Generalized AKM: Theory and Evidence*

Francesco Del Prato[†] Yaroslav Korobka[‡] Paolo Zacchia[§]

February 2026

Preliminary draft. Please do not circulate it.

Abstract

This paper develops an estimator for quadratic forms in semiparametric models where covariates enter through an unknown function. The leading application is a generalized AKM wage decomposition: we estimate worker and firm variance components and sorting while flexibly partialling out observed worker and firm characteristics. The estimator combines leave-one-out bias correction with series approximation, is robust to heteroskedasticity and many fixed effects, and is consistent when both fixed effects and basis terms grow with sample size under a strengthened smoothness condition. Simulations show large bias reductions relative to plug-in estimators in nonlinear settings. In linked Portuguese employer–employee data, adding rich controls to linear AKM generates large discontinuities, whereas the generalized estimator yields coherent movements across control sets and basis richness, with sorting attenuating toward zero as flexibility increases.

Keywords: wage decomposition, many regressors, semiparametric model, series estimator

*We are grateful to Edoardo M. Acabbi, Stanislav Anatolyev, Andrea Alati, Matias D. Cattaneo, Michal Kolesár, Mikkel Plagborg-Møller, Ulrich Müller, and seminar participants at Princeton University, XXIII Brucchi Luchino conference, and IRLE Berkeley for fruitful comments and suggestions. This paper results from research funded under the umbrella of the ERC-CZ project No. LL2319. Funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 870245 is also gratefully acknowledged.

[†] *Department of Economics and Business Economics, Aarhus University*. *e*-mail: francesco.delprato@econ.au.dk.

[‡] *CERGE-EI (Charles University & Czech Academy of Sciences)*. *e*-mail: yaroslav.korobka@cerge-ei.cz.

[§] *Department of Economics, Ca' Foscari University of Venice*; *CERGE-EI*; *CEPR*; and *IZA*. *e*-mail: Paolo.Zacchia@cerge-ei.cz.

1. Introduction

Why do some firms pay more than others, and how much of wage dispersion reflects workers, firms, or sorting? With linked employer–employee data, the AKM framework (Abowd, Kramarz, and Margolis 1999) is the workhorse reduced-form application for answering this question. Yet, standard plug-in variance-component estimators rely on linear covariate effects and can be biased under heteroskedasticity or limited mobility (Andrews et al. 2008; Kline, Saggio, and Sølvsten 2020; Cattaneo, Jansson, and Newey 2018b; Jochmans 2022). Modern applications feature many covariates and high-dimensional fixed effects; theory and evidence point to nonlinear wage profiles and worker–firm complementarities (Bonhomme, Lamadon, and Manresa 2019), so linear additivity—and common low-order fixes like cubic age profiles—can be a strong and potentially misspecified restriction, especially when several covariates are nonlinear (Mincer 1974; Card et al. 2018). This raises a question: how should one flexibly partial out many observables in AKM, and—once that is done correctly—how much wage variance is attributed to workers, firms, and sorting?

This paper develops estimation and inference for variance components (quadratic forms) when covariates enter the model through an unspecified function. In particular, we extend the familiar plug-in approach by Kline, Saggio, and Sølvsten (2020) based on a leave-one-out, bias-corrected estimator that is robust to arbitrary heteroskedasticity and to designs with many fixed effects. Our estimator further controls for a rich set of basis terms meant to approximate the actual effect of observed covariates. We establish consistency as the number of basis terms and fixed effects grow, and a chi-square limit under fixed rank, under a strengthened smoothness condition. Furthermore, we show asymptotic normality in the growing rank case. Simulations show a negligible bias. Empirically, flexible controls produce coherent movements in the decomposition rather than discontinuities: generalized estimates adjust systematically as controls and basis richness are enriched, while the linear AKM specification with the same controls displays large jumps. While there is semiparametric theory for estimators of linear functionals with growing bases, we are not aware of a framework that delivers validity for variance-component estimators when covariates enter through an unspecified function. Our results fill this gap by giving conditions under which variance components are consistently estimated and by characterizing their large-sample distributions.

We develop asymptotic results tailored to modern designs. Consistency holds when both the number of basis terms used to approximate the covariate effect and the number of fixed effects grow with the sample. For fixed-rank quadratic forms (e.g., testing a few linear contrasts of fixed effects), the statistic has a chi-square limit; for growing rank (e.g., analysis of variance across many groups), a Gaussian approximation applies. A key message is that quadratic targets are more demanding than linear ones: to control the approximation bias from the flexible covariate effect, we require a stronger smoothness condition than is standard for slope functionals. Moreover, the estimator scales to large datasets. A random-projection implementation approximates the leverage and influence quantities that enter the correction with negligible loss of accuracy, making high-order interactions and large panels computationally feasible.

Our work connects two strands of literature. On the "many regressors" side, we complement results on inference with many covariates and robust variance estimation (Cattaneo, Jansson, and Newey 2018b; Anatolyev 2012; Anatolyev and Sølvsten 2023), and connect to leave-one-out corrections and variance-component estimation in linear models (Kline, Saggio, and Sølvsten 2020; Jochmans 2022). On the semiparametric side, we build on series estimators with growing bases (Donald and Newey 1994; Cattaneo, Jansson, and Newey 2018a), showing that quadratic targets require stronger smoothness than slope functionals, and relate our asymptotics to many-instruments and small-bandwidth regimes (Hansen, Hausman, and Newey 2008; Cattaneo, Crump, and Jansson 2014). We also discuss an alternative kernel-based construction that fits naturally within the same leave-one-out logic (Cattaneo, Crump, and Jansson 2014).

We show the finite-sample behavior of our estimation through simulations. When the true covariate effect is nonlinear, the plug-in estimator exhibits non-negligible bias that can be amplified by heteroskedasticity and many fixed effects. The leave-one-out correction removes this bias across a range of designs, with accurate coverage for the proposed standard errors.

Finally, we revisit the AKM wage decomposition using linked employer–employee data with rich worker and firm observables. The allocation of variance matters because firm premia correlate with productivity, capital intensity, and downsizing or outsourcing policies (Card, Heining, and Kline 2013; Goldschmidt and Schmieder 2017; Bertheau et al. 2023). Benchmarks in the literature place the firm component around 15–25%, while recent bias-corrected methods suggest 5–16% (Bonhomme et al. 2023). In our data, the linear specification changes sharply when moving from a parsimonious to a richer control set: both worker and firm components jump, and estimated sorting changes sign. By contrast, under our generalized approach, changes are gradual and interpretable: richer controls and greater flexibility lower estimated worker and firm components and move sorting closer to zero. This contrast suggests that much of the instability comes from forcing rich, collinear controls through a linear additive structure. Allowing flexible, group-specific nonlinear adjustments absorbs this variation without the large distortions seen in the linear decomposition.

Outline. The remainder of the paper is organized as follows. Section 2 introduces the semiparametric model and formalizes the variance component targets. Section 3 derives the estimator and its finite-sample representation. Section 4 establishes consistency and limiting distributions under fixed and growing ranks, and discusses the strengthened smoothness condition. Section 5 presents the random projection approximation. Section 6 derives the limiting distribution of our proposed estimator. Section 7 reports our simulation exercise, while Section 8 illustrates an empirical application on linked employer-employee data. Section 9 concludes.

2. Setup

We consider the following semiparametric model,

$$y_i = x_i'\beta + f(z_i) + e_i, \quad i = 1, \ldots, n, \tag{1}$$

where the regressors $x_i \in \mathbb{R}^p$, $z_i \in \mathbb{R}^d$ are non-random. The unknown function $f(z)$ belongs to the class of smooth and real-valued functions, $f \in \mathcal{F}$. The unobserved errors $\{e_i\}_{i=1}^n$ are mutually independent and obey $\mathbb{E}[e_i] = 0$, but may possess observation-specific variances $\mathbb{E}[e_i^2] = \sigma_i^2$.

Our object of interest is a quadratic form $\theta := \beta' A\beta$ for some known non-random symmetric matrix $A \in \mathbb{R}^{p\times p}$ of rank r. This is a quantity of interest in a wide range of economic applications: two key examples are summarized next.

Example 1 (Generalized analysis of variance)**.** We discuss a particular example of the generalized analysis of covariance model where our results allow for unspecified functional relationship between the exogenous covariates and the outcome variable. Consider having observations arranged into N groups with T_g observations in the g-th group. Since Fisher (1928) originally introduced the analysis of variance methodology, it has been common since to assume the following model:

$$y_{gt} = \alpha_g + z_{gt}'\delta + \varepsilon_{gt}, \quad g = 1, \ldots, N, \quad t = 1, \ldots, T_g,$$

where α_g are group-specific fixed effects, and z_{gt} is a vector of exogenous covariates. The focus here is the variability in the outcome variable attributable to groups, or

$$\sigma_\alpha^2 = \frac{1}{n}\sum_{g=1}^N T_g\left(\alpha_g - \bar{\alpha}\right)^2$$

with $n := \sum_{g=1}^N T_g$, and $\bar{\alpha} := n^{-1}\sum_{g=1}^N T_g\alpha_g$. Our approach relaxes the linearity assumption on covariates and instead assumes any form of unspecified functional dependence, that is

$$y_{gt} = \alpha_g + f(z_{gt}) + \varepsilon_{gt}, \quad f \in \mathcal{F}, \quad g = 1, \ldots, N, \quad t = 1, \ldots, T_g.$$

We can represent it as in (1), defining $i := i(g,t)$ with $i(\cdot,\cdot)$ being a bijective function, $y_i := y_{gt}$, $z_i := z_{gt}$ and $e_i := \varepsilon_{gt}$,

$$x_i := d_i, \quad \beta := (\alpha_1, \ldots, \alpha_N)', \quad d_i := (\mathbb{1}\{g = 1\}, \ldots, \mathbb{1}\{g = N\})'.$$

Our object of interest σ_α^2 can be represented here as $\beta' A\beta$, with

$$A := \begin{pmatrix} A_d' A_d & 0 \\ 0 & 0 \end{pmatrix}, \quad A_d := \frac{1}{\sqrt{n}}(d_1 - \bar{d}, \ldots, d_n - \bar{d}), \quad \bar{d} := \frac{1}{n}\sum_{i=1}^n d_i.$$

Example 2 (Generalized AKM)**.** Our second and leading example is a classic wage decomposition model proposed in Abowd, Kramarz, and Margolis (1999). It models the log wage determination as an additive function of worker fixed effects, firm fixed effects, and a linear function of strictly exogenous covariates. Specifically,

$$y_{gt} = \alpha_g + \psi_{j(g,t)} + z_{gt}'\delta + \varepsilon_{gt}, \quad g = 1, \ldots, N, \quad t = 1, \ldots, T_g.$$

Here, α_g and $\psi_{j(g,t)}$ capture the gth worker and jth firm unobserved heterogeneity component respectively, and z_{gt} is a vector of exogenous regressors. The bijective function $j(\cdot, \cdot) : \{1, \ldots, N\} \times \{1, \ldots, \max_g T_g\} \to \{0, \ldots, J\}$ maps $n = \sum_{g=1}^{N} T_g$ person-year observations to one of $J + 1$ firms. One of the model's objectives is to quantify how much of the variability in log wages is determined by firms,

$$\sigma_\psi^2 = \frac{1}{n} \sum_{g=1}^{N} \sum_{t=1}^{T_g} \left(\psi_{j(g,t)} - \bar{\psi} \right)^2,$$

where $\bar{\psi} = \frac{1}{n} \sum_{g=1}^{N} \sum_{t=1}^{T_g} \psi_{j(g,t)}$. Given the empirical evidence that the relationship between the explanatory worker characteristics and the log wage is nonlinear (Mincer 1974; Card et al. 2018), the dominance of worker and firm fixed effects might partially stem from the functional linear restriction on the exogenous covariates. Our methodology allows for more plausible unspecified functional dependence between the outcome variable and regressors, that is,

$$y_{gt} = \alpha_g + \psi_{j(g,t)} + f(z_{gt}) + \varepsilon_{gt}, \quad f \in \mathcal{F}, \quad g = 1, \ldots, N, \quad t = 1, \ldots, T_g.$$

This formulation allows, for example, to capture rich patterns of non-linear interaction between both worker-level and firm-level characteristics, which possibly vary across observed groups in the population.

Given that the common covariates z_{gt} and the firm assignments $j(\cdot, \cdot)$ obey a strict exogeneity condition, we can rewrite the equation above as in (1) with

$$x_i := (d_i', h_i')', \quad \beta := (\alpha', \psi')', \quad \alpha := (\alpha_1, \ldots, \alpha_N)' + \mathbb{1}_N'\psi_0, \quad \psi = (\psi_1, \ldots, \psi_J)' - \mathbb{1}_J'\psi_0,$$

defining y_i, z_i, and e_i as in Example 1, and $h_i := (\mathbb{1}\{j(g,t) = 1\}, \ldots, \mathbb{1}\{j(g,t) = J\})'$. The parameter of interest σ_ψ^2 then can be rewritten as $\beta' A_\psi \beta$ with

$$A_\psi := \begin{pmatrix} 0 & 0 & 0 \\ 0 & A_h' A_h & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad A_h := \frac{1}{\sqrt{n}}(h_1 - \bar{h}, \ldots, h_n - \bar{h}), \quad \bar{h} := \frac{1}{n} \sum_{i=1}^{n} h_i.$$

3. Finite-sample properties

From now on, we restrict the analysis to model (1). To derive an estimator of β, one must regress y_i on x_i and *functions* of z_i. To this end, let $p^1(z), \ldots, p^k(z)$ be some *approximating* functions, and let $p_k(z) := (p^1(z), \ldots, p^k(z)) \in \mathbb{R}^k$ be a vector of covariates $p_k(z_i)$ aimed at approximating $f(z_i)$. We assume that the class of functions $\mathcal{F}$ to which f belongs can be well approximated by linear combinations of elements in $p_k(z_i)$.

Define M_{ij} as a (i,j)th element of $M := I_n - P_k(P_k'P_k)^{-1}P_k'$ with

$$P_k := (p_k(z_1), \ldots, p_k(z_n)) \in \mathbb{R}^{n \times k},$$

and let the partialed-out design matrix $S_{xx} := \sum_{i=1}^n \sum_{j=1}^n M_{ij} x_i x_j'$ have a full column rank. Then, the estimator of β is defined as

$$\hat{\beta} := S_{xx}^{-1} \sum_{i=1}^n \sum_{j=1}^n M_{ij} x_i y_j.$$

Under some regularity conditions, the bias of $\hat{\beta}$ is negligible for large k. Rewrite it as:

$$\begin{aligned}
\hat{\beta} &= S_{xx}^{-1} \sum_{i=1}^n \sum_{j=1}^n M_{ij} x_i \left(x_j'\beta + f(z_j) + e_j \right) \\
&= S_{xx}^{-1} \sum_{i=1}^n \sum_{j=1}^n M_{ij} x_i x_j \beta' + S_{xx}^{-1} \sum_{i=1}^n \sum_{j=1}^n M_{ij} x_i f(z_j) + S_{xx}^{-1} \sum_{i=1}^n \sum_{j=1}^n M_{ij} x_i e_j \\
&= \beta + S_{xx}^{-1} \sum_{i=1}^n \sum_{j=1}^n M_{ij} x_i f(z_j) + S_{xx}^{-1} \sum_{i=1}^n \sum_{j=1}^n M_{ij} x_i e_j := \beta + B + U.
\end{aligned}$$

Because of the zero-mean assumption on errors, $\mathbb{E}[U] = 0$ holds, and the bias in the estimator is reflected in B term only. Under some conditions discussed below,

$$B = o(1)$$

as the number of approximating functions goes to infinity, $k \to \infty$.

