

Generalized AKM: Flexible Controls and Interactions in Wage Decompositions*

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

July 2026

Abstract

How much wage dispersion is attributed to workers, firms, and their sorting depends on how wages are adjusted for observed characteristics. Standard AKM decompositions impose a known linear adjustment. We develop Generalized AKM, a framework that permits an unknown smooth covariate function and group-specific nonlinear interactions while preserving the original variance components. We prove consistency and asymptotic normality with heteroskedastic errors and many fixed effects, and characterize the stronger smoothness required for quadratic forms. In Portuguese employer–employee data, adding worker and firm-input controls lowers the bias-corrected worker-effect variance from 0.551 to 0.474 of total wage variance, firm-effect variance from 0.144 to 0.121, and sorting from 0.080 to 0.047. Across three group-specific nonlinear bases, firm-effect variance remains between 0.114 and 0.117 and sorting between 0.041 and 0.042, while worker-effect variance ranges from 0.474 to 0.491. Which controls enter matters more for firm variance and sorting than how flexibly they enter; worker variance remains more sensitive to the basis.

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

JEL codes: J31, J62, C14, C23, C55

*We are grateful to Edoardo M. Acabbi, Stanislav Anatolyev, Andrea Alati, Matias D. Cattaneo, Patrick Kline, 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

How much wage dispersion is attributed to workers, firms, and the sorting of workers across firms? AKM decompositions answer this question by assigning wage variation to worker effects, firm effects, and their covariance after adjusting wages for observed characteristics (Abowd, Kramarz, and Margolis 1999). These components have become standard measures of worker heterogeneity, firm pay differences, and sorting in matched employer–employee data. So how well wages are adjusted for observables, before the decomposition is even run, shapes how these components should be read. Standard AKM estimators, including the leave-out correction of Kline, Saggio, and Sølvssten (2020), impose a known linear control function. Yet returns to experience and schooling are nonlinear (Mincer 1974; Card et al. 2018), and firm inputs need not be additively separable from worker characteristics. If the control function is misspecified, its omitted nonlinearities and interactions can be reassigned to worker effects, firm effects, or sorting. At stake is whether the decomposition isolates persistent worker and firm heterogeneity or partly relabels observed wage schedules as fixed effects.

We develop Generalized AKM, a semiparametric estimator for the original AKM variance components when observed covariates enter through an unknown smooth function. The estimator approximates this function with a growing polynomial or spline basis and permits the resulting nonlinear profiles to vary across worker groups. Worker and firm fixed effects, year effects, and all group-specific series terms enter one augmented regressor set, rather than adjusting wages for covariates in a first step and decomposing the residual.¹ The construction preserves the familiar worker, firm, and sorting components rather than replacing them with latent types or a different earnings model. It answers two related questions: how much does this added flexibility change the decomposition in practice, and when can these components be estimated at all without treating the covariate function as known?

The empirical application answers the first question. We use Portuguese linked employer–employee data for 2008–2018 and estimate a five-step ladder of increasingly flexible specifications. The common estimation sample is a leave-match-out set (the largest connected set of workers and firms that supports the match-level correction) of 9.4 million person-year observations, 1.7 million workers, and 105,000 firms. Adding the observed controls—quadratic and cubic age terms, education, qualification, and firm inputs—in a linear additive specification lowers worker-effect variance from 0.551 to 0.474 of total wage variance, firm-effect variance from 0.144 to 0.121, and sorting from 0.080 to 0.047.² The proportional movement is largest for sorting, which falls by more than two fifths: a substantial part of what the uncontrolled decomposition attributes to high-wage workers matching with high-wage firms is instead accounted for by education, qualification, and

1. The two-step alternative introduces a higher-order bias, because the first-step estimation error enters both the adjusted outcome and the residual used in the bias correction (Kline 2024).

