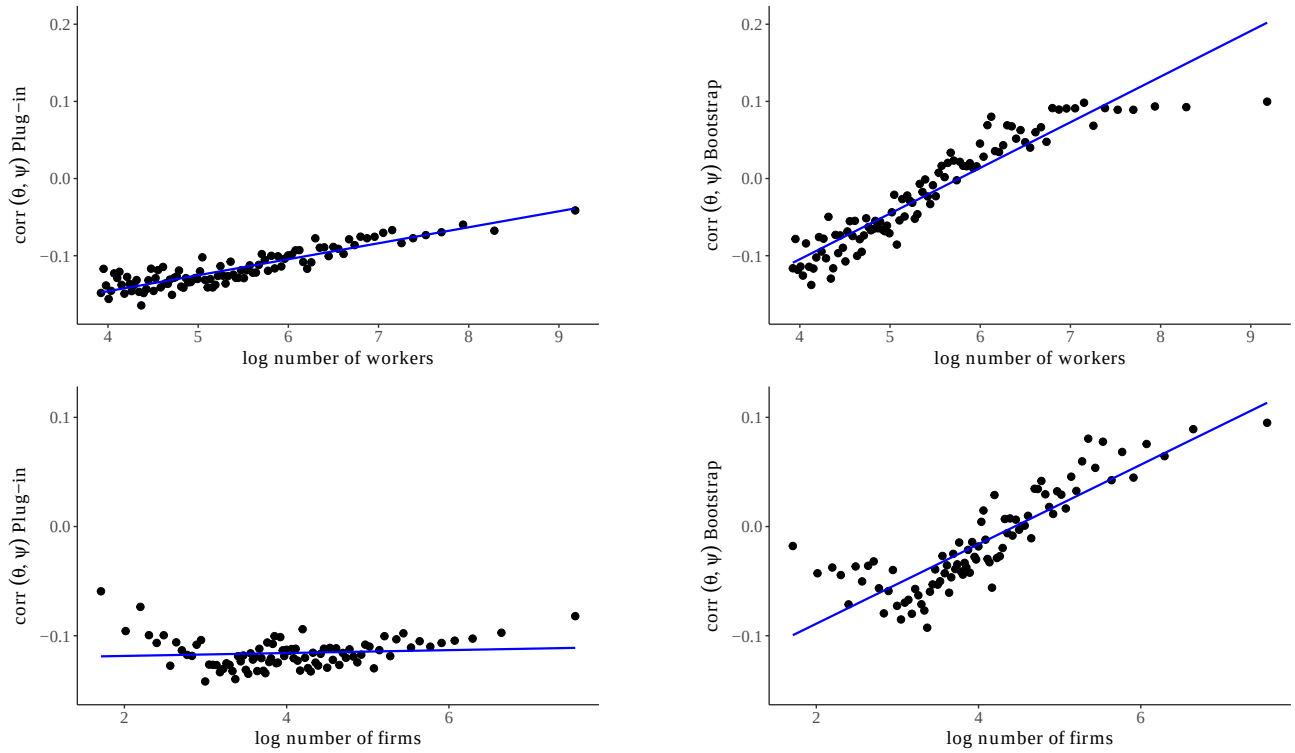
In this section, we present additional tables and figures that complement the analysis in the main text. First, we show that the results concerning sorting and labor market size remain robust when using an alternative and more granular definition of labor markets; here, they are defined as the combination of commuting zones and 4-digit occupations, rather than 2-digit occupations. Second, we do the same exercise but defining labor markets as commuting zones like Dauth et al. (2022), Leknes et al. (2022), and Pérez et al. (2023).

F.1 Labor markets: commuting zones $\times$ 4-digit occupations

Here we present the same figures as in the main text using the 2015-2019 sample but with the alternative definition of labor markets as the combination of commuting zones and 4-digit occupations. We first show the relationship between sorting direction and labor market size, followed by the relationship between sorting direction and labor market size.