Analogously to the fully parametric setup examined by Kline, Saggio, and Sølvsten (2020), the plug-in estimator of the quadratic form $\hat{\theta}_{\text{PI}} := \hat{\beta}' A \hat{\beta}$ is biased, because

$$\mathbb{E}[\hat{\theta}_{\text{PI}} - \theta] = \operatorname{trace}(A \operatorname{var}[\hat{\beta}]) = \sum_{i=1}^n B_{ii} \sigma_i^2,$$

where $B_{ii} := \sum_{j=1}^n M_{ij} x_j' S_{xx}^{-1} A S_{xx}^{-1} \sum_{j=1}^n M_{ij} x_j$ measures the influence of the ith squared error e_i^2 on $\hat{\theta}_{\text{PI}}$. In cases when $p/n \approx 0$, individual elements B_{ii} are close to zero, and the bias is negligible. However, with increasing p, the bias is more pronounced and needs to be accounted for.

We introduce some notation to describe our estimator. Collect x_i in a matrix as $X := (x_1, \ldots, x_n)' \in \mathbb{R}^{n \times p}$. Then, $W := (X, P_k) \in \mathbb{R}^{n \times p+k}$ is a matrix containing full set of regressors, with individual vectors w_i, and $W = (w_1, \ldots, w_n)'$. Define

$$\hat{\gamma} := \left(\sum_{i=1}^n w_i w_i' \right)^{-1} \sum_{i=1}^n w_i y_i,$$

and the leave-one-out version of it,

$$\hat{\gamma}_{-i} := \left(\sum_{i=1}^{n} w_i w_i' - w_i w_i'\right)^{-1} \left(\sum_{i=1}^{n} w_i y_i - w_i y_i\right).$$

Also, note that first p elements of $\hat{\gamma}$ are $\hat{\beta}$, and the remaining k elements are basis coefficients from the function approximation. Similarly to Kline, Saggio, and Sølvsten (2020), we propose to estimate θ as

$$\hat{\theta} := \hat{\beta}' A \hat{\beta} - \sum_{i=1}^{n} B_{ii} \hat{\sigma}_i^2, \tag{2}$$

where $\hat{\sigma}_i^2$ is a leave-one-out estimator of the individual variance of e_i defined as

$$\hat{\sigma}_i^2 := y_i \left(y_i - w_i' \hat{\gamma}_{-i}\right). \tag{3}$$

We note that computing $\hat{\gamma}_{-i}$ for each $i = 1, \ldots, n$ is computationally costly in large-scale applications. To avoid this, we represent (3) as

$$\hat{\sigma}_i^2 = \frac{y_i \hat{e}_i}{M_{W,ii}}, \tag{4}$$

where $M_{W,ii}$ is ith diagonal element of the matrix $M_W := I_n - W(W'W)^{-1}W'$ projecting onto the complement of the column space spanned by x_i and functions of z_i, and $\hat{e}_i := \sum_{j=1}^{n} M_{ij}(y_j - x_j \hat{\beta})$ are residuals.

Our estimator $\hat{\theta}$ is *exactly* unbiased if *(i)* $\hat{\beta}$ is unbiased, *(ii)* $\hat{\sigma}_i^2$ is unbiased for each i. For a moment, suppose that $\mathbb{E}[\hat{\beta}] = \beta$ holds and $f \in \mathcal{F}_{\text{linear}}$: that is, the estimator of the "slope" β is unbiased, and the unknown function $f(z) = z'\alpha$ is linear in z. Under these conditions, we have

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}_i^2] &= \mathbb{E}\left[y_i \left(y_i - x_i' \hat{\beta}_{-i} - p_k(z_i)' \hat{\alpha}_{-i}\right)\right] \\
&= \mathbb{E}\left[\left(x_i'\beta + f(z_i) + e_i\right)\left(x_i'\beta + f(z_i) + e_i - x_i'\hat{\beta}_{-i} - p_k(z_i)'\hat{\alpha}_{-i}\right)\right] \\
&= \mathbb{E}\left[\left(x_i'\beta + f(z_i) + e_i\right)\left(x_i'(\beta - \hat{\beta}_{-i}) + (f(z_i) - p_k(z_i)'\hat{\alpha}_{-i}) + e_i\right)\right] \\
&= \left(x_i'\beta x_i' + f(z_i) x_i'\right) \mathbb{E}[\beta - \hat{\beta}_{-i}] + \left(x_i'\beta + f(z_i)\right) \mathbb{E}[f(z_i) - p_k(z_i)'\hat{\alpha}_{-i}] + \sigma_i^2 \\
&= \sigma_i^2
\end{aligned}$$

where we use the fact that $\mathbb{E}[e_i x_i'(\beta - \hat{\beta}_{-i})] = \mathbb{E}[e_i]\mathbb{E}[x_i'(\beta - \hat{\beta}_{-i})] = 0$, because $\hat{\beta}_{-i}$ is independent of e_i, and similarly $\mathbb{E}[e_i(f(z_i) - p_k(z_i)'\hat{\alpha}_{-i})] = \mathbb{E}[e_i]\mathbb{E}[f(z_i) - p_k(z_i)'\hat{\alpha}_{-i}] = 0$, because $\hat{\alpha}_{-i}$ is independent of e_i.

However, for general classes of functions $\mathcal{F}$, the exact unbiasedness of $\hat{\sigma}_i^2$ does not hold. Instead, we rely on asymptotic approximations where we let the number of approximating functions to go to infinity, $k \to \infty$. Hence, the estimator of the slope $\hat{\beta}$, of the individual variances $\hat{\sigma}_i^2$ and, as a result, of the quadratic form $\hat{\theta}$ will be unbiased *asymptotically*[1], that

1. For expositional purposes, we sometimes drop the Euclidean norm on the vector $\|v\|$, and write asymptotic results as, say, $v = o_p(1)$ instead of $\|v\| = o_p(1)$.

is

$$\mathbb{E}[\hat{\beta}] - \beta = o(1), \quad \mathbb{E}[\hat{\sigma}_i^2] - \sigma_i^2 = o_p(1), \quad \mathbb{E}[\hat{\theta}] - \theta = o_p(1),$$

as $k \to \infty$.

For these purposes, we conventionally assume that the unknown function $f(z)$ belongs to the functional class $\mathcal{F}$ of smooth and real-valued functions that can be well-approximated by the functions of their arguments.

Assumption 1. *We assume that $f \in \mathcal{F}$, where*

$$\mathcal{F} := \left\{ f : \min_{\alpha \in \mathbb{R}^k} \mathbb{E}\left[|f(z_i) - p_k(z_i)'\alpha|^2 \right] \leq Ck^{-2\alpha_f}, \alpha_f > 1 \right\} \tag{5}$$

for some absolute constant $C < \infty$.

This particular implies that

$$\sup_{f \in \mathcal{F}} \mathbb{E}\left[|f(z_i) - p_k(z_i)'\hat{\alpha}_{-i}|^2 \right] = \mathcal{O}_p(k^{-2\alpha_f}),$$

so that all the functions in the class are if z_i has a compact support.

The constant α_f plays an important role in our asymptotic analysis and highlights the point of departure from the results in the previous literature on semiparametric estimation. Specifically, to obtain the consistency result of our estimator, we need to strengthen the assumption on the constant α_f from being strictly positive (Donald and Newey 1994; Cattaneo, Jansson, and Newey 2018a) to being strictly greater than one. Intuitively, because we are interested in the estimator of the quadratic form, the bias should decrease to zero faster than in cases when we are solely interested in estimators that are linear in the outcome variable (for example, the estimator of β in fixed-p designs). This extra assumption is the cost of relaxing the linearity assumption on the covariates, and bears practical implications in the form of the number of exogeneous variables, and assumptions on the degree of smoothness of $f(z)$ the researcher is willing to consider. [2]

For example, if the support of z_i is compact, while the basis functions (say, polynomials or splines), satisfy Assumption 1 with $\alpha_f = s_f/d$, where s_f is the number of continuous derivatives of $f(z)$. Given that $\alpha_f > 1$, it implies that such functions will satisfy the assumption if $s_f > d$, that is, if the function possesses more continuous derivatives than the dimensionality of z_i (Chen 2007). Practically speaking, it implies that if the researcher has relatively many exogenous covariates, she implicitly assumes that the underlying function is very smooth. We stress that there are studies (Armstrong and Kolesár 2018) that highlight the impossibility of the degree of smoothness estimation from the data, so our practical recommendation in empirical applications is to carefully look for an *ad hoc* appropriate balance between the number of covariates and robustness of the results.

2. We leave discussions on whether $\alpha_f > 1$ is a sufficient or a necessary condition for consistency for future research.

4. Consistency

In this section, we prove the consistency result for the proposed estimator $\hat{\theta}$. We study the asymptotic behavior of $\hat{\theta}$ assuming that x_i, z_i, and A are sequences of constants so that the only source of randomness is e_i. We adopt the conditional perspective approach as in Scheffe (1959), Searle, Casella, and McCulloch (2009), and Kline, Saggio, and Sølvsten (2020). This allows us to be agnostic about the potential dependency between x_i and z_i, and A. Here is also where our analysis is different from Cattaneo, Jansson, and Newey (2018b) where the authors consider sequences of random variables and conduct analysis conditional on z_i only. Limits are taken assuming the number of observations goes to infinity, $n \to \infty$, the number of approximating functions goes to infinity, $k \to \infty$ (so that the finite-sample bias of $\hat{\beta}$ and $\hat{\sigma}_i^2$ is asymptotically of negligible order), and the dimensionality of x_i goes to infinity, $p \to \infty$, to model the limited mobility bias. We make the following additional assumptions:

Assumption 2. *(i)* $\max_i(\mathbb{E}[e_i^4] + \sigma_i^{-2}) = \mathcal{O}(1)$; *(ii) there exists a* $c < 1$ *such that* $\max_i P_{W,ii} < c$ *for all* n*, where* $P_{W,ii} := I_n - M_{W,ii}$; *(iii)* $\max_i(x_i'\beta)^2 = \mathcal{O}(1)$.

Part *(i)* excludes heavy-tailed distributions of the errors as is typically assumed in the literature on OLS estimation (Cattaneo, Jansson, and Newey 2018b; Kline, Saggio, and Sølvsten 2020). Parts *(ii)* and *(iii)* coupled with the Assumption 1 imply that the leave-one-out estimator $\hat{\sigma}_i^2$ is well-defined and has bounded variance. Part *(ii)* also implies that $\frac{p+k}{n} \leq c < 1$ for all n.

Assumption 3. *We assume that* $k \to \infty$ *as* $n \to \infty$.

Below we formalize the consistency result for the proposed estimator.

Lemma 3. *If Assumption 1, 2, and 3 hold, A is positive semi-definite,* $\theta = \beta' A \beta = \mathcal{O}(1)$*, and* $trace(\tilde{A}^2) = \sum_{\ell=1}^{r} \lambda_\ell^2 = o(1)$*, then*

$$\hat{\theta} - \theta \xrightarrow{p} 0.$$

4.1. Alternative estimation with kernel methods

In this subsection we briefly discuss an alternative way for estimating the quadratic form θ: one based on kernel methods. By still assuming that the data-generating process takes the form in (1), we now estimate the unknown function $f(z)$ at the point z as

$$\hat{f}(z) = \frac{\sum_{i=1}^{n} K\left(\frac{z - z_i}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{z - z_i}{h}\right)}, \quad \hat{f}_{-i}(z) = \frac{\sum_{j=1, j \neq i}^{n} K\left(\frac{z - z_j}{h}\right) y_j}{\sum_{j=1, j \neq i}^{n} K\left(\frac{z - z_j}{h}\right)}, \tag{6}$$

for some suitable kernel function $K(\cdot)$ that fulfills regularity requirements. This form of the "leave-one-out" kernel estimator is extensively used in the nonparametric literature on cross-validation or estimation of density-weighted average derivatives (Cattaneo, Crump, and Jansson, 2014). After obtaining the estimate of the function $\hat{f}_{-i}(z)$, we construct the estimator of the slope as:

$$\hat{\beta}_{\mathrm{a}} = \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1} \sum_{i=1}^{n} x_i \hat{\xi}_i, \quad \hat{\xi}_i := y_i - \hat{f}(z_i),$$

and the estimator of the quadratic form as:

$$\hat{\theta}_{\mathrm{a}} := \hat{\beta}_{\mathrm{a}}' A \hat{\beta}_{\mathrm{a}} - \sum_{i=1}^{n} B_{ii} \hat{\sigma}_{i,\mathrm{a}}^2$$

with

$$\hat{\sigma}_{i,\mathrm{a}}^2 := y_i \left(y_i - x_i' \hat{\beta}_{-i,\mathrm{a}} - \hat{f}_{-i}(z_i)\right), \quad \hat{\beta}_{-i,\mathrm{a}} := \left(\sum_{i=1}^{n} x_i x_i' - x_i x_i'\right)^{-1} \left(\sum_{i=1}^{n} x_i \hat{\xi}_i - x_i \hat{\xi}_i\right).$$

The leave-one-out kernel estimator (6) eliminates the "self-observation" bias when constructing residuals $\hat{\xi}_i$ in the first step. This form of sample-splitting is for example also used (and for similar reasons at that) in the literature on double-debiased machine learning (Chernozhukov et al. 2018). We conjecture that given conventional asymptotic conditions on the bandwidth and the sample size used in nonparametrics, namely $n \to \infty$, $h \to 0$, and $nh \to \infty$, the estimator $\hat{\theta}_{\mathrm{a}}$ is consistent, in line with the Lemma 3.