2. These are bias-corrected point estimates. The correction used in the application treats each worker–firm match as a cluster, which places it outside the observation-level inference theory of Section 6; we therefore report no standard errors and read differences across specifications as descriptive.

firm inputs. Allowing a common degree-5 polynomial and then interacting the basis with gender, education, and qualification changes firm variance and sorting much less. Across polynomial, Hermite, and cubic B-spline implementations of the group-specific nonlinear model, firm variance lies between 0.114 and 0.117 and sorting between 0.041 and 0.042; worker variance ranges from 0.474 to 0.491. Thus which observables enter matters more for measured firm variance and sorting than how flexibly they are allowed to enter once they are included. The worker component is more sensitive to the nonlinear basis.

The limited movement across flexible specifications is itself economically informative. Predictive content and decomposition movement are distinct objects. A richer wage adjustment changes the AKM decomposition only when its fitted component reallocates wage variation along the worker–firm network: it must change the second moments of the worker and firm effects or their covariance. Nonlinear terms can improve conditional wage fit while leaving those moments nearly unchanged. On a held-out sample drawn so that the training data keep the worker–firm network connected, the fully interacted specification lowers mean squared error by 1.1% relative to heterogeneous linear controls, yet firm variance and sorting remain nearly unchanged. Our specification ladder separates two margins that standard robustness exercises often combine. The large movement from adding education, qualification, and firm inputs shows that observed worker composition and firm conditions matter. The smaller movement from letting those controls enter more flexibly indicates that measured firm heterogeneity and sorting are not mainly consequences of imposing additive linear controls; the greater basis sensitivity of worker variance identifies where the choice of approximating basis still matters. Generalized AKM turns this interpretation into an estimable restriction by asking whether the conventional components survive a much broader class of wage adjustments.

The second question is theoretical: when can these components be estimated at all without treating the covariate function as known? The problem is not a routine extension of either leave-out estimation or semiparametric series regression. Three sources of error must be handled jointly: many-regressor bias from high-dimensional fixed effects, heteroskedasticity, and approximation error from the unknown control function. Existing leave-out estimators address the first two under linear controls, while standard series results study approximation error for linear functionals of the coefficients, such as a single slope. A variance component is not of that kind: it is quadratic in the worker and firm effects, being a sum of squares and cross-products rather than a weighted average. That difference is what makes the problem hard: approximation error then enters both the plug-in quadratic form and its leave-out bias correction. We derive conditions under which both terms vanish together and show that the estimator is consistent and asymptotically normal for the original variance components. The key smoothness requirement is stronger than for linear functionals: the unknown function must have more continuous derivatives than the number of continuous covariates entering flexibly. The condition is nonetheless mild in wage applications. With age as the only continuous covariate, it asks that the wage–age profile be twice continuously differentiable; with the three firm inputs added, that

the wage surface in age and inputs have five derivatives. We also establish a chi-squared limit when the target has fixed rank, a normal limit when its rank grows with sample size, and a random-projection approximation for large administrative datasets.

The simulations show what this added flexibility buys. When the control function is strongly nonlinear, higher-degree bases reduce the bias, especially in high-leverage designs. When the relationship is linear or only mildly nonlinear, the flexible estimator performs similarly to the linear alternatives, so extra basis terms impose little cost. In a calibration that holds the Portuguese worker–firm graph fixed, the leave-out correction has point-estimate bias below 0.2% in all four nuisance designs, compared with 1.2–1.6% for the homoskedastic correction and 1.9–2.4% for the plug-in estimator. Finite-sample confidence intervals remain less reliable in several high-leverage stress designs, so the strongest numerical evidence concerns the decomposition itself.

The central contribution is to make AKM variance components estimable with a flexible and interacted covariate function without changing the objects being estimated. Relative to Kline, Saggio, and Sølvssten (2020), the estimator allows the approximation space to grow and states when leave-out validity survives the resulting approximation error. Relative to the semiparametric series literature (Donald and Newey 1994; Cattaneo, Jansson, and Newey 2018a, 2018b), the target is a quadratic form rather than a slope coefficient, which generates the stronger smoothness condition. The empirical application connects this result to work on the sensitivity of AKM decompositions (Andrews et al. 2008; Bonhomme et al. 2023; Bonhomme, Lamadon, and Manresa 2019): it separates the effect of adding observed controls from the effect of allowing nonlinearities and interactions among them. Unlike grouped-firm and latent-type approaches (Bonhomme, Lamadon, and Manresa 2019), Generalized AKM retains the original worker and firm variance components and asks how their estimated magnitudes change when the control function is no longer imposed.