Figure F1 and Table F1 show that when defining a labor market using 4-digit occupations, the slope is also greater after correcting for bias. Moreover, when measuring labor market size

Figure F1: Sorting direction and labor market size: CZ $\times$ 4-digit occupations



Notes: Binned scatter plots between sorting direction—the correlation between worker (θ) and firm (ψ) fixed effects—and labor market (combination of commuting zone and 4-digit occupations) size. x-axis: two different measures of size by the logarithm of the (i) number of workers for the top figures, and (ii) number of firms for the bottom figures. y-axis: on the left, plug-in estimates, on the right, bias-corrected estimates.

by the number of firms, the slope changes sign, becoming positive only with the bias-corrected estimates.

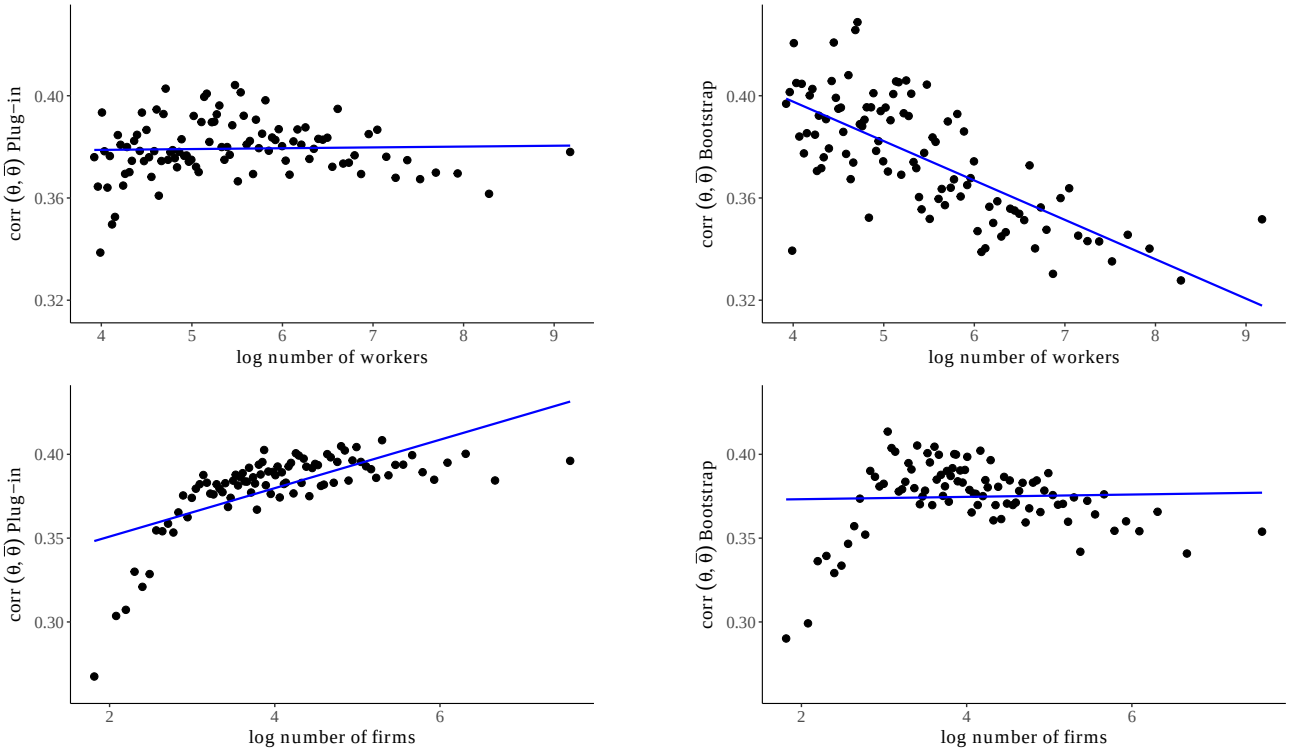
Regarding the relationship of sorting intensity and labor market size, Figure F2 and Table F1 show that the result in the main text is robust to this alternative definition of the labor market: when using the corrected estimates, the relationship changes sign and turns negative.