5. Large-scale approximation

In this section we discuss an alternative estimator that allows for fast computation in typical large-scale applications such as those based on administrative linked employer-employee data. We follow Kline, Saggio, and Sølvsten (2020) by considering the random projection method of Achlioptas (2003). To describe it, fix $m \in \mathbb{N}$ and generate matrices $R_B, R_P \in \mathbb{R}^{m \times n}$ where each (i, j) coordinate is a random draw from the following Rademacher distribution:

$$R_{\cdot,ij} = \begin{cases} +1 & \text{with probability } 1/2, \\ -1 & \text{with probability } 1/2. \end{cases}$$

Decompose $A = 1/2(A_1' A_2 + A_2' A_1)$ for $A_1, A_2 \in \mathbb{R}^{n \times p}$, where $A_1 = A_2$ if A is positive semi-definite. Denote

$$\hat{P}_{W,ii} := \frac{1}{m} \left\| R_P W S_{ww}^{-1} w_i \right\|^2, \quad \hat{B}_{ii} := \frac{1}{m} \left(R_B A_1 S_{xx}^{-1} \sum_{j=1}^{n} M_{ij} x_j\right)' \left(R_B A_2 S_{xx}^{-1} \sum_{j=1}^{n} M_{ij} x_j\right).$$

The proposed estimator is then:

$$\hat{\theta}_{\mathrm{JLA}} := \hat{\beta}' A \hat{\beta} - \sum_{i=1}^{n} \hat{B}_{ii} \hat{\sigma}_{i,\mathrm{JLA}}^2, \quad \hat{\sigma}_{i,\mathrm{JLA}}^2 := \frac{y_i (y_i - w_i' \hat{\gamma})}{1 - \hat{P}_{W,ii}} \left(1 - \frac{1}{m} \frac{3\hat{P}_{W,ii}^3 + \hat{P}_{W,ii}^2}{1 - \hat{P}_{W,ii}}\right). \tag{7}$$

As in Kline, Saggio, and Sølvsten (2020), the term $\frac{1}{m}\frac{3\hat{P}_{W,ii}^3+\hat{P}_{W,ii}^2}{1-\hat{P}_{W,ii}}$ removes a non-linearity bias from the approximation of $P_{W,ii}$ by $\hat{P}_{W,ii}$.

Lemma 4. *If Assumptions 1, 2, and 3 are satisfied, $n/m^4 = o(1)$, $\mathrm{var}[\hat{\theta}]^{-1} = \mathcal{O}(n)$, and one of the following conditions hold, then $\mathrm{var}[\hat{\theta}]^{-1/2}(\hat{\theta}_{JLA} - \hat{\theta} - \mathrm{B}_m) = o_p(1)$ where $|\mathrm{B}_m| \leq \frac{1}{m}\sum_{i=1}^n P_{W,ii}^2|B_{ii}|\sigma_i^2$:*

(i) A is positive semi-definite and $\mathbb{E}[\hat{\beta}'A\hat{\beta}] - \theta = \sum_{i=1}^n B_{ii}\sigma_i^2 = \mathcal{O}(1)$.

(ii) $A = 1/2(A_1'A_2 + A_2'A_1)$ where $\theta_1 = \beta'A_1'A_1\beta$ and $\theta_2 = \beta'A_2'A_2\beta$ satisfy (i) and $\frac{\mathrm{var}[\hat{\theta}_1]\,\mathrm{var}[\hat{\theta}_2]}{n\,\mathrm{var}[\hat{\theta}]^2} = \mathcal{O}(1)$.

6. Limiting distributions

This section examines two different results about the limiting distribution of $\hat{\theta}$. The first assumes that the rank of A: r, is fixed. Instead, the second result allows for growing r.

6.1. Fixed rank

First, we derive the limiting distribution of $\hat{\theta}$ for fixed r. We represent the estimator as

$$\hat{\theta} = \sum_{\ell=1}^{r} \lambda_\ell \left(\hat{b}_\ell^2 - \widehat{\mathrm{var}}[\hat{b}_\ell]\right),$$

where $\hat{b} := \sum_{i=1}^n v_i y_i$, $\widehat{\mathrm{var}}[\hat{b}] := \sum_{i=1}^n v_i v_i'\hat{\sigma}_i^2$, and $v_i := Q'S_{xx}^{-1/2}\sum_{j=1}^n M_{ij}x_j$.

Below we state a formal result stating that the distribution of $\hat{\theta}$ is a sum of potentially dependent chi-squared random variables with non-centralities $b = (b_1, \ldots, b_r)'$, and that the estimator of the variance of $\hat{b}$ is consistent.

Theorem 5. *If Assumptions 1, 2, and 3 hold, r is fixed, and $\max_i v_i'v_i = o(1)$, then*

(i) $\mathrm{var}[\hat{b}]^{-1/2}(\hat{b} - b) \xrightarrow{d} \mathcal{N}(0, I_r)$, where $b := Q'S_{xx}^{1/2}\beta$,

(ii) $\mathrm{var}[\hat{b}]^{-1}\widehat{\mathrm{var}}[\hat{b}] \xrightarrow{p} I_r$,

(iii) $\hat{\theta} = \sum_{\ell=1}^r \lambda_\ell \left(\hat{b}_\ell^2 - \mathrm{var}[\hat{b}_\ell]\right) + o_p(\mathrm{var}[\hat{\theta}]^{1/2})$.

This result is standard, though arguably of limited interest in empirical settings like the AKM model, where the rank of A grows with the sample size.

6.2. Growing rank

We can also motivate $\hat{\theta}$ using a generalized "full set" of regressors $\tilde{w}_i := \check{A}S_{ww}^{-1}w_i$, where $\check{A} := \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix}$ and $\gamma := \begin{pmatrix} \beta & \alpha \end{pmatrix}'$. Indeed,

$$\theta = \gamma'\check{A}\gamma = \gamma'S_{ww}S_{ww}^{-1}\check{A}\gamma = \sum_{i=1}^{n} \gamma' w_i \tilde{w}_i'\gamma$$

and the asymptotically unbiased leave-one-out estimator is then

$$\hat{\theta} = \sum_{i=1}^{n} y_i \tilde{w}_i' \hat{\gamma}_{-i},$$

where the difference between $\gamma' w_i$ and its proxy y_i goes in expectation squared to zero as the number of approximating functions becomes large. Now we can represent

$$\hat{\theta} = \sum_{i=1}^{n} \sum_{\ell \neq i} C_{i\ell} y_i y_\ell,$$

where $C_{i\ell} := B_{W,i\ell} - 2^{-1} M_{W,i\ell} \left(B_{W,ii} M_{W,ii}^{-1} + B_{W,\ell\ell} M_{W,\ell\ell}^{-1} \right)$. Hence, the estimator of the quadratic form is a second-order U-statistic with the kernel $C_{i\ell}$. Below we show, using the supporting Lemma 7 provided in the Appendix, that the kernel varies with the sample size n in a way that the individual contributions are small enough leading to the Gaussian approximation. The analogical result is shown in Kline, Saggio, and Sølvsten (2020) in the context of quadratic form estimation under the linear specification, and Cattaneo, Jansson, and Newey (2018a) in the context of slope estimation in the semiparametric model.

To show normality of the proposed estimator, we decompose $\hat{\theta} - \theta$ into components that satisfy conditions of Lemma 7, and then rely on results of the Lemma to imply the joint normality of individual components. Compared to Kline, Saggio, and Sølvsten (2020), the difference between the estimator and the estimand now includes the error from the functional approximation, which is, however, of negligible order under the Assumption 1 and the assumption that the number of approximating functions becomes large.

Theorem 6. *If Assumption 1, 2, and 3 hold, and the following conditions are satisfied,*

$$\text{(i) } \operatorname{var}[\hat{\theta}]^{-1} \max_i \left((\tilde{w}_i' \gamma)^2 + (\breve{w}_i' \gamma)^2 \right) = o(1), \quad \text{(ii) } \frac{\lambda_1^2}{\sum_{\ell=1}^{r} \lambda_\ell^2} = o(1),$$

then $\operatorname{var}[\hat{\theta}]^{-1/2} (\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, 1)$.

7. Simulation study

We conduct a simulation study to verify the unbiasedness of the proposed estimator in different contexts. Our data-generating process follows (1) with the unknown function defined as $f(z_i) = \|z_i\|^7$. We generate $n = 500$ observations with $z_{\ell i} \sim$ i.i.d. $\mathcal{U}(-1, 1)$, and $x_{\ell i} = \exp\left(0.3 \|z_i\|^2 + 0.7\mathcal{N}(0,1)\right)$, fixing dimensionality of z_i to $d = 4$ and varying dimensionality of x_i as $p = \{20, 90, 180, 300\}$. The slope coefficients are generated as $\beta_j \sim \mathcal{U}(-0.5, 0.5)$ for $j = 1, \ldots, p$, and we set the A matrix to $A = x_1 x_1' + x_2 x_2'$. The explanatory power of x_i remains fixed to $R_x^2 = 0.7$ as we vary its dimensionality.

We evaluate the performance of the proposed estimator in *(i)* homoskedastic regime with errors generated as $e_i \sim$ i.i.d. $\mathcal{N}(0, 1)$, *(ii)* heteroskedastic regime with errors generated as $e_i = \mathcal{N}(0,1) + x_{1i}/12$, and different values of k for approximating $f(z_i)$, $k = \{4, 30, 50\}$. Table 1 displays the values of $|\frac{1}{s}\sum_{\ell=1}^{s}\hat{\theta}_\ell - \theta|$ for different settings, where $\hat{\theta}_\ell$ is an estimate of the quadratic form in ℓth simulation as defined in (2), and the total number of simulations is $s = 1000$.

Table 1: Simulation study: results

$k \setminus p$	20	90	180	300
		Homoskedastic		
4 (linear)	2.19	3.50	5.47	2.55
30 (3rd power + int.)	0.10	0.04	0.00	0.63
50 (5th power + int.)	0.06	0.00	0.10	0.51
		Heteroskedastic		
4 (linear)	2.50	8.59	7.24	11.04
30 (3rd power + int.)	0.36	0.32	0.59	0.06
50 (5th power + int.)	0.08	0.01	0.22	0.02

Note: Entries report absolute bias of the estimated quadratic form, averaged across $s = 1000$ Monte Carlo replications with sample size $n = 500$ in each replication. Rows vary basis richness for the nonparametric component: $k = 4$ is linear, $k = 30$ uses third-order polynomials with interactions, and $k = 50$ uses fifth-order polynomials with interactions. Columns vary the number of linear regressors p.

In sum, given a high nonlinearity of the underlying function, we observe that a large number of approximating functions is required to eliminate the bias. A linear specification leads to a severe bias, while a cubic specification seems to not capture the nonlinearities fully, still leading to a bias (though of a much smaller magnitude). Approximation relying on polynomials and interactions of the fifth power appears to be the most robust solution, addressing estimation issues due to heteroskedasticity, high dimensionality of x_i, as well as the (unknown) nonparametric component of the model.

We highlight that approximations based on cubic polynomials are fairly common in practice: see e.g. Card et al. (2018). Our simulation study shows that in some situations this might not be sufficient to fully eliminate the bias, potentially leading to distorted final estimates in actual applications.

8. Empirical application

We apply our estimator to decompose wage dispersion in linked employer–employee data. The exercise illustrates that relaxing linearity in the covariates yields stable and interpretable variance components, in contrast to the standard linear AKM model once rich controls are introduced.

8.1. Data and estimation

Our empirical analysis combines two Portuguese sources of administrative data. The matched employer–employee dataset *Quadros de Pessoal* (QP) reports worker characteristics (age, gender, occupation, qualification, education, contract type, hours, compensation) and firm characteristics (location, industry, revenues) for the universe of private-sector firms. The firm accounts dataset *Central de Balanços* (CB) provides annual balance-sheet and income-statement information (including total assets and sales) for non-financial corporations. We link QP and CB at the firm level to obtain a panel with rich worker- and firm-side observables.

We focus on the period 2008–2018 and on full-time workers aged 20–65. The model's dependent variable is the logarithm of individual hourly wages. As in the typical "AKM" literature, we decompose the cross-sectional variance of wages into the variance of worker fixed effects (σ_α^2), the variance of firm fixed effects (σ_ψ^2), and their covariance $\mathrm{cov}(\alpha, \psi)$, which captures sorting.

We compare two empirical specifications. The first is the standard linear two-way fixed effects (AKM) model with covariates entering additively and linearly. The second implements our generalized semiparametric estimator, which approximates $f(z)$ via polynomial bases of degree 1, 3, and 5 (denoted Generalized (1)/(3)/(5)). All polynomial bases are interacted with a fully saturated set of observed worker categorical variables, allowing for group-specific nonlinear effects. The Generalized (1) model already differs from Standard AKM: while the latter enters continuous covariates additively without interactions, Generalized (1) allows group-specific linear effects of each continuous variable.

To connect this specification to the theoretical framework of Section 2, let $s_i \in \{1, \ldots, G\}$ denote the worker categorical group (e.g., gender, or gender $\times$ education $\times$ qualification). The group-specific model

$$y_i = x_i'\beta + \sum_{s=1}^{G} \mathbb{1}\{s_i = s\}\, f_s(z_i^{\text{cont}}) + e_i$$

is a special case of (1) with the unknown function redefined on the expanded argument $\tilde{z}_i := (z_i^{\text{cont}\prime}, s_i)'$. The natural basis for this piecewise-smooth function is the set of indicator-weighted polynomials $\{\mathbb{1}\{s_i = s\} \cdot p^j(z_i^{\text{cont}})\}_{s,j}$, which is exactly the interacted basis P_k used in our implementation. Because the discrete indicators partition the sample rather than adding continuous dimensions, the smoothness condition (Assumption 1) applies within each group at rate $\alpha_f = s_f / d_{\text{cont}}$, where d_{cont} is the number of continuous covariates and s_f is the number of their continuous derivatives. Note that the interaction groups consist of worker categorical variables, not firm identifiers: including firm identifiers would make P_k nearly collinear with the firm fixed-effect indicators in X, jeopardizing the full-rank condition on $S_{xx} = X'MX$ required by Assumption 2.

We report estimates under a parsimonious control set (gender, age) and an extended set that adds education and qualification at the worker level and revenue and assets at the firm level.[3]

3. In practice, we compute all estimates by first partialling out covariates (which, in the "Generalized" cases, amount to the fully interacted set of polynomial bases P_k); subsequently, we compute the bias-corrected variance components of interest on the resulting residual. All variance components in Table 3 are reported as shares of total log-wage variance.

Table 2: Firm controls: within-firm time variation and missingness

	Non-missing firm-year	Firms (≥2y)	Mean years (firms ≥2y)	Within share (log) (%)	Zero within SD (%)
A. Firm-control time variation in estimation sample					
Revenue	1,874,310	290,932	6.22	18.79%	1.67%
Assets	1,874,310	290,932	6.22	6.43%	0.72%
Workers	2,397,757	383,679	5.99	7.13%	27.11%
B. Missingness by year (firm-year panel)					
Year	Firm-years	Revenue miss.	Assets miss.	Workers miss.	
2008	253,005	26.27%	26.27%	0.00%	–
2009	245,077	25.37%	25.37%	0.00%	–
2010	223,267	23.66%	23.66%	0.00%	–
2011	218,229	23.01%	23.01%	0.00%	–
2012	202,744	22.47%	22.47%	0.00%	–
2013	198,597	21.20%	21.20%	0.00%	–
2014	202,625	20.36%	20.36%	0.00%	–
2015	207,037	19.89%	19.89%	0.00%	–
2016	211,892	19.39%	19.39%	0.00%	–
2017	215,908	18.80%	18.80%	0.00%	–
2018	219,376	18.24%	18.24%	0.00%	–

Note: Panel A reports non-missing firm-year observations and within-firm time variation of firm controls in the estimation sample (2008–2018), restricting the within-variation calculations to firms observed for at least two years. The within-share statistic is the share of total log variance attributable to within-firm over-time variation. Panel B reports year-by-year missingness rates for the same firm controls in the firm-year panel; worker counts use a MEE-based fallback when SCIE firm controls are missing.