Roadmap. Section 2 defines the semiparametric decomposition and explains why existing estimators do not directly apply. Sections 3–6 develop the leave-out estimator, its large-sample properties, and its random-projection approximation. The simulation in Section 7 isolates the gains from flexibility, and the Portuguese application in Section 8 shows how adding observables, nonlinearities, and interactions changes the decomposition.

2. Setup

2.1. Model and parameter of interest

We consider the following semiparametric model,

$$y_i = x_i' \beta + f(z_i) + e_i, \quad i = 1, \dots, n, \quad (1)$$

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

$\mathbb{E}[e_i^2] = \sigma_i^2$. In the wage application, x_i collects the high-dimensional worker and firm indicators whose dispersion we ultimately want to measure, while z_i collects the observed covariates (age, firm inputs, and so on) whose effect on wages we want to control for without committing to a functional form.

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 . The matrix A selects which feature of β to isolate: different choices pick out the variance of worker effects, the variance of firm effects, or their covariance.

2.2. Examples

The quadratic form covers many quantities of economic interest. The two examples below are the analysis-of-variance and two-way fixed-effects cases studied by Kline, Saggio, and Sølvesten (2020), whose notation and framing we follow so that the estimands remain directly comparable; the difference in each case is that the covariates enter through the unknown f rather than linearly.

Example 1 (Generalized analysis of variance). Let the data consist of N groups, with group g contributing T_g observations. Since Fisher (1928), the standard analysis of variance model assumes:

$$y_{gt} = \alpha_g + z'_{gt}\delta + \varepsilon_{gt}, \quad g = 1, \dots, N, \quad t = 1, \dots, T_g,$$

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

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

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

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

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 $e_i := \varepsilon_{gt}$,

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

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

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

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,

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

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

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

where $\bar{\psi} = \frac{1}{n} \sum_{g=1}^N \sum_{t=1}^{T_g} \psi_{j(g,t)}$. Given evidence that returns to worker characteristics are nonlinear (Mincer 1974; Card et al. 2018), part of the measured variance of worker and firm fixed effects may stem from the linear restriction on z_{gt} . Our methodology replaces it with unspecified functional dependence,

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

This formulation can capture nonlinear interactions between worker-level and firm-level characteristics that 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

$$x_i := (d'_i, h'_i)', \quad \beta := (\alpha', \psi')', \quad \alpha := (\alpha_1, \dots, \alpha_N)' + \mathbb{1}'_N \psi_0, \quad \psi = (\psi_1, \dots, \psi_J)' - \mathbb{1}'_J \psi_0,$$

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

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

2.3. Why existing estimators do not directly apply

Throughout, our object of interest is the quadratic form $\theta = \beta' A_\psi \beta$, defined on the high-dimensional fixed-effect coefficients, while the control function $f(z)$ is estimated through a growing approximation space. Several existing estimation strategies address parts of this setting; we review each in turn and explain why none covers the full problem.

KSS-style leave-out estimation. Kline, Saggio, and Sølvesten (2020), hereafter KSS, provide valid leave-out estimation of the same quadratic form $\theta = \beta' A \beta$ in finite-dimensional linear many-regressor models under arbitrary heteroskedasticity. Our setting preserves that object, but drops the assumption that observables enter through a finite-dimensional linear control. Once $f(z)$ is replaced by a series approximation $p_k(z)' \alpha$ with $k \rightarrow \infty$, validity is no longer automatic: both $\hat{\beta}$ and the leave-out variance estimators inherit approximation error, so one must re-establish that leave-out validity survives series approximation.³

Semiparametric series and orthogonalization. Series estimators in partially linear models (Donald and Newey 1994; Cattaneo, Jansson, and Newey 2018a) and orthogonalized or cross-fit procedures (Chernozhukov et al. 2018; Bonhomme, Jochmans, and Weidner 2025) allow an unknown nuisance function, but they estimate low-dimensional slope coefficients or other approximately linear functionals. None provides inference for quadratic forms under the high-dimensional fixed-effect structure and heteroskedasticity that define our setting.