Table F1: Gradient of sorting on labor market size: CZ × 4-digit occupations

	Sorting Direction				Sorting Intensity			
	Plug-in		Bootstrap		Plug-in		Bootstrap	
log No. Workers	0.0208		0.0595		0.0003		-0.0175	
	(0.0010)		(0.0017)		(0.0008)		(0.0017)	
log No. Firms		0.0013		0.0359		0.0163		-0.0007
		(0.0010)		(0.0018)		(0.0008)		(0.0017)
Number of Markets	39,776				39,520			

Notes: Slope coefficients of an OLS regression of sorting direction—worker-firm correlation—, and sorting intensity—worker-coworker correlation—with different measures of labor market (combination of commuting zone and 4-digit occupations) size. Standard errors in parenthesis. *Plug-in*: slope estimate using plug-in estimates. *Bootstrap*: slope estimate using bootstrap-corrected estimates with the leave-match out covariance matrix estimator.

Figure F2: Sorting intensity and labor market size: CZ × 4-digit occupations



Notes: Binned scatter plots between sorting intensity—the correlation between worker fixed effects (θ) and the average of coworkers ($\bar{\theta}$)—and labor market (combination of commuting zone and 4-digit occupations) size. x-axis: two different measures of size by the logarithm of the (i) number of workers for the top figures, and (ii) number of firms for the bottom figures. y-axis: on the left, plug-in estimates, on the right, bias-corrected estimates.

F.2 Labor markets: commuting zones

We repeat the exercises from the previous section using the 2015-2019 sample, but now define labor markets based only on commuting zones, rather than using combinations of commuting zones and occupational codes.

When defining labor markets as combinations of commuting zones and occupations—whether at the 2-digit or 4-digit level—we found that corrected estimates strengthen the relationship between sorting direction and market size, while weakening the relationship between sorting intensity and market size. However, when defining labor markets using only commuting zones, these patterns are reversed. As shown in Figures F3 and F4, and Table F2, the relationship between market size and sorting direction weakens after correcting for limited mobility bias, while the relationship between sorting intensity and market size becomes stronger with the corrected estimates.

Previous studies have used this same definition of labor markets, allowing us to compare our results with theirs. For example, using data from Germany, Dauth et al. (2022) found a slope coefficient of 0.06 between the uncorrected worker-firm correlation and the log of population for the 2008-2014 period—slightly larger than our estimate of 0.05.[33] They reported a similar estimate after applying KSS corrections city-by-city. In contrast, we found a slope estimate of 0.023 after correcting for limited mobility bias, less than half the slope found with uncorrected correlations.