Detailed sample descriptives and connected-component coverage are reported in Appendix Tables 4 and 5. The main patterns are that the cleaned panel is large, and the largest connected set retains most usable observations and workers.

Table 2 characterizes the firm-level controls that distinguish the two panels. Within-firm time variation in logs is limited: about 18.8% for revenue, 6.4% for assets, and 7.1% for workers. Missingness is material though declining for revenue and assets (roughly 26% in 2008 to 18% in 2018), while workers missingness is 0% as we impute back missing employment from the matched employer-employee data. Because most of the variation in these controls is cross-sectional, adding them as linear regressors in the presence of firm fixed effects introduces near-collinearity—a point that will be central to interpreting the results below.

8.2. Results

Table 3 presents the estimates of the variance components for each model and set of controls.

Within Panel A (parsimonious controls), comparisons across methods are already informative. Relative to Standard, the generalized specifications produce substantially lower worker variance (σ^2_α: 0.7175 $\to$ 0.4853 at degree 5) and moderately lower firm variance (σ^2_ψ: 0.1818 $\to$ 0.1466). Note that sorting slightly increases from Standard to Generalized (1) (cov(α, ψ): 0.0815 $\to$ 0.0957): the standard model's restricted age profile

Table 3: Wage variance decomposition

Model	σ^2_α	σ^2_ψ	$\text{cov}(\alpha, \psi)$
Panel A: Parsimonious controls			
Standard	0.7175	0.1818	0.0815
Generalized (1)	0.5000	0.1570	0.0957
Generalized (3)	0.4890	0.1482	0.0875
Generalized (5)	0.4853	0.1466	0.0863
Panel B: Extended controls			
Standard	2.8077	0.9420	−0.1390
Generalized (1)	0.4667	0.1600	0.0456
Generalized (3)	0.3152	0.1397	0.0144
Generalized (5)	0.2654	0.1270	0.0120

Note: Entries report bias-corrected (KSS) variance components as shares of total log-wage variance: worker-effect variance (σ^2_α / $\text{var}(y)$), firm-effect variance (σ^2_ψ / $\text{var}(y)$), and their covariance ($\text{cov}(\alpha, \psi)$ / $\text{var}(y)$). Values above one indicate that the estimated component exceeds total wage dispersion, which is possible because variance components need not sum to one (the residual absorbs the remainder and can be negative in bias-corrected decompositions). Panel A uses parsimonious controls (gender, age); Panel B adds education, qualification, revenue, and assets. "Standard" is the linear AKM specification; "Generalized (1/3/5)" use semiparametric approximations of increasing polynomial degree with group-specific interactions.

partially conflates observable nonlinearities with fixed effects, and allowing group-specific linear effects reallocates some of this variation, temporarily raising measured sorting. Within the generalized family, increasing the polynomial degree from 1 to 5 then yields monotone declines in all three moments (σ^2_α: 0.5000 $\rightarrow$ 0.4853, σ^2_ψ: 0.1570 $\rightarrow$ 0.1466, $\text{cov}(\alpha, \psi)$: 0.0957 $\rightarrow$ 0.0863), consistent with richer approximation absorbing nonlinear observable variation that would otherwise load on fixed effects.

The central contrast appears when moving to Panel B (extended controls). Under Standard AKM, all moments shift sharply: σ^2_α jumps from 0.7175 to 2.8077, σ^2_ψ rises from 0.1818 to 0.9420, and $\text{cov}(\alpha, \psi)$ flips sign from $+0.0815$ to -0.1390. Variance components that exceed total log-wage variance and a sign reversal in sorting are hallmarks of a pathological decomposition, indicating that the linear specification is severely destabilized by the addition of rich continuous controls.

Under the generalized estimator, Panel B behaves much more coherently. At degree 1, moments remain close to their Panel A magnitudes ($\sigma^2_\alpha = 0.4667$, $\sigma^2_\psi = 0.1600$, $\text{cov} = 0.0456$). At higher degrees, worker and firm variances decline further and sorting attenuates toward zero (Generalized (5): $\sigma^2_\alpha = 0.2654$, $\sigma^2_\psi = 0.1270$, $\text{cov} = 0.0120$). Richer controls and richer basis functions thus reduce measured premia and sorting in a systematic, monotone way, rather than generating sign reversals and scale explosions.

Why does the linear specification break down so sharply? Revenue and assets are predominantly cross-sectional: as documented in Table 2, within-firm variance shares are only 6–19%. Adding these variables as linear regressors therefore introduces strong collinearity with firm fixed effects. In the bias-corrected (KSS) estimator, near-collinearity inflates the leverage and influence quantities that enter the correction, amplifying rather than removing instability. The generalized estimator sidesteps this problem because the flexible, group-specific basis P_k can absorb the nonlinear relationship between firm characteristics and wages without forcing that variation through the fixed effects.

Overall, the relevant comparison is not "no movement" versus "movement," but coherent versus incoherent movement. The generalized estimates move in a structured, interpretable way across panels and degrees; the linear AKM estimates exhibit large discontinuities.

8.3. Interpreting the role of flexible controls

Our generalized specification interacts polynomial bases of continuous variables (age in Panel A; age, revenue, and assets in Panel B) with saturated worker categorical groups (gender in Panel A; gender $\times$ education $\times$ qualification in Panel B). This means the effective covariate adjustment varies across demographic groups: the return to age or to working at a high-revenue firm can differ for, say, high-educated men versus low-educated women.

Under this specification, moving from Panel A to Panel B lowers both fixed-effect variances and especially the covariance term, which approaches zero at higher degree. This suggests that part of what appears as worker–firm sorting in linear AKM reflects misspecified functional form in observed characteristics rather than a genuine correlation between unobserved worker and firm types.

Several caveats apply. First, this conclusion speaks only to heterogeneity along observed z_{it}. Match-specific premia or complementarities outside z_{it} can still matter (e.g., a manager–worker fit that raises pay on a particular team). Second, because the interaction groups are defined by worker categories rather than firm identifiers, the model does not estimate firm-specific pay schedules *per se*. Firms differ in their effective covariate adjustment only to the extent that they employ different mixes of worker types and have different levels of the continuous firm controls. Allowing firm-level interaction groups is a natural extension that we leave for future work. Third, our decomposition is static: time-varying firm policies or match dynamics can exist without affecting the estimated variance components.

Overall, the evidence points to a decomposition in which flexible treatment of observables materially changes the allocation between premia and explained variation, but does so in a coherent way. The key empirical message is that allowing nonlinear, group-specific returns to observables prevents the large distortions that arise when rich controls are forced through a linear AKM structure.

9. Conclusions

This paper develops a semiparametric framework for estimating variance components, (as in the workhorse AKM model of wage decomposition) when covariates enter flexibly through an unknown function. We extend extant leave-one-out bias corrections to this setting and show that, under strengthened smoothness conditions, the resulting estimator is consistent and asymptotically normal. The approach remains computationally feasible in large-scale applications.

Simulations demonstrate that our proposed method eliminates the bias of plug-in estimators under heteroskedasticity and nonlinear covariate effects. In an empirical application to matched employer–employee data, generalized estimates move coherently as controls are enriched and basis flexibility increases, while linear AKM estimates display large discontinuities (including sharp inflation in both worker and firm variance components and a covariance sign reversal). This evidence suggests that instability in standard decompositions is primarily due to functional-form restrictions rather than substantive nonlinearities in pay determination.

Relative to conventional methods, our estimator requires additional assumptions and comes with a slightly higher computational cost. In actual applications, these disadvantages must be weighed against the estimator's ability to deliver results that are both more nuanced and different relative to standard approaches. To provide more guidance on this issue, in future work we plan on comparing the empirical implications of our method across a diverse set of administrative datasets, all of which shall ideally feature rich sets of observed covariates.

References

Abowd, J., F. Kramarz, and D. Margolis. 1999. "High wage workers and high wage firms." *Econometrica* 67 (2): 251–333.

Anatolyev, S. 2012. "Inference in regression models with many regressors." *Journal of Econometrics* 170 (2): 368–382.

Anatolyev, S., and M. Sølvsten. 2023. "Testing many restrictions under heteroskedasticity." *Journal of Econometrics* 236 (1): 1–19.

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 (3): 673–697.

Armstrong, T., and M. Kolesár. 2018. "Optimal inference in a class of regression models." *Econometrica* 86 (2): 655–683.

Bertheau, A., E. M. Acabbi, C. Barceló, A. Gulyas, S. Lombardi, and R. Saggio. 2023. "The unequal consequences of job loss across countries." *American Economic Review: Insights* 5, no. 3 (September): 393–408.

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 (2): 291–322.

Bonhomme, S., T. Lamadon, and E. Manresa. 2019. "A distributional framework for matched employer employee data." *Econometrica* 87 (3): 699–739.

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 (1): 13–70.

Card, D., J. Heining, and P. Kline. 2013. "Workplace heterogeneity and the rise of West German wage inequality." *The Quarterly Journal of Economics* 128 (3): 967–1015.

Cattaneo, M. D., R. K. Crump, and M. Jansson. 2014. "Small bandwidth asymptotics for density-weighted average derivatives." *Econometric Theory* 30 (1): 176–200.

Cattaneo, M. D., M. Jansson, and W. K. Newey. 2018a. "Alternative asymptotics and the partially linear model with many regressors." *Econometric Theory* 34 (2): 277–301.

———. 2018b. "Inference in linear regression models with many covariates and heteroscedasticity." *Journal of the American Statistical Association* 113 (523): 1350–1361.

Chen, X. 2007. "Large sample sieve estimation of semi-nonparametric models." *Handbook of Econometrics* 6:5549–5632.

Donald, S. G., and W. K. Newey. 1994. "Series estimation of semilinear models." *Journal of Multivariate Analysis* 50 (1): 30–40.

Fisher, Ronald. 1928. *Statistical Methods for Research Workers.* 5th ed. Oliver / Boyd.

Goldschmidt, D., and J. Schmieder. 2017. “The rise of domestic outsourcing and the evolution of the German wage structure.” *The Quarterly Journal of Economics* 132 (3): 1165–1217.

Hansen, C., J. Hausman, and W. K. Newey. 2008. “Estimation with many instrumental variables.” *Journal of Business & Economic Statistics* 26 (4): 398–422.

Jochmans, K. 2022. “Heteroscedasticity-robust inference in linear regression models with many covariates.” *Journal of the American Statistical Association* 117 (538): 887–896.

Kline, P., R. Saggio, and M. Sølvsten. 2020. “Leave-out estimation of variance components.” *Econometrica* 88 (5): 1859–1898.

Mincer, Jacob. 1974. *Schooling, Experience, and Earnings.* National Bureau of Economic Research, Inc.

A. Proofs

Here, we provide proofs of the main results found in the paper and additional derivations.

Proof of Lemma 3. The variance of $\hat{\beta}$ is

$$\begin{aligned}
\operatorname{var}[\hat{\beta}] &= \operatorname{var}\left[S_{xx}^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}x_i f(z_j) + S_{xx}^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}x_i e_j\right] \\
&= S_{xx}^{-1}\operatorname{var}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}x_i f(z_j) + \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}x_i e_j\right] S_{xx}^{-1} \\
&= S_{xx}^{-1}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}^2 x_i x_i' \sigma_j^2\right) S_{xx}^{-1}
\end{aligned}$$

under the assumption of fixed x_i and z_i.

First, rewrite the difference between the estimator and the estimand as

$$\begin{aligned}
\hat{\theta} - \theta &= \hat{\beta}' A \hat{\beta} - \beta' A \beta - \sum_{i=1}^{n} B_{ii}\hat{\sigma}_i^2 \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij} y_j x_i' S_{xx}^{-1} A S_{xx}^{-1} \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij} x_i y_j - \beta' A \beta - \sum_{i=1}^{n} B_{ii}\hat{\sigma}_i^2.
\end{aligned}$$

Given a data-generating process for y_i, we can expand $\hat{\theta} - \theta$ further as

$$\begin{aligned}
\hat{\theta} - \theta &= \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}\left(x_j'\beta + f(z_j) + e_j\right) x_i' S_{xx}^{-1} A S_{xx}^{-1} \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij} x_i \left(x_j'\beta + f(z_j) + e_j\right) - \beta' A \beta - \sum_{i=1}^{n} B_{ii}\hat{\sigma}_i^2 \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij} x_j'\beta x_i S_{xx}^{-1} A S_{xx}^{-1} \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij} x_i x_j'\beta + \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij} f(z_j) x_i' S_{xx}^{-1} A S_{xx}^{-1} \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij} x_i f(z_j) \\
&+ \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij} e_j x_i' S_{xx}^{-1} A S_{xx}^{-1} \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij} x_i e_j + \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij} x_j'\beta x_i' S_{xx}^{-1} A S_{xx}^{-1} \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij} x_i f(z_j) \\
&+ \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij} f(z_j) x_i' S_{xx}^{-1} A S_{xx}^{-1} \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij} x_i x_j'\beta + \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij} x_j'\beta x_i S_{xx}^{-1} A S_{xx}^{-1} \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij} x_i e_j \\
&+ \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij} e_j x_i' S_{xx}^{-1} A S_{xx}^{-1} \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij} x_i x_j'\beta + \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij} f(z_j) x_i' S_{xx}^{-1} A S_{xx}^{-1} \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij} x_i e_j \\
&+ \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij} e_j x_i' S_{xx}^{-1} A S_{xx}^{-1} \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij} x_i f(z_j) - \beta' A \beta - \sum_{i=1}^{n} B_{ii}\hat{\sigma}_i^2.
\end{aligned}$$

Next, we use the definition $B_{i\ell} := \sum_{j=1}^{n} M_{ij} x_j' S_{xx}^{-1} A S_{xx}^{-1} \sum_{j=1}^{n} M_{\ell j} x_j$, so that

$$\begin{aligned}
\hat{\theta} - \theta &= \sum_{i=1}^{n}\sum_{\ell=1}^{n} B_{i\ell} f(z_i)^2 + \sum_{i=1}^{n}\sum_{\ell=1}^{n} B_{i\ell} e_i^2 - \sum_{i=1}^{n} B_{ii}\hat{\sigma}_i^2 \\
&\quad + 2\sum_{i=1}^{n}\sum_{\ell=1}^{n} B_{i\ell} x_\ell'\beta f(z_i) + 2\sum_{i=1}^{n}\sum_{\ell=1}^{n} B_{i\ell} x_\ell'\beta e_i + 2\sum_{i=1}^{n}\sum_{\ell=1}^{n} B_{i\ell} e_\ell f(z_i).
\end{aligned}$$

Finally, rearranging, we have

$$\begin{aligned}\hat{\theta}-\theta=&\sum_{i=1}^{n}\sum_{\ell=1}^{n}B_{i\ell}f(z_i)^2+\sum_{i=1}^{n}\sum_{\ell\neq i}B_{i\ell}e_ie_\ell+\sum_{i=1}^{n}B_{ii}(e_i^2-\hat{\sigma}_i^2)\\&+2\sum_{i=1}^{n}\sum_{\ell=1}^{n}B_{i\ell}x_\ell'\beta f(z_i)+2\sum_{i=1}^{n}\sum_{\ell=1}^{n}B_{i\ell}x_\ell'\beta e_i+2\sum_{i=1}^{n}\sum_{\ell=1}^{n}B_{i\ell}e_\ell f(z_i).\end{aligned}\tag{8}$$

In general, it holds that

$$\mathbb{E}[|\hat{\theta}-\theta|^2]=|\mathbb{E}[\hat{\theta}-\theta]|^2+\text{trace}(\text{var}[\hat{\theta}-\theta]).$$

To show that $\hat{\theta}$ is consistent for θ, we need to show that the bias and the variance of the difference in (8) goes to zero. The main idea is to compute bounds on each term and their variances, and show that these bounds are asymptotically negligible. Then, convergence in the quadratic mean would imply convergence in probability.