Quadratic-functional estimation. Breunig and Chen (2022) study a closely related object: a leave-one-out sieve estimator of a quadratic functional, with minimax and adaptive rates, in a nonparametric instrumental variables model. Their setting has no high-dimensional fixed effects, and their smoothness threshold reflects the degree of ill-posedness of the inverse problem rather than many-regressor bias. Our contribution is the analogous threshold for a quadratic form in a many-fixed-effects leave-out design.

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 $p^1(z), \dots, p^k(z)$ be some *approximating* functions, and let $p_k(z) := (p^1(z), \dots, p^k(z)) \in \mathbb{R}^k$ collect them into a vector. We assume that the class of functions $\mathcal{F}$ to which f belongs can be well approximated by linear combinations of the basis functions $p^1(z), \dots, p^k(z)$.

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

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

and let the partialled-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

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

3. A separate strand of the literature relaxes the economic structure of AKM itself. For example, Bonhomme, Lamadon, and Manresa (2019) model latent worker and firm types and reduce dimensionality through grouping. These approaches replace the original variance component with a different earnings structure, rather than preserving $\theta = \beta' A \beta$ under an unknown nuisance function.

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

$$\begin{aligned}
\hat{\beta} &= S_{xx}^{-1} \sum_{i=1}^n \sum_{j=1}^n M_{ij} x_i (x'_j \beta + f(z_j) + e_j) \\
&= S_{xx}^{-1} \sum_{i=1}^n \sum_{j=1}^n M_{ij} x_i x'_j \beta + S_{xx}^{-1} \sum_{i=1}^n \sum_{j=1}^n M_{ij} x_i f(z_j) + S_{xx}^{-1} \sum_{i=1}^n \sum_{j=1}^n M_{ij} x_i e_j \\
&= \beta + S_{xx}^{-1} \sum_{i=1}^n \sum_{j=1}^n M_{ij} x_i f(z_j) + S_{xx}^{-1} \sum_{i=1}^n \sum_{j=1}^n M_{ij} x_i e_j := \beta + \mathcal{B} + \mathcal{U}.
\end{aligned} \tag{2}$$

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

$$\mathcal{B} = o(1)$$

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

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

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

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 i th 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. The intuition is that $\hat{\beta}' A \hat{\beta}$ squares the estimation noise in $\hat{\beta}$ together with the signal, so it overstates the true dispersion $\beta' A \beta$. The leave-out correction estimates that noise term observation by observation and subtracts it, using for unit i a residual variance that never draws on unit i 's own outcome.

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

$$\hat{\gamma} := \left(\sum_{j=1}^n w_j w'_j \right)^{-1} \sum_{j=1}^n w_j y_j,$$

and the leave-one-out version of it,

$$\hat{\gamma}_{-i} := \left(\sum_{j=1}^n w_j w'_j - w_i w'_i \right)^{-1} \left(\sum_{j=1}^n w_j y_j - w_i y_i \right).$$

Note that the 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 Sølvesten (2020), we

estimate θ as

$$\hat{\theta} := \hat{\beta}' A \hat{\beta} - \sum_{i=1}^n B_{ii} \hat{\sigma}_i^2, \quad (3)$$

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

$$\hat{\sigma}_i^2 := y_i (y_i - w_i' \hat{\gamma}_{-i}). \quad (4)$$

We note that computing $\hat{\gamma}_{-i}$ for each $i = 1, \dots, n$ is computationally costly in large-scale applications. To avoid this, we represent (4) as

$$\hat{\sigma}_i^2 = \frac{y_i \hat{e}_i}{M_{W,ii}}, \quad (5)$$

where $M_{W,ii}$ is i th 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 is general in the sense that it nests the linear leave-out estimator of KSS as a special case. The generality comes from enlarging the assumed function class from the linear class $\mathcal{F}_{\text{linear}}$ to the broader class $\mathcal{F}$ of nonlinear functions satisfying Assumption 1, so that $\mathcal{F}_{\text{linear}} \subseteq \mathcal{F}$.