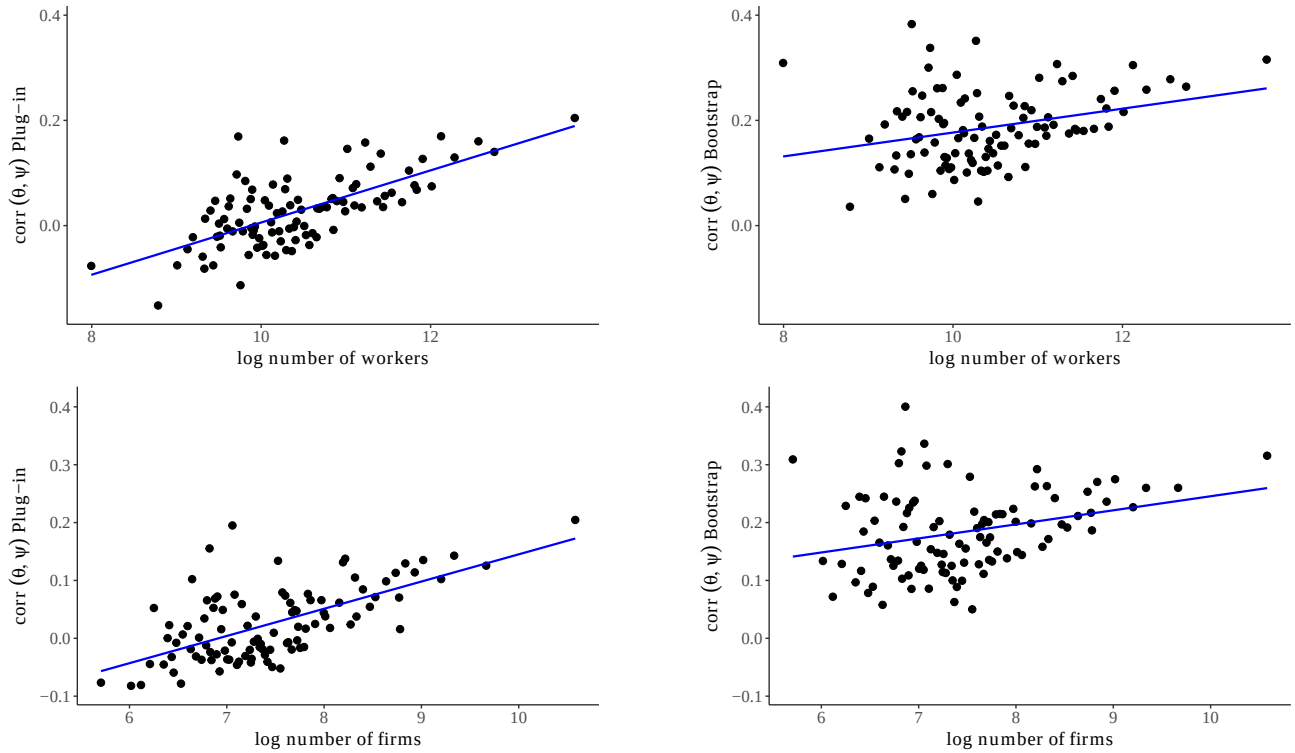
Similarly, Pérez et al. (2023), using data from Mexico, found a slope estimate of 0.045 between the log of the commuting zone population and the uncorrected worker-firm correlation. Like us, they observed a smaller slope coefficient (0.0244) after correcting for limited mobility bias.

Using data from Norway, Leknes et al. (2022) reported a slope coefficient of 0.025 based on uncorrected correlations. When using historical mining sites as instruments for current city populations, they found a larger slope estimate of 0.039.

Lastly, Dauth et al. (2022) study the relationship between sorting intensity and labor market size using uncorrected estimates (Table B.1 in their Online Appendix). They found a slope estimate of 0.047, higher than both our uncorrected (0.02) and corrected (0.03) slope estimates.

[33] We obtained similar slope estimates using our 2009-2014 sample.