Applying expectations on both sides, and using independence and mean-zero properties of errors, we have

$$\mathbb{E}[\hat{\theta}-\theta]=\sum_{i=1}^{n}\sum_{\ell=1}^{n}B_{i\ell}f(z_i)^2+2\sum_{i=1}^{n}\sum_{\ell=1}^{n}B_{i\ell}x_\ell'\beta f(z_i)+\sum_{i=1}^{n}B_{ii}\mathbb{E}[\sigma_i^2-\hat{\sigma}_i^2].$$

We now prove that each term goes to zero in probability as $n\to\infty$, $k\to\infty$, and $p\to\infty$.

Denote $B:=(B_{i\ell})_{i,\ell=1}^n\in\mathbb{R}^{n\times n}$, and $F:=(f(z_1),\ldots,f(z_n))'\in\mathbb{R}^n$. Then

$$\begin{aligned}\sum_{i=1}^{n}\sum_{\ell=1}^{n}B_{i\ell}f(z_i)^2&=F'BF\\&=F'MXS_{xx}^{-1}AS_{xx}^{-1}X'MF\\&=F'MXS_{xx}^{-1}A^{1/2}A^{1/2}S_{xx}^{-1}X'MF\\&=(F'MXS_{xx}^{-1}A^{1/2})^2,\end{aligned}$$

where we use $A=A^{1/2}A^{1/2}$ because A is symmetric. By the Markov inequality, assumptions on $\mathcal{F}$, M being idempotent, and the Cauchy-Schwartz inequality,

$$\begin{aligned}\left\|\frac{1}{n}F'MXS_{xx}^{-1}A^{1/2}\right\|&\leq\text{trace}\left(\frac{1}{n}F'MF\right)^{1/2}\cdot\text{trace}\left(\frac{1}{n}A^{1/2}S_{xx}^{-1}X'MXS_{xx}^{-1}A^{1/2}\right)^{1/2}\\&=\text{trace}\left(\frac{1}{n}F'MF\right)^{1/2}\cdot\text{trace}\left(\frac{1}{n}A^{1/2}S_{xx}^{-1}A^{1/2}\right)^{1/2}\\&=\mathcal{O}(k^{-\alpha_f}\sqrt{p/n}).\end{aligned}$$

Thus, using the Assumption 1, we have

$$\sum_{i=1}^{n}\sum_{\ell=1}^{n}B_{i\ell}f(z_i)^2=\mathcal{O}(k^{-2\alpha_f}np)=\mathcal{O}(k^{-\alpha_f}n)\to 0.$$

Now, for the second term in the main decomposition,

$$\begin{aligned}\sum_{i=1}^{n}\sum_{\ell=1}^{n} B_{i\ell}x_\ell'\beta f(z_i) &= F'BX\beta \\ &= F'MXS_{xx}^{-1}AS_{xx}^{-1}X'MX\beta \\ &= F'MXS_{xx}^{-1}A\beta.\end{aligned}$$

Given that $\beta' AS_{xx}^{-1}A\beta = \mathcal{O}(1)$, by the Markov inequality, assumption on $\mathcal{F}$, M being idempotent, and the Cauchy-Schwartz inequality we have

$$\begin{aligned}\left\|\frac{1}{n}F'MXS_{xx}^{-1}A\beta\right\| &\leq \text{trace}\left(\frac{1}{n}F'MF\right)^{1/2}\cdot\text{trace}\left(\frac{1}{n}\beta'AS_{xx}^{-1}X'MXS_{xx}^{-1}A\beta\right)^{1/2} \\ &= \text{trace}\left(\frac{1}{n}F'MF\right)^{1/2}\cdot\text{trace}\left(\frac{1}{n}\beta'AS_{xx}^{-1}A\beta\right)^{1/2} \\ &= \mathcal{O}(k^{-\alpha_f}/\sqrt{n}).\end{aligned}$$

So that using the Assumption 1, we have

$$\sum_{i=1}^{n}\sum_{\ell=1}^{n} B_{i\ell}x_\ell'\beta f(z_i) = \mathcal{O}(k^{-\alpha_f}\sqrt{n}) \to 0.$$

To bound the third term, we should show that

$$\mathbb{E}[\hat{\sigma}_i^2 - \sigma_i^2] = \left(x_i'\beta + f(z_i)\right)\left(x_i'\mathbb{E}[\beta - \hat{\beta}_{-i}] + \mathbb{E}[f(z_i) - p_k(z_i)'\hat{\alpha}_{-i}]\right)$$

goes in probability to zero. This is equivalent to bounding the bias of the estimator of the slope coefficient and the bias of the function approximation.

The bias of the slope estimator is

$$\mathbb{E}[\hat{\beta} - \beta] = \mathbb{E}\left[\left(\frac{1}{n}X'MX\right)^{-1}\frac{1}{n}X'MF\right].$$

By the Markov inequality, assumption on $\mathcal{F}$, M being idempotent, and the Cauchy-Schwartz inequality we have that:

$$\begin{aligned}\left\|\frac{1}{n}X'MF\right\| &\leq \text{trace}\left(\frac{1}{n}X'MX\right)^{1/2}\cdot\text{trace}\left(\frac{1}{n}F'MF\right)^{1/2} \\ &= \mathcal{O}(k^{-\alpha_f}\sqrt{p/n}),\end{aligned}$$

and, similarly,

$$\left\|\frac{1}{n}X'MX\right\| = \mathcal{O}(p/n),$$

so that using Assumption 1, we ultimately have:

$$\mathbb{E}[\|\hat{\beta} - \beta\|] = \mathcal{O}(\sqrt{n}/k^{\alpha_f}\sqrt{p}) \to 0.$$

Because of the assumption on the functional class $\mathcal{F}$, we can bound the bias of the function approximation using Jensen's inequality for convex $x \mapsto x^2$,

$$\begin{aligned}\mathbb{E}[|f(z_i) - p_k(z_i)'\hat{\alpha}_{-i}|] &\leq \mathbb{E}[|f(z_i) - p_k(z_i)'\hat{\alpha}_{-i}|^2]^{1/2} \\ &\leq (Ck^{-2\alpha_f})^{1/2} = \mathcal{O}_p(k^{-\alpha_f}),\end{aligned}$$

so that:

$$\mathbb{E}[|\hat{\sigma}_i^2 - \sigma_i^2|] = \mathcal{O}_p(\sqrt{n}/k^{\alpha_f}\sqrt{p}) \xrightarrow{p} 0.$$

From this it follows that:

$$\mathbb{E}[|\hat{\theta} - \theta|] \xrightarrow{p} 0.$$

We next turn our attention to bounding the variance. Let matrix

$$\tilde{A} := S_{xx}^{-1/2} A S_{xx}^{-1/2},$$

and let $\lambda_1, \ldots, \lambda_r$ be its nonzero eigenvalues. We assume that $\lambda_1^2 \geq \ldots \geq \lambda_r^2$, and that each eigenvalue appears as many times as its algebraic multiplicity. Under these assumptions, we can spectrally decompose $\tilde{A} = QDQ'$, where Q is a matrix of orthonormal vectors and $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_r)$.

The variance of $2\sum_{i=1}^n \sum_{\ell=1}^n B_{i\ell} x_\ell' \beta e_i$ is

$$\begin{aligned}4\sum_{i=1}^n \left(\sum_{\ell=1}^n B_{i\ell} x_\ell' \beta\right)^2 \sigma_i^2 &\leq \max_i \sigma_i^2 \beta' X' B^2 X \beta = \max_i \sigma_i^2 \beta' A S_{xx}^{-1} A \beta \\ &\leq \max_i \sigma_i^2 \lambda_1 \theta = o(1).\end{aligned}$$

To explain why the last inequality holds, define $\tilde{\beta} = S_{xx}^{1/2}\beta$ so that $\beta = S_{xx}^{-1/2}\tilde{\beta}$,

$$\beta' A S_{xx}^{-1} A \beta = \tilde{\beta}' S_{xx}^{-1/2} A S_{xx}^{-1/2} \tilde{\beta} = \tilde{\beta}'(S_{xx}^{-1/2} A S_{xx}^{-1/2})^2 \tilde{\beta} = \tilde{\beta}' \tilde{A}^2 \tilde{\beta},$$

and

$$\theta = \tilde{\beta}' S_{xx}^{-1/2} A S_{xx}^{-1/2} \tilde{\beta} = \tilde{\beta}' \tilde{A} \tilde{\beta}.$$

Then, using the Rayleigh quotient argument, it holds that

$$\frac{\tilde{\beta}' \tilde{A}^2 \tilde{\beta}}{\tilde{\beta}' \tilde{A} \tilde{\beta}} \leq \lambda_{\max}(\tilde{A}),$$

because $\tilde{A}$ is positive semi-definite, $\tilde{\beta} \neq 0$, and $\lambda_{\max}(A)$ is the largest eigenvalue of some matrix A. From this it follows that we can bound

$$\beta' A S_{xx}^{-1} A \beta \leq \lambda_1 \theta,$$

use assumptions $\theta = \mathcal{O}(1)$, $\lambda_1 \leq \operatorname{trace}(\tilde{A}^2)^{1/2} = o(1)$, and the fact that the variance is bounded. The conclusion then follows.

The variance of $2\sum_{i=1}^n \sum_{\ell=1}^n B_{i\ell} e_\ell f(z_i)$ is

$$4\sum_{i=1}^n \sum_{\ell=1}^n B_{i\ell}^2 \sigma_\ell^2 f(z_i)^2 \leq \max_i 4\sigma_i^2 f(z_i)^2 \sum_{i=1}^n \sum_{\ell=1}^n B_{i\ell}^2 = \max_i 4\sigma_i^2 f(z_i)^2 \text{trace}(\tilde{A}^2) = o(1),$$

which follows since

$$\begin{aligned}
\sum_{i=1}^n \sum_{\ell=1}^n B_{i\ell}^2 = \|B\|_F^2 = \text{trace}(B'B) \\
&= \text{trace}(MXS_{xx}^{-1}AS_{xx}^{-1}AS_{xx}^{-1}X'M) \\
&= \text{trace}(MXS_{xx}^{-1/2}S_{xx}^{-1/2}AS_{xx}^{-1/2}S_{xx}^{-1/2}AS_{xx}^{-1/2}S_{xx}^{-1/2}X'M) \\
&= \text{trace}(S_{xx}^{-1/2}AS_{xx}^{-1/2}S_{xx}^{-1/2}AS_{xx}^{-1/2}S_{xx}^{-1/2}X'MXS_{xx}^{-1/2}) \\
&= \text{trace}(S_{xx}^{-1/2}AS_{xx}^{-1/2}S_{xx}^{-1/2}AS_{xx}^{-1/2}) = \text{trace}(\tilde{A}^2),
\end{aligned}$$

and we use the assumptions as above but now, instead of the bounded variance, we assume that $\mathcal{F}$ is a class of bounded functions so that $\max |f(z_i)| < C$ for $i = 1, \ldots, n$ for some absolute constant C (this is implied by the main assumption on $\mathcal{F}$).

Because $B_{i\ell}^2 = B_{\ell i}^2$ for any i, ℓ, the variance of $\sum_{i=1}^n \sum_{\ell \neq i} B_{i\ell} e_i e_\ell$ is

$$2\sum_{i=1}^n \sum_{\ell \neq i} B_{i\ell}^2 \sigma_i^2 \sigma_\ell^2 \leq \max_i 2\sigma_i^4 \sum_{i=1}^n \sum_{\ell=1}^n B_{i\ell}^2 = \max_i 2\sigma_i^4 \text{trace}(\tilde{A}^2) = o(1).$$

To compute the variance of $\sum_{i=1}^n B_{ii}(e_i^2 - \hat{\sigma}_i^2)$, we represent the leave-one-out variance estimator as

$$\begin{aligned}
\hat{\sigma}_i^2 &= y_i(y_i - w_i'\hat{\gamma}_{-i}) = y_i M_{W,ii}^{-1}(y_i - w_i'\hat{\gamma}) \\
&= y_i M_{W,ii}^{-1}\hat{e}_i = y_i M_{W,ii}^{-1} \sum_{\ell=1}^n M_{W,i\ell} y_\ell \\
&= y_i M_{W,ii}^{-1} \sum_{\ell=1}^n M_{W,i\ell} \left(x_\ell'\beta + f(z_\ell) + e_\ell\right) \\
&= y_i M_{W,ii}^{-1} \sum_{\ell=1}^n M_{W,i\ell} \left(x_\ell'\beta + f(z_\ell)\right) + y_i M_{W,ii}^{-1} \sum_{\ell=1}^n M_{W,i\ell} e_\ell.
\end{aligned}$$

Thus, the variance is now expressed as:

$$\begin{aligned}
\sum_{i=1}^n \left(\sum_{\ell=1}^n M_{W,\ell\ell}^{-1} B_{\ell\ell} M_{W,i\ell} \left(x_\ell'\beta + f(z_\ell)\right) \right)^2 \sigma_i^2 + 2\sum_{i=1}^n \sum_{\ell \neq i} M_{W,ii}^{-2} B_{ii}^2 M_{W,i\ell}^2 \sigma_i^2 \sigma_\ell^2 \\
\leq \frac{1}{c^2} \max_i \sigma_i^2 \max_i \left(x_i'\beta + f(z_i)\right)^2 \sum_{i=1}^n B_{ii}^2 + \frac{2}{c} \max_i \sigma_i^4 \sum_{i=1}^n B_{ii}^2 = o(1),
\end{aligned}$$

because $\min_i M_{W,ii} \geq c > 0$, $\sum_{i=1}^n B_{ii}^2 \leq \text{trace}(\tilde{A}^2) = o(1)$, and $\max_i \left(x_i'\beta + f(z_i)\right)^2 \leq 2\max_i (x_i'\beta)^2 + 2\max_i f(z_i)^2 = \mathcal{O}(1)$.