To see this more clearly, note that our estimator $\hat{\theta}$ is *exactly* unbiased if the underlying functional class is linear, $f \in \mathcal{F}_{\text{linear}}$. It follows from decomposition as in (2) and noting that the $\mathcal{B}$ term disappears by the properties of the M matrix that projects onto the complement of the space spanned by z_i (where we set $P_k := P_d = (z_1, \dots, z_n) \in \mathbb{R}^{n \times d}$), in which case $\sum_{i=1}^n \sum_{j=1}^n M_{ij} x_i z_j' \alpha = 0$. Under the condition that the unknown function $f(z_i) = z_i' \alpha$ is linear in z (as in Kline, Saggio, and Solvsten (2020)), as a result $\mathbb{E}[\hat{\beta}] = \beta$ holds, and we have

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

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 more general classes of functions $\mathcal{F}$, the exact unbiasedness of $\hat{\beta}$ and $\hat{\sigma}_i^2$ does not hold. Instead, we study an asymptotic sequence in which the number of approximating functions diverges, $k \rightarrow \infty$, and show that the bias terms vanish under

explicit conditions. Concretely,⁴

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

along sequences satisfying the assumptions below.

For these purposes, we assume that the unknown function $f(z)$ belongs to a class $\mathcal{F}$ of smooth functions that can be well approximated by linear combinations of the basis functions $p^1(z), \dots, p^k(z)$.

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

$$\mathcal{F} := \left\{ f : \min_{\alpha \in \mathbb{R}^k} \mathbb{E} [|f(z_i) - p_k(z_i)' \alpha|^2] \leq C k^{-2\alpha_f}, \alpha_f > 1 \right\} \quad (6)$$

for some absolute constant $C < \infty$. Furthermore, $\sup_{f \in \mathcal{F}} f < M$ for some absolute constant $M < \infty$.

This implies that

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

so the mean-square approximation error decays at rate $k^{-2\alpha_f}$ uniformly over $\mathcal{F}$.

The constant α_f governs how fast approximation bias decays. To obtain consistency for the quadratic form studied here, we strengthen the usual semiparametric requirement from $\alpha_f > 0$ in Donald and Newey (1994) and Cattaneo, Jansson, and Newey (2018a) to $\alpha_f > 1$: compared with slope-type estimands, quadratic forms require faster decay of approximation error. For polynomial or spline series on compact support with d_{cont} continuous covariates, $\alpha_f = s_f/d_{\text{cont}}$ where s_f is the number of continuous derivatives (Chen 2007), so the assumption reduces to $s_f > d_{\text{cont}}$: the unknown function must have more continuous derivatives than the number of continuous covariates entering flexibly. Discrete covariates such as gender, occupation, or region do not count toward d_{cont} because they are absorbed by group interactions rather than by the polynomial basis. In the wage applications of Section 8, the parsimonious sample has $d_{\text{cont}} = 1$ (age), so $s_f > 1$ suffices—the wage–age profile need only be twice continuously differentiable. The Panel B specification adds employment, fixed assets, and intermediate inputs, giving $d_{\text{cont}} \leq 4$ and requiring $s_f > 4$. Both conditions are mild for smooth economic relationships. The condition is sufficient but not shown to be necessary. This mirrors the elbow phenomenon in minimax estimation of quadratic functionals, where reaching the parametric rate requires more smoothness than a linear functional of the same nuisance would need (Breunig and Chen 2022). Even though the assumption on α_f is strengthened compared to the standard linear setups, it is equivalent to the assumption commonly placed in settings with nonlinear structure, as in, for example, Hirano, Imbens, and Ridder (2003) and Farrell, Liang, and Misra (2021).