Figure F3: Sorting direction and labor market size: CZ



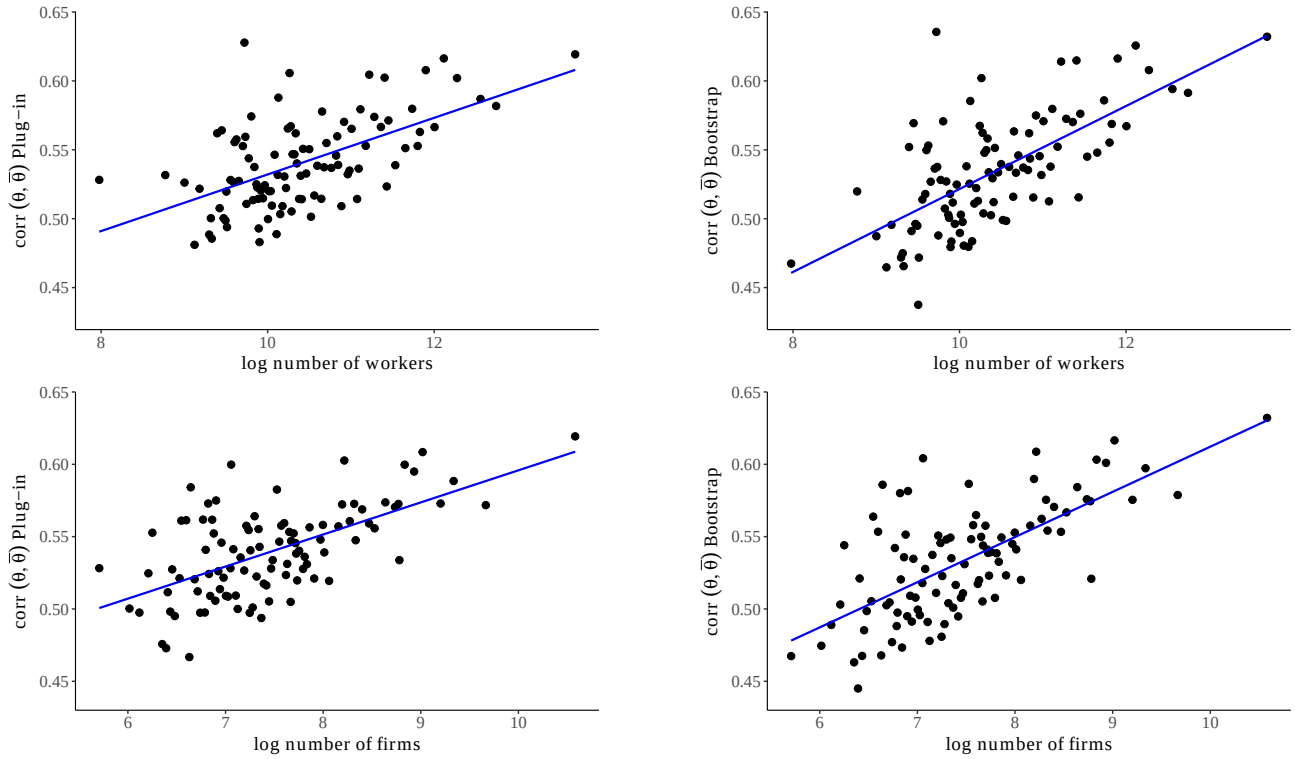
Notes: Binned scatter plots between sorting direction—the correlation between worker (θ) and firm (ψ) fixed effects—and labor market (commuting zone) size. x-axis: two different measures of size by the logarithm of the (i) number of workers for the top figures, and (ii) number of firms for the bottom figures. y-axis: on the left, plug-in estimates, on the right, bias-corrected estimates.

Table F2: Gradient of sorting on labor market size: CZ

	Sorting Direction				Sorting Intensity			
	Plug-in		Bootstrap		Plug-in		Bootstrap	
log No. Workers	0.0502		0.0232		0.0204		0.0300	
	(0.0048)		(0.0068)		(0.0026)		(0.0030)	
log No. Firms		0.0483		0.0235		0.0220		0.0313
		(0.0054)		(0.0073)		(0.0028)		(0.0032)
Number of Markets	287				287			

Notes: Slope coefficients of an OLS regression of sorting direction—worker-firm correlation—, and sorting intensity—worker-coworker correlation—with different measures of labor market (commuting zones) size. Standard errors in parenthesis. *Plug-in*: slope estimate using plug-in estimates. *Bootstrap*: slope estimate using bootstrap-corrected estimates with the leave-match out covariance matrix estimator.

Figure F4: Sorting intensity and labor market size: CZ



Notes: Binned scatter plots between sorting intensity—the correlation between worker fixed effects (θ) and the average of coworkers ($\bar{\theta}$)—and labor market (commuting zones) size. x-axis: two different measures of size by the logarithm of the (i) number of workers for the top figures, and (ii) number of firms for the bottom figures. y-axis: on the left, plug-in estimates, on the right, bias-corrected estimates.