Because we have that

$$\mathbb{E}[\hat{\theta} - \theta] \xrightarrow{p} 0, \quad \text{var}[\hat{\theta} - \theta] \xrightarrow{p} 0,$$

the proposed estimator $\hat{\theta}$ is consistent. $\square$

Proof of Lemma 4. We prove the result by considering a second-order approximation of $\hat{\theta}_{\text{JLA}} - \hat{\theta}$ around $\hat{a}_i := (1 - P_{W,ii})^{-1}(\hat{P}_{W,ii} - P_{W,ii})$ as

$$(\hat{\theta}_{\text{JLA}} - \hat{\theta})_2 := \sum_{i=1}^{n} \hat{\sigma}_i^2 \left(B_{ii} - \hat{B}_{ii} - \hat{B}_{ii}\hat{a}_i - \hat{B}_{ii} \left(\hat{a}_i^2 - \frac{1}{m} \frac{3P_{W,ii}^3 + P_{W,ii}^2}{1 - P_{W,ii}} \right) \right),$$

and an approximation error that we show to be negligible,

$$\text{AE}_2 := \sum_{i=1}^{n} \hat{\sigma}_i^2 \hat{B}_{ii} \left(\frac{1}{m} \frac{3\hat{P}_{W,ii}^3 + \hat{P}_{W,ii}^2 - (3P_{W,ii}^3 + P_{W,ii}^2)(1 - \hat{a}_i)^2}{(1 - \hat{a}_i)^2(1 - P_{W,ii})} - \frac{\hat{a}_i^3}{1 - \hat{a}_i} \right).$$

To decompose $\hat{\theta}_{\text{JLA}} - \hat{\theta} = (\hat{\theta}_{\text{JLA}} - \hat{\theta})_2 + \text{AE}_2$, note that

$$\begin{aligned}
\hat{\sigma}_{i,\text{JLA}}^2 &= \frac{y_i(y_i - w_i'\hat{\gamma})}{1 - \hat{P}_{W,ii}} \left(1 - \frac{1}{m} \frac{3\hat{P}_{W,ii}^3 + \hat{P}_{W,ii}^2}{1 - \hat{P}_{W,ii}} \right) \\
&= \frac{1 - P_{W,ii}}{1 - \hat{P}_{W,ii}} \hat{\sigma}_i^2 \left(1 - \frac{1}{m} \frac{3\hat{P}_{W,ii}^3 + \hat{P}_{W,ii}^2}{1 - \hat{P}_{W,ii}} \right),
\end{aligned}$$

so that

$$\begin{aligned}
\hat{\theta}_{\text{JLA}} - \hat{\theta} &= \sum_{i=1}^{n} B_{ii}\hat{\sigma}_i^2 - \hat{B}_{ii}\hat{\sigma}_{i,\text{JLA}}^2 \\
&= \sum_{i=1}^{n} B_{ii}\hat{\sigma}_i^2 - \hat{B}_{ii}\hat{\sigma}_i^2 \left(1 - \frac{1}{m} \frac{3\hat{P}_{W,ii}^3 + \hat{P}_{W,ii}^2}{1 - \hat{P}_{W,ii}} \right) \\
&= \sum_{i=1}^{n} B_{ii}\hat{\sigma}_i^2 + \hat{B}_{ii}\hat{\sigma}_i^2 \left(\frac{1}{m} \frac{1 - P_{W,ii}}{1 - \hat{P}_{W,ii}} \frac{3\hat{P}_{W,ii}^3 + \hat{P}_{ii}^2}{1 - \hat{P}_{W,ii}} - \frac{1 - P_{W,ii}}{1 - \hat{P}_{W,ii}} \right).
\end{aligned}$$

Add and subtract $m^{-1}\hat{\sigma}_i^2 \hat{B}_{ii}(1 - P_{W,ii})^{-1}(3P_{W,ii}^3 + P_{W,ii}^2)$ to obtain

$$\begin{aligned}
\hat{\theta}_{\text{JLA}} - \hat{\theta} = \sum_{i=1}^{n} & B_{ii}\hat{\sigma}_i^2 + \frac{1}{m}\hat{\sigma}_i^2 \hat{B}_{ii} \frac{3P_{W,ii}^3 + P_{W,ii}^2}{1 - P_{W,ii}} + \hat{B}_{ii}\hat{\sigma}_i^2 \left(\frac{1}{m} \frac{1 - P_{W,ii}}{1 - \hat{P}_{W,ii}} \frac{3\hat{P}_{W,ii}^3 + \hat{P}_{ii}^2}{1 - \hat{P}_{W,ii}} - \frac{1 - P_{W,ii}}{1 - \hat{P}_{W,ii}} \right) \\
& - \frac{1}{m}\hat{\sigma}_i^2 \hat{B}_{ii} \frac{3P_{W,ii}^3 + P_{W,ii}^2}{1 - P_{W,ii}} \\
&= \sum_{i=1}^{n} \hat{\sigma}_i^2 \left(B_{ii} + \frac{1}{m}\hat{\sigma}_i^2 \hat{B}_{ii} \frac{3P_{W,ii}^3 + P_{W,ii}^2}{1 - P_{W,ii}} \right) \\
& + \sum_{i=1}^{n} \hat{\sigma}_i^2 \hat{B}_{ii} \left(\frac{1}{m} \frac{(1 - P_{W,ii})^2(3\hat{P}_{W,ii}^3 + \hat{P}_{W,ii}^2) - (1 - \hat{P}_{W,ii})^2(3P_{W,ii}^3 + P_{W,ii}^2)}{(1 - \hat{P}_{W,ii})^2(1 - P_{W,ii})} - \frac{1 - P_{W,ii}}{1 - \hat{P}_{W,ii}} \right),
\end{aligned}$$

and using $(1-\hat{P}_{W,ii})^{-1}(1-P_{W,ii}) = (1-\hat{a}_i)^{-1}$, and expanding up to the third order as

$$\begin{aligned}\frac{1-P_{W,ii}}{1-\hat{P}_{W,ii}} &= 1 + \frac{1-P_{W,ii}}{1-\hat{P}_{W,ii}}\hat{a}_i \\ &= 1 + \hat{a}_i + \hat{a}_i^2 + \frac{1-P_{W,ii}}{1-\hat{P}_{W,ii}}\hat{a}_i^3,\end{aligned}$$

we have that

$$\begin{aligned}\hat{\theta}_{\text{JLA}} - \hat{\theta} &= \sum_{i=1}^n \hat{\sigma}_i^2 \left(B_{ii} + \frac{1}{m}\hat{B}_{ii}\frac{3P_{W,ii}^3 + P_{W,ii}^2}{1-P_{W,ii}} \right) \\ &\quad + \sum_{i=1}^n \hat{\sigma}_i^2 \hat{B}_{ii} \left(\frac{1}{m}\frac{3\hat{P}_{W,ii}^3 + \hat{P}_{W,ii}^2 - (1-\hat{a}_i)^2(3P_{W,ii}^3 + P_{W,ii}^2)}{(1-\hat{a}_i)^2(1-P_{W,ii})} - \left(1 + \hat{a}_i + \hat{a}_i^2 + \frac{\hat{a}_i^3}{1-\hat{a}_i}\right)\right) \\ &= (\hat{\theta}_{\text{JLA}} - \hat{\theta})_2 + \text{AE}_2.\end{aligned}$$

To describe the bias, we note that $\hat{P}_{W,ii}$, $\hat{B}_{ii}$, and $\hat{\sigma}_i^2$ are independent of each other, $\mathbb{E}[\hat{P}_{W,ii}] = P_{W,ii}$, $\mathbb{E}[\hat{B}_{ii}] = B_{ii}$, $\mathbb{E}[\hat{\sigma}_i^2] = \sigma_i^2 + \mathcal{O}_p(\sqrt{n}/k^{\alpha_f}\sqrt{p})$, and using properties of the Rademacher random variables,

$$\text{var}[\hat{a}_i] = \frac{2}{m}\frac{P_{W,ii} - \sum_{\ell=1}^n P_{W,i\ell}^4}{(1-P_{W,ii})^2} = \frac{1}{m}\frac{3P_{W,ii}^3 + P_{W,ii}^2}{1-P_{W,ii}} + \frac{P_{W,ii}(1-P_{W,ii})^2 - 2\sum_{\ell\neq i}^n P_{W,i\ell}^4}{m(1-P_{W,ii})^2}.$$

Therefore, in total we have:

$$\mathbb{E}[(\hat{\theta}_{\text{JLA}} - \hat{\theta})_2] = -\sum_{i=1}^n \sigma_i^2 B_{ii}\left(\text{var}[\hat{a}_i] - \frac{1}{m}\frac{3P_{W,ii}^3 + P_{W,ii}^2}{1-P_{W,ii}}\right) + \mathcal{O}_p(\sqrt{n}/k^{\alpha_f}\sqrt{p}),$$

or, assuming $k \to \infty$,

$$\mathbb{E}[(\hat{\theta}_{\text{JLA}} - \hat{\theta})_2] = \text{B}_m + o(1), \quad \text{B}_m := \sum_{i=1}^n B_{ii}\sigma_i^2\left(\frac{2\sum_{\ell\neq i} P_{W,i\ell}^4 - P_{W,ii}^2(1-P_{W,ii})^2}{m(1-P_{W,ii})^2}\right).$$

Focusing on the variance next, denote $y := (y_1, \dots, y_n)'$, so that

$$\begin{aligned}\text{var}\left[\sum_{i=1}^n \hat{\sigma}_i^2(B_{ii} - \hat{B}_{ii})\right] &= \mathbb{E}\left[\text{var}\left[\sum_{i=1}^n \hat{\sigma}_i^2\hat{B}_{ii}\right]\middle| y\right] + \text{var}\left[\mathbb{E}\left[\sum_{i=1}^n \hat{\sigma}_i^2\hat{B}_{ii}\right]\middle| y\right] = \mathbb{E}\left[\text{var}\left[\sum_{i=1}^n \hat{\sigma}_i^2\hat{B}_{ii}\right]\middle| y\right] \\ &\leq 2m^{-1}\sum_{i=1}^n\sum_{\ell=1}^n B_{i\ell}^2\mathbb{E}[\hat{\sigma}_i^2\hat{\sigma}_\ell^2] = \mathcal{O}\left(m^{-1}\text{trace}(\tilde{A}^2)\right),\end{aligned}$$

$$\begin{aligned}\text{var}\left[\sum_{i=1}^n \hat{\sigma}_i^2\hat{B}_{ii}\hat{a}_i\right] &= \mathbb{E}\left[\text{var}\left[\sum_{i=1}^n \hat{\sigma}_i^2\hat{B}_{ii}\hat{a}_i\middle| y, R_B\right]\right] \leq 2m^{-1}\sum_{i=1}^n\sum_{\ell=1}^{\ell} P_{W,i\ell}^2\frac{\mathbb{E}[\hat{B}_{ii}\hat{B}_{\ell\ell}]\mathbb{E}[\hat{\sigma}_i^2\hat{\sigma}_\ell^2]}{(1-P_{W,ii})(1-P_{W,\ell\ell})} \\ &= \mathcal{O}\left(m^{-1}\text{trace}(\tilde{A}^2) + m^{-2}\text{trace}(\tilde{A}_1^2)^{1/2}\text{trace}(\tilde{A}_2^2)^{1/2}\right)\end{aligned}$$

for $\tilde{A}_k := S_{xx}^{-1/2} A_k' A_k S_{xx}^{-1/2}$ for $k = 1, 2$. Regarding the ensuing terms, it holds that:

$$\begin{aligned}
\operatorname{var}\left[\sum_{i=1}^{n} \hat{\sigma}_i^2 \hat{B}_{ii}\left(\hat{a}_i^2 - \operatorname{var}[\hat{a}_i]\right)\right] &= \sum_{i=1}^{n}\sum_{\ell=1}^{n} \mathbb{E}[\hat{B}_{ii}\hat{B}_{\ell\ell}]\mathbb{E}[\hat{\sigma}_i^2\hat{\sigma}_\ell^2] \operatorname{cov}[\hat{a}_i^2, \hat{a}_\ell^2] \\
&= \mathcal{O}\left(m^{-2}\operatorname{trace}(\tilde{A}^2) + m^{-3}\operatorname{trace}(\tilde{A}_1^2)^{1/2}\operatorname{trace}(\tilde{A}_2^2)^{1/2}\right),
\end{aligned}$$

$$\begin{aligned}
\operatorname{var}\left[\sum_{i=1}^{n} \hat{\sigma}_i^2 (\hat{B}_{ii} - B_{ii}) \frac{2\sum_{\ell\neq i}^{n} P_{\mathrm{W},i\ell}^4 - P_{\mathrm{W},ii}(1-P_{\mathrm{W},ii})^2}{m(1-P_{\mathrm{W},ii})^2}\right] &= \mathcal{O}\left(m^{-3}\operatorname{trace}(\tilde{A}^2)\right), \\
\operatorname{var}\left[\sum_{i=1}^{n} B_{ii}\left(\hat{\sigma}_i^2 - \sigma_i^2\right) \frac{2\sum_{\ell\neq i}^{n} P_{\mathrm{W},i\ell}^4 - P_{\mathrm{W},ii}(1-P_{\mathrm{W},ii})^2}{m(1-P_{\mathrm{W},ii})^2}\right] &= \mathcal{O}\left(m^{-2}\operatorname{var}[\hat{\theta}]\right).
\end{aligned}$$

Because $\operatorname{trace}(\tilde{A}^2) = \mathcal{O}(\operatorname{var}[\hat{\theta}])$ and $m^{-4}\operatorname{var}[\hat{\theta}]^{-2}\operatorname{var}[\hat{\theta}_1]\operatorname{var}[\hat{\theta}_2] = o(1)$, it can be ultimately established that $\operatorname{var}[\hat{\theta}]^{-1/2}\left((\hat{\theta}_{\mathrm{JLA}} - \hat{\theta})_2 - \mathrm{B}_m\right) = o_p(1)$.