4. 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 $\|v\| = o_p(1)$.

Remark 1. *The results cover a fixed number of continuous covariates and a fixed collection of discrete interaction groups, as in Section 8. They do not cover a growing continuous dimension, a growing number of groups, or nuisance functions estimated by machine learning—random forests, boosted trees, or neural networks—whose approximating class is not the span of a known, pre-specified basis.*⁵

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 of Scheffé (1959), Searle, Casella, and McCulloch (2009), and Kline, Saggio, and Sølvesten (2020), treating x_i , z_i , and A as fixed. This allows us to be agnostic about potential dependence between x_i , z_i , and A . Our analysis differs from Cattaneo, Jansson, and Newey (2018b), who consider sequences of random variables and condition on z_i only. Limits are taken assuming the number of observations goes to infinity, $n \rightarrow \infty$, the number of approximating functions goes to infinity, $k \rightarrow \infty$ (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, $p \rightarrow \infty$, to model the limited mobility bias that arises when workers move between only a few firms, so that worker and firm effects are estimated from limited variation. We make the following additional assumptions:

Assumption 2 (Data-generating process). *(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 $P_{W,ii} := 1 - M_{W,ii}$; (iii) $\max_i (x_i' \beta)^2 = \mathcal{O}(1)$.*

Part (i) excludes errors without a uniformly bounded fourth moment. The standardized Student- t_5 design in Section 7 is therefore a finite-sample tail-stress case, not a violation of the assumption. Together with Assumption 1, parts (ii) and (iii) 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 (Growth rate conditions). *We assume that $k \rightarrow \infty$ and $p \rightarrow \infty$ as $n \rightarrow \infty$, and $k = o(n)$, $k = o(p)$, $n = o(k^{\alpha_f})$, and $p = \mathcal{O}(n)$.*

This assumption places restrictions on the relationship between growth rates of the number of observations, linear covariates, and approximating functions. In particular, it reflects empirical practice in the AKM literature where the number of observations is of a much larger magnitude than the potential number of series terms. It also places a lower bound on the number of series terms, since k must be large enough that the n/k^{α_f} term is small. Intuitively, the window $n^{1/\alpha_f} \ll k \ll n$ balances zero asymptotic bias against a negligible added variance contribution. Relaxing the assumption to cover $k = \mathcal{O}(n)$ would be of independent theoretical interest, though of limited empirical relevance in the AKM setting.

5. Bakirov, Del Prato, and Zacchia (2026) take that route on the same Portuguese data, replacing the worker and firm fixed effects with gradient-boosted trees fit to observable characteristics.

We conjecture that letting the number of series terms grow proportionally with the number of observations would add a further term to the asymptotic variance of $\hat{\theta}$, in line with the theory on linear functionals with growing dimension (Cattaneo, Jansson, and Newey 2018b). We leave a proof to future work.

Shared rate restrictions for the proved results. The results below all rely on the same core asymptotics: the series dimension grows ($k \rightarrow \infty$), total regressor dimension remains bounded away from the sample size through $\max_i P_{W,ii} < c < 1$, and approximation bias decays fast enough through $\alpha_f > 1$. The theorem-specific conditions then add the restrictions needed for each result: Lemma 1 requires $\text{trace}(\tilde{A}^2) = o(1)$, Theorem 3 adds the fixed-rank leverage condition $\max_i v_i' v_i = o(1)$, and Theorem 4 adds the growing-rank negligibility conditions on $\max_i ((\tilde{w}_i' \gamma)^2 + (\check{w}_i' \gamma)^2)$ and on the leading eigenvalue share $\lambda_1^2 / \sum_{\ell=1}^r \lambda_\ell^2$. The tradeoff behind the fixed-dimensional series results is simple: k must grow fast enough for approximation error to vanish, but not so fast that the combined regressor dimension $p + k$ drives leverage close to one.