Using that $\mathbb{E}[\hat{a}_i^3] = \mathcal{O}(m^{-2})$, $\mathbb{E}[\hat{a}_i^4] = \mathcal{O}(m^{-2})$, and $\max_i |\hat{a}_i| = o_p(\log n/\sqrt{m})$, the terms in the approximation error are as follows:

$$\begin{aligned}
\sum_{i=1}^{n} \hat{\sigma}_i^2 \hat{B}_{ii}\hat{a}_i^3 + \sum_{i=1}^{n} \hat{\sigma}_i^2 \hat{B}_{ii}\hat{a}_i^4 &= m^{-2}\mathcal{O}_p\left(\mathbb{E}[\hat{\theta}_{1,\mathrm{PI}} - \theta_1] + \mathbb{E}[\hat{\theta}_{2,\mathrm{PI}} - \theta_2]\right), \\
\sum_{i=1}^{n} \hat{\sigma}_i^2 \hat{B}_{ii}\frac{\hat{a}_i^5}{1-\hat{a}_i} &= \frac{\log n}{m^{5/4}}\mathcal{O}_p\left(\mathbb{E}[\hat{\theta}_{1,\mathrm{PI}} - \theta_1] + \mathbb{E}[\hat{\theta}_{2,\mathrm{PI}} - \theta_2]\right), \\
\frac{1}{m}\sum_{i=1}^{n} \hat{\sigma}_i^2 \hat{B}_{ii} & \frac{3\hat{P}_{\mathrm{W},ii}^3 + \hat{P}_{\mathrm{W},ii}^2 - (3P_{\mathrm{W},ii}^3 + P_{\mathrm{W},ii}^2)(1-\hat{a}_i)^2}{(1-\hat{a}_i)^2(1-P_{\mathrm{W},ii})} \\
&= \left(m^{-2} + \frac{\log n}{p^{5/4}}\right)\mathcal{O}_p\left(\mathbb{E}[\hat{\theta}_{1,\mathrm{PI}} - \theta_1] + \mathbb{E}[\hat{\theta}_{2,\mathrm{PI}} - \theta_2]\right).
\end{aligned}$$

□

Proof of Theorem 5. Representation in the theorem holds because

$$\sum_{\ell=1}^{r} \lambda_\ell \hat{b}_\ell^2 = \hat{\beta}' S_{xx}^{1/2} Q D Q' S_{xx}^{1/2}\hat{\beta} = \hat{\beta}' S_{xx}^{1/2} S_{xx}^{-1/2} A S_{xx}^{-1/2} S_{xx}^{1/2}\hat{\beta} = \hat{\beta}' A \hat{\beta},$$

and

$$\sum_{i=1}^{n} B_{ii}\hat{\sigma}_i^2 = \operatorname{trace}(A\widehat{\operatorname{var}}[\hat{\beta}]) = \operatorname{trace}(D\widehat{\operatorname{var}}[\hat{b}]) = \sum_{\ell=1}^{r} \lambda_\ell \widehat{\operatorname{var}}[\hat{b}_\ell].$$

We prove the theorem in three steps.
Approximation. Equivalently, represent $\hat{\theta}$ as

$$\hat{\theta} = \sum_{\ell=1}^{r} \lambda_\ell\left(\hat{b}_\ell^2 - \operatorname{var}[\hat{b}_\ell]\right) + \sum_{i=1}^{n} B_{ii}(\sigma_i^2 - \hat{\sigma}_i^2),$$

and below we show that the second term is asymptotically dominated by the variance of the estimator, $\operatorname{var}[\hat{\theta}]$. Given that, the asymptotic distribution of $\hat{\theta}$ is then driven by the joint distribution of random vector $\hat{b}$.

We need to show that the second term is asymptotically mean-zero and is of smaller order than variance of $\hat{\theta}$. The first claim is immediate because we have shown in Lemma 3 that

$$\mathbb{E}[\hat{\sigma}_i^2 - \sigma_i^2] = \mathcal{O}_p(\sqrt{n}/k^{\alpha_f}\sqrt{p}) \xrightarrow{p} 0,$$

so that

$$\mathbb{E}\left[\sum_{i=1}^n B_{ii}(\sigma_i^2 - \hat{\sigma}_i^2)\right] \xrightarrow{p} 0.$$

Now, rewrite

$$\begin{aligned}\sum_{i=1}^n B_{ii}(\hat{\sigma}_i^2 - \sigma_i^2) &= \sum_{i=1}^n B_{ii} M_{W,ii}^{-1} x_i'\beta \sum_{\ell=1}^n M_{W,i\ell} e_\ell + \sum_{i=1}^n (e_i^2 - \sigma_i^2) \\ &+ \sum_{i=1}^n B_{ii} M_{W,ii}^{-1} f(z_i) \sum_{\ell=1}^n M_{W,i\ell} e_\ell + \sum_{i=1}^n B_{ii} M_{W,ii}^{-1} \sum_{\ell \neq i} M_{W,i\ell} e_i e_\ell.\end{aligned}$$

The variances of the first and the third terms are

$$\sum_{\ell=1}^n \sigma_\ell^2 \left(\sum_{i=1}^n M_{W,i\ell} B_{ii} M_{W,ii}^{-1} x_i'\beta\right)^2 \leq \max_i \sigma_i^2 \sum_{i=1}^n B_{ii}^2 M_{W,ii}^{-2} (x_i\beta)^2 \leq \max_i \sigma_i^2 \max_i (x_i'\beta)^2 M_{W,ii}^{-2} \sum_{i=1}^n B_{ii}^2,$$

$$\sum_{\ell=1}^n \sigma_\ell^2 \left(\sum_{i=1}^n M_{W,i\ell} B_{ii} M_{W,ii}^{-1} f(z_i)\right)^2 \leq \max_i \sigma_i^2 \sum_{i=1}^n B_{ii}^2 M_{W,ii}^{-2} f(z_i)^2 \leq \max_i \sigma_i^2 \max_i f(z_i)^2 M_{W,ii}^{-2} \sum_{i=1}^n B_{ii}^2,$$

and of the second and the forth

$$\sum_{i=1}^n B_{ii}^2 \operatorname{var}[e_i^2] \leq \max_i \mathbb{E}[e_i^4] \sum_{i=1}^n B_{ii}^2,$$

$$\sum_{i=1}^n \sum_{\ell \neq i} (B_{ii}^2 M_{W,ii}^{-2} + B_{ii} M_{W,ii}^{-1} B_{\ell\ell} M_{W,\ell\ell}^{-1}) M_{i\ell}^2 \sigma_i^2 \sigma_\ell^2 \leq 2 \max_i \sigma_i^4 M_{W,ii}^{-2} \sum_{i=1}^n B_{ii}^2.$$

Because each variance is bounded by $C\sum_{i=1}^n B_{ii}^2$, to show that it is of smaller order than the variance of $\hat{\theta}$, we need $\operatorname{var}[\hat{\theta}]^{-1}\sum_{i=1}^n B_{ii}^2 = o(1)$. It holds because

$$\operatorname{var}[\hat{\theta}]^{-1} \sum_{i=1}^n B_{ii}^2 \leq \max_i v_i'v_i \operatorname{var}[\hat{\theta}]^{-1} \sum_{\ell=1}^r \lambda_\ell^2 \leq \max_i v_i'v_i \max_i \sigma_i^{-4} = o(1).$$

Variance estimator consistency. Now we show that the variance estimator is consistent, i.e. $\operatorname{var}[\hat{b}]^{-1}\widehat{\operatorname{var}}[\hat{b}] \xrightarrow{p} I_r$. For it we need to show that

$$\operatorname{var}[\vartheta'\hat{b}]^{-1}\left(\widehat{\operatorname{var}}[\vartheta'\hat{b}] - \operatorname{var}[\vartheta'\hat{b}]\right) = o_p(1), \quad \vartheta \in \mathbb{R}^r, \quad \vartheta'\vartheta = 1$$

for some non-random ϑ. Rewrite the expression above as

$$\delta(\vartheta) := \sum_{i=1}^n v_i(\vartheta)(\hat{\sigma}_i^2 - \sigma_i^2), \tag{9}$$

where

$$v_i(\vartheta) := \frac{(\vartheta' v_i)^2}{\sum_{i=1}^n \sigma_i^2 (\vartheta' v_i)^2}.$$

We know that $\mathbb{E}[\delta(\vartheta)] = o_p(1)$ because by the triangle inequality, the Cauchy-Schwartz inequality, and $|\mathbb{E}[\hat{\sigma}_i^2 - \sigma_i^2]| = o_p(1)$,

$$|v(\vartheta)\mathbb{E}[\hat{\sigma}_i^2 - \sigma_i^2]| \leq (v(\vartheta)^2)^{1/2} \cdot \left(|\mathbb{E}[\hat{\sigma}_i^2 - \sigma_i^2]|^2\right)^{1/2} = o_p(1)$$

for $i = 1, \ldots, n$. The variance of $\delta(\vartheta)$ is

$$\begin{aligned}
\sum_{i=1}^n \delta(\vartheta)^2 \operatorname{var}[\hat{\sigma}_i^2] &\leq \sum_{i=1}^n v_i(\vartheta)^4 \\
&\leq \max_i \sigma_i^{-4} \max_i v_i v_i' \frac{\vartheta'\vartheta}{\sum_{i=1}^n v_i v_i' \vartheta' \vartheta} \\
&= \max_i \sigma_i^{-4} \max_i v_i v_i' = o(1)
\end{aligned}$$

because $\max_i v_i v_i' = o(1)$ by assumption.

Asymptotic normality. Our objective is to prove that

$$\operatorname{var}[\vartheta'\hat{b}]^{-1/2} \left(\vartheta'\hat{b} - \vartheta' b\right) \xrightarrow{d} \mathcal{N}(0,1),$$

where $\hat{b} := Q' S_{xx}^{1/2} \hat{\beta}$, and $b := Q' S_{xx}^{1/2} \beta$. Lyapunov's condition implies that it is sufficient to show that

$$\operatorname{var}[\vartheta'\hat{b}]^{-2} \sum_{i=1}^n \mathbb{E}\left[\left(\vartheta'(\hat{b} - b)\right)^4\right] = o_p(1).$$

Because we have that

$$\begin{aligned}
\vartheta'(\hat{b} - b) &= \vartheta'(Q' S_{xx}^{1/2} \hat{\beta} - Q' S_{xx}^{1/2} \beta) \\
&= \vartheta' \left(Q' S_{xx}^{1/2} \sum_{j=1}^n M_{ij} x_j f(z_i) + Q' S_{xx}^{1/2} \sum_{j=1}^n M_{ij} x_j e_i \right) \\
&= \vartheta'(v_i f(z_i) + v_i e_i),
\end{aligned}$$

and $(a+b)^4 \leq C(a^4 + b^4)$ for some constant C, Lyapunov's condition is equivalent to

$$\operatorname{var}[\vartheta'\hat{b}]^{-2} \sum_{i=1}^n C \left(f(z_i)^4 + \mathbb{E}[e_i^4]\right) \cdot (\vartheta' v_i)^4 = o_p(1).$$

It holds because $\max_i |f(z_i)| = \mathcal{O}(1)$, and $\max_i \mathbb{E}[e_i^4] = \mathcal{O}(1)$ by assumption, so that $\max_i f(z_i)^4 + \mathbb{E}[e_i^4] = \mathcal{O}(1)$, also $\max_i (\vartheta' v_i)^2 \leq \max_i v_i' v_i = o(1)$, $\sum_{i=1}^n (\vartheta' v_i)^2 = \sum_{i=1}^n \vartheta' v_i v_i' \vartheta = 1$, and $\operatorname{var}[\vartheta'\hat{b}]^{-2} \leq \max_i \sigma_i^{-2} = \mathcal{O}(1)$. □

Proof of Theorem 6. We derive the limiting distribution of $\hat{\theta}$ with growing rank based on the following result regarding the joint normality of independent and not necessarily identical random variables as in Kline, Saggio, and Sølvsten (2020).

Let $\{q_{n,i}\}_{i,n}$ be a triangular array of row-wise independent random variables with $\mathbb{E}[q_{n,i}] = 0$ and $\text{var}[q_{n,i}] = \sigma^2_{n,i}$, let $\{\dot{w}_{n,i}\}_{i,n}$ be a triangular array of non-random weights that satisfy $\sum_{i=1}^n \dot{w}_{n,i}\sigma^2_{n,i} = 1$ for $\forall n$, and let $(Q_n)_n$ be a sequence of symmetric non-random matrices in $\mathbb{R}^{n\times n}$ with zeros on the diagonal and having $2\sum_{i=1}^n \sum_{\ell\neq i} Q^2_{n,i\ell}\sigma^2_{n,i}\sigma^2_{n,\ell} = 1$. Define

$$\mathcal{S}_n := \sum_{i=1}^n \dot{w}_{n,i}q_{n,i}, \quad \mathcal{U}_n := \sum_{i=1}^n \sum_{\ell\neq i} Q_{n,i\ell}q_{n,i}q_{n,\ell}.$$

Lemma 7. *If* $\max_i \mathbb{E}[q^4_{n,i}] + \sigma^{-2}_{n,i} = O(1)$, *(i)* $\max_i \dot{w}^2_{n,i} = o(1)$, *and (ii) trace*$(Q^4_n) = o(1)$, *then* $(\mathcal{S}_n, \mathcal{U}_n)' \xrightarrow{d} \mathcal{N}(0, I_2)$.

Proof. See Appendix B in Kline, Saggio, and Sølvsten (2020) and Appendix A2 in Solvsten (2020). □

The U-statistic representation holds because

$$\begin{aligned}
\hat{\theta} &= \sum_{i=1}^n y_i \tilde{w}_i' \hat{\gamma}_{-i} = \sum_{i=1}^n y_i \tilde{w}_i' \left(S^{-1}_{ww} - w_i w_i'\right) \sum_{\ell\neq i} w_\ell y_\ell \\
&= \sum_{i=1}^n y_i \tilde{w}_i' \left(S^{-1}_{ww} + \frac{S^{-1}_{ww} w_i w_i' S^{-1}_{ww}}{1 - w_i' S^{-1}_{ww} w_i}\right) \sum_{\ell\neq i} w_\ell y_\ell \\
&= \sum_{i=1}^n y_i \tilde{w}_i' \left(S^{-1}_{ww} + M^{-1}_{W,ii} S^{-1}_{ww} w_i w_i' S^{-1}_{ww}\right) \sum_{\ell\neq i} w_\ell y_\ell \\
&= \sum_{i=1}^n y_i w_i' S^{-1}_{ww} \check{A} S^{-1}_{ww} \sum_{\ell\neq i} w_\ell y_\ell + \sum_{i=1}^n y_i w_i' S^{-1}_{ww} \check{A} M^{-1}_{W,ii} S^{-1}_{ww} w_i w_i' S^{-1}_{ww} \sum_{\ell\neq i} w_\ell y_\ell \\
&= \sum_{i=1}^n \sum_{\ell\neq i} y_i y_\ell w_i' S^{-1}_{ww} \check{A} S^{-1}_{ww} w_\ell + \sum_{i=1}^n \sum_{\ell\neq i} y_i y_\ell M^{-1}_{W,ii} w_i' S^{-1}_{ww} \check{A} S^{-1}_{ww} w_i w_i' S^{-1}_{ww} w_\ell \\
&= \sum_{i=1}^n \sum_{\ell\neq i} y_i y_\ell B_{W,i\ell} + y_i y_\ell M^{-1}_{W,ii} B_{W,ii}(1 - M_{W,i\ell}) = \sum_{i=1}^n \sum_{\ell\neq i} C_{i\ell} y_i y_\ell.
\end{aligned}$$

Now, define $\check{e}_i := f(z_i) - p_k(z_i)'\alpha$ to be an approximation error (it is implicitly indexed by the unknown function but we omit this dependence for brevity). We can write the difference

$$\begin{aligned}
\hat{\theta} - \theta &= \sum_{i=1}^n \sum_{\ell\neq i} C_{i\ell} y_i y_\ell - \sum_{i=1}^n \gamma' w_i \tilde{w}_i' \gamma \\
&= \sum_{i=1}^n \sum_{\ell\neq i} C_{i\ell} \left(x_i'\beta + f(z_i) + e_i\right) \left(x_\ell'\beta + f(z_\ell) + e_\ell\right) - \sum_{i=1}^n \gamma' w_i \tilde{w}_i' \gamma \\
&= \sum_{i=1}^n \sum_{\ell\neq i} C_{i\ell} \left(\gamma' w_i + e_i + \check{e}_i\right) \left(\gamma' w_\ell + e_\ell + \check{e}_\ell\right) - \sum_{i=1}^n \gamma' w_i \tilde{w}_i' \gamma \\
&= \sum_{i=1}^n \sum_{\ell\neq i} C_{i\ell} \gamma' w_i \gamma' w_\ell + \sum_{i=1}^n \sum_{\ell\neq i} C_{i\ell}(\gamma' w_i e_\ell + \gamma' w_\ell e_i) - \sum_{i=1}^n \gamma' w_i \tilde{w}_i' \gamma \\
&\quad + \sum_{i=1}^n \sum_{\ell\neq i} C_{i\ell}(\gamma' w_i \check{e}_\ell + \gamma' w_\ell \check{e}_i + e_i \check{e}_\ell + e_\ell \check{e}_i) + \sum_{i=1}^n \sum_{\ell\neq i} C_{i\ell} e_i e_\ell + \sum_{i=1}^n \sum_{\ell\neq i} C_{i\ell} \check{e}_i \check{e}_\ell
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} e_i \sum_{\ell \neq i} (\gamma' w_i + \gamma' w_\ell) C_{i\ell} + \sum_{i=1}^{n} \check{e}_i \sum_{\ell \neq i} (\gamma' w_i + \gamma' w_\ell + e_i + e_\ell) \\
&\quad + \sum_{i=1}^{n} \sum_{\ell \neq i} C_{i\ell} e_i e_\ell + \sum_{i=1}^{n} \sum_{\ell \neq i} C_{i\ell} \check{e}_i \check{e}_\ell \\
&= \sum_{i=1}^{n} (2\tilde{w}_i'\gamma - \check{w}_i'\gamma) e_i + \sum_{i=1}^{n} \sum_{\ell \neq i} C_{i\ell} e_i e_\ell + \mathcal{O}_p(nk^{-\alpha_f}),
\end{aligned}
$$

where the last equality follows from defining $\check{w}_i := \sum_{\ell=1}^{n} M_{W,i\ell} \frac{B_{W,\ell\ell}}{1-P_{W,\ell\ell}} w_\ell$, and the fact that by the Assumption 1

$$
\check{e}_i^2 \leq \big(f(z_i) - p_k(z_i)'\alpha\big)^2 = \mathbb{E}[\big(f(z_i) - p_k(z_i)'\alpha\big)^2] \leq \min_{\alpha \in \mathbb{R}^k} \mathbb{E}[\big(f(z_i) - p_k(z_i)'\alpha\big)^2] \leq Ck^{-2\alpha_f}.
$$

As $k \to \infty$, we have that

$$
\hat{\theta} - \theta = \sum_{i=1}^{n} \left(2\tilde{w}_i'\gamma - \check{w}_i'\gamma\right) e_i + \sum_{i=1}^{n} \sum_{\ell \neq i} C_{i\ell} e_i e_\ell + o_p(1).
$$

Having dispensed with asymptotically negligible contributions to $\hat{\theta}$, asymptotic variance is

$$
\operatorname{var}[\hat{\theta}] = \sum_{i=1}^{n} (2\tilde{w}_i'\gamma - \check{w}_i'\gamma)^2 \sigma_i^2 + 2 \sum_{i=1}^{n} \sum_{\ell \neq i} C_{i\ell} \sigma_i^2 \sigma_\ell^2,
$$

that is, a sum of two components given by

$$
\mathcal{V}_s := \sum_{i=1}^{n} (2\tilde{w}_i'\gamma - \check{w}_i'\gamma)^2 \sigma_i^2, \quad \mathcal{V}_u := 2 \sum_{i=1}^{n} \sum_{\ell \neq i} C_{i\ell} \sigma_i^2 \sigma_\ell^2.
$$

The normalized difference is then given by

$$
\operatorname{var}[\hat{\theta}]^{-1/2} (\hat{\theta} - \theta) = \omega_1 \mathcal{S}_n + \omega_2 \mathcal{U}_n,
$$

with $\omega_1 := \operatorname{var}[\hat{\theta}]^{-1/2} \mathcal{V}_s^{1/2}$, $\omega_2 := \operatorname{var}[\hat{\theta}]^{-1/2} \mathcal{V}_u^{1/2}$, and

$$
\mathcal{S}_n := \mathcal{V}_s^{-1/2} \sum_{i=1}^{n} (2\tilde{w}_i'\gamma - \check{w}_i'\gamma) e_i, \quad \mathcal{U}_n := \mathcal{V}_u^{-1/2} \sum_{i=1}^{n} \sum_{\ell \neq i} C_{i\ell} e_i e_\ell.
$$

Consider the case where the limit of ω_1 is nonzero. If it is not, then asymptotic normality of the difference $\operatorname{var}[\hat{\theta}]^{-1/2}(\hat{\theta} - \theta)$ is implied by asymptotic normality of $\mathcal{U}_n$. Using notation of Lemma 7, we have that $\dot{w}_i = \mathcal{V}_s^{-1/2}(2\tilde{w}_i'\gamma - \check{w}_i'\gamma)$, and $Q_{i\ell} = \mathcal{V}_u^{-1/2} C_{i\ell}$.

To verify the condition *(i)* of Lemma 7, note that

$$
\begin{aligned}
\max_i \dot{w}_i^2 = \max_i \mathcal{V}_s (2\tilde{w}_i'\gamma - \check{w}_i'\gamma)^2 &\leq \max_i 4\mathcal{V}_s \left((\tilde{w}_i'\gamma)^2 + (\check{w}_i'\gamma)^2\right) \\
&= \max_i 4\omega_1^{-2} \frac{(\tilde{w}_i'\gamma)^2 + (\check{w}_i'\gamma)^2}{\operatorname{var}[\hat{\theta}]} = o(1),
\end{aligned}
$$

where the last equality follows from Theorem 6 *(i)*, and the nonzero limit of ω_1.

For the condition *(ii)* of Lemma 7, denote $\tilde{A}_W := S_{ww}^{-1/2}\check{A}S_{ww}^{-1/2}$, and note that the first r eigenvalues of the $\tilde{A}_W$ matrix are equal to eigenvalues of the $\tilde{A}$ matrix. Now, with constants c_U and c_L not dependent on n, we have that $\text{trace}(C^4) \leq c_U \cdot \text{trace}(B_W^4) = c_U \cdot \text{trace}(\tilde{A}_W^4) \leq c_U \lambda_1^2 \cdot \text{trace}(\tilde{A}_W^2)$, and $\mathcal{V}_u \geq c_L \min_i \sigma_i^4 \cdot \text{trace}(\tilde{A}_W)$, which implies

$$\text{trace}(Q^4) \leq \frac{c_U \lambda_1^2 \cdot \text{trace}(\tilde{A}_W^2)}{\left(c_L \min_i \sigma_i^4 \cdot \text{trace}(\tilde{A}_W^2)\right)^2} = \mathcal{O}\left(\frac{\lambda_1^2}{\text{trace}(\tilde{A}_W^2)}\right) = o(1),$$

where the last equality follows from Theorem 6 *(ii)*. □

Sherman-Morrison representation. The representation in (4) holds due to application of the Sherman-Morrison formula to the leave-one-out estimator of γ. Denoting $S_{ww} := \sum_{i=1}^n w_i w_i'$, it holds that

$$\left(S_{ww} - w_i w_i'\right)^{-1} = S_{ww}^{-1} + \frac{S_{ww}^{-1} w_i w_i' S_{ww}^{-1}}{1 - w_i' S_{ww}^{-1} w_i}.$$

Plugging the above into the definition of $\hat{\gamma}_{-i}$, we obtain

$$\begin{aligned}
\hat{\gamma}_{-i} &= \left(S_{ww}^{-1} + \frac{S_{ww}^{-1} w_i w_i' S_{ww}^{-1}}{1 - w_i' S_{ww}^{-1} w_i}\right)\left(\sum_{i=1}^n w_i y_i - w_i y_i\right) \\
&= S_{ww}^{-1} \sum_{i=1}^n w_i y_i - S_{ww} w_i y_i + \frac{S_{ww}^{-1} w_i w_i' S_{ww}^{-1} \sum_{i=1}^n w_i y_i}{1 - w_i' S_{ww}^{-1} w_i} - \frac{S_{ww}^{-1} w_i w_i' S_{ww}^{-1} w_i y_i}{1 - w_i' S_{ww}^{-1} w_i} \\
&= \hat{\gamma} + S_{ww}^{-1} w_i \left(\frac{w_i' \hat{\gamma}}{1 - w_i' S_{ww}^{-1} w_i} - y_i - \frac{w_i' S_{ww}^{-1} w_i y_i}{1 - w_i' S_{ww}^{-1} w_i}\right) \\
&= \hat{\gamma} + S_{ww}^{-1} w_i \left(\frac{w_i' \hat{\gamma} - y_i}{1 - w_i' S_{ww}^{-1} w_i}\right) \\
&= \hat{\gamma} - S_{ww}^{-1} w_i \frac{(y_i - w_i' \hat{\gamma})}{1 - w_i' S_{ww}^{-1} w_i}.
\end{aligned}$$

Substituting for $\hat{\gamma}_{-i}$ in (3), we have

$$\begin{aligned}
\hat{\sigma}_i^2 &= y_i \left(y_i - w_i' \hat{\gamma}_{-i}\right) \\
&= y_i \left(y_i - w_i' \left(\hat{\gamma} - S_{ww}^{-1} w_i \frac{(y_i - w_i' \hat{\gamma})}{1 - w_i' S_{ww}^{-1} w_i}\right)\right) \\
&= y_i \left(y_i - w_i' \hat{\gamma} + w_i' S_{ww}^{-1} w_i \frac{(y_i - w_i' \hat{\gamma})}{1 - w_i' S_{ww}^{-1} w_i}\right) \\
&= y_i \left((y_i - w_i' \hat{\gamma}) \left(\frac{1}{1 - w_i' S_{ww}^{-1} w_i}\right)\right) \\
&= \frac{y_i \hat{e}_i}{M_{W,ii}}.
\end{aligned}$$

B. Additional Tables

Table 4: Descriptive statistics of the cleaned worker–firm panel

	Mean	SD	p25	Median	p75
A. Worker-year panel					
Age	39.20	10.52	31.00	38.00	47.00
Months worked	8.19	1.69	8.00	9.00	9.00
Tenure (months)	103.57	103.76	22.00	69.00	153.00
Annual wage (k€)	13.92	10.46	7.99	10.53	16.02
Hourly wage (€)	7.17	5.17	4.15	5.33	7.99
Number of jobs	1.01	0.12	1.00	1.00	1.00
B. Worker composition and mobility					
Job-to-job moves (count)	1,803,534	–	–	–	–
Job-to-job movers (%)	9.85%	–	–	–	–
Education: At most primary (%)	47.91%	–	–	–	–
Education: Secondary (%)	34.25%	–	–	–	–
Education: At least bachelor's (%)	17.52%	–	–	–	–
Education missing (%)	0.32%	–	–	–	–
Qualification: Specialized workers (%)	60.44%	–	–	–	–
Qualification: Generic workers (%)	15.27%	–	–	–	–
Qualification missing (%)	11.88%	–	–	–	–
Qualification: Top managers (%)	7.10%	–	–	–	–
Qualification: Middle managers (%)	5.31%	–	–	–	–
Lisboa (%)	35.13%	–	–	–	–
Norte (%)	35.06%	–	–	–	–
Centro (%)	18.28%	–	–	–	–
Alentejo (%)	4.54%	–	–	–	–
Algarve (%)	3.67%	–	–	–	–
Region (NUT2) other (%)	3.32%	–	–	–	–
C. Firm-year panel					
Revenue (M€)	1.64	31.48	0.08	0.20	0.56
Assets (M€)	2.34	73.33	0.07	0.20	0.61
Workers	11.26	110.04	2.00	4.00	7.00
D. Panel counts					
Unique workers	3,532,497	–	–	–	–
Unique firms	484,704	–	–	–	–
Worker-firm observations	18,314,740	–	–	–	–
E. Estimation variable coverage (all runs)					
Missing: Age (%)	0.00%	–	–	–	–
Missing: Assets (%)	8.26%	–	–	–	–
Missing: Education (%)	0.32%	–	–	–	–
Missing: Firm ID (%)	0.00%	–	–	–	–
Missing: Log hourly wage (%)	0.01%	–	–	–	–
Missing: Qualification (%)	11.88%	–	–	–	–
Missing: Revenue (%)	8.26%	–	–	–	–
Missing: Gender (%)	0.00%	–	–	–	–
Missing: Worker ID (%)	0.00%	–	–	–	–
Missing: Year (%)	0.00%	–	–	–	–

Note: Descriptive moments are computed on the cleaned analysis sample (2008–2018). Panels A and C report distributional statistics for worker-year and firm-year variables, respectively. Panels B and D report composition, mobility, and sample-size counts. Percentages are sample shares in the corresponding panel.

Table 5: Coverage of the largest connected component

	Full sample	Largest CC	Share (%)
Observations	18,312,694	17,256,654	94.23
Workers	3,532,350	3,282,116	92.92
Firms	484,696	333,180	68.74
Mean obs per worker	5.18	5.26	–
Mean spells per worker	1.51	1.54	–
B. Coverage for variables used across all runs			
Non-missing: Age	18,312,694	17,256,654	94.23
Non-missing: Assets	16,799,218	16,061,565	95.61
Non-missing: Education	18,253,325	17,200,972	94.23
Non-missing: Firm ID	18,312,694	17,256,654	94.23
Non-missing: Log hourly wage	18,312,694	17,256,654	94.23
Non-missing: Qualification	16,137,639	15,173,011	94.02
Non-missing: Revenue	16,799,218	16,061,565	95.61
Non-missing: Gender	18,312,694	17,256,654	94.23
Non-missing: Worker ID	18,312,694	17,256,654	94.23
Non-missing: Year	18,312,694	17,256,654	94.23

Note: The table compares the full cleaned estimation sample with the largest connected component used for fixed-effects estimation. "Share (%)" is the fraction retained in the largest connected component relative to the full sample for each count measure.