Relative to linear leave-out estimation, replacing linear controls by a growing basis adds one additional task: one must show that the approximation terms from the unknown nuisance are negligible both for $\hat{\beta}$ and for the leave-out residual-variance correction. Lemma 1 and Theorems 3–4 establish exactly this under Assumptions 1–3.

Corollary 1. *Suppose the covariates entering the unknown function can be written as $z_i = (u_i', s_i)'$, where $u_i \in \mathbb{R}^{d_{\text{cont}}}$ has fixed dimension, $s_i \in \{1, \dots, G\}$ indexes a fixed collection of discrete groups, and*

$$f(z_i) = \sum_{g=1}^G \mathbf{1}\{s_i = g\} f_g(u_i).$$

Assume the support of u_i is compact and that, for the specified basis $\{p^j(u_i)\}_{j=1}^k$, each f_g satisfies

$$\inf_{\alpha_g \in \mathbb{R}^k} \mathbb{E} \left[|f_g(u_i) - p_k(u_i)' \alpha_g|^2 \right] \leq C k^{-2s_f/d_{\text{cont}}}$$

for a common constant $C < \infty$ and common smoothness index $s_f > d_{\text{cont}}$, uniformly over g . Then the interacted basis $\{\mathbf{1}\{s_i = g\} p^j(u_i)\}_{g,j}$ satisfies Assumption 1 with $\alpha_f = s_f/d_{\text{cont}} > 1$. If, in addition, Assumptions 2 and 3 hold together with the theorem-specific leverage and eigenvalue conditions, then the conclusions of Lemma 1, Theorem 3, and Theorem 4 apply to this interacted specification.

Let $\tilde{A} := S_{xx}^{-1/2} A S_{xx}^{-1/2}$ and write its eigendecomposition as $\tilde{A} = Q \Lambda Q'$, where $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_r)$ collects the r nonzero eigenvalues and $Q \in \mathbb{R}^{p \times r}$ the corresponding eigenvectors.

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

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

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 Sølvesten (2020) by considering the random projection method of Achlioptas (2003), known as the Johnson–Lindenstrauss approximation (JLA), which names the estimator below. 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

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

The proposed estimator is then:

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

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

(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 $\frac{\text{var}[\hat{\theta}_1] \text{var}[\hat{\theta}_2]}{n \text{var}[\hat{\theta}]^2} = \mathcal{O}(1)$.*

6. Limiting distributions

This section derives two limiting distribution results for $\hat{\theta}$. The first assumes that the rank r of A is fixed; the second allows r to grow with n .

6.1. Fixed rank

First, we derive the limiting distribution of $\hat{\theta}$ for fixed r . Using the eigendecomposition $\tilde{A} = Q\Lambda Q'$ introduced before Lemma 1, we represent the estimator as

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

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

The next theorem gives the fixed-rank limit for $\hat{\theta}$ and consistency of $\widehat{\text{var}}[\hat{b}]$.

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

- (i) $\text{var}[\hat{b}]^{-1/2}(\hat{b} - b) \xrightarrow{d} \mathcal{N}(0, I_r)$, where $b := Q' S_{xx}^{1/2} \beta$,
- (ii) $\text{var}[\hat{b}]^{-1} \widehat{\text{var}}[\hat{b}] \xrightarrow{p} I_r$,
- (iii) $\hat{\theta} = \sum_{\ell=1}^r \lambda_{\ell} \left(\hat{b}_{\ell}^2 - \text{var}[\hat{b}_{\ell}] \right) + o_p(\text{var}[\hat{\theta}]^{1/2})$.

The growing-rank case is more relevant for AKM-type applications, so we turn to that setting next.

6.2. Growing rank

Define $\tilde{w}_i := \check{A} S_{ww}^{-1} w_i$, where $\check{A} := \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix}$ and $\gamma := \begin{pmatrix} \beta & \alpha \end{pmatrix}'$. Then

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

and the leave-one-out estimator can be written as

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

where the approximation error from replacing $\gamma' w_i$ by y_i is controlled by Assumption 1 and 3. A direct algebraic rearrangement yields

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

where $C_{i\ell} := B_{W,i\ell} - 2^{-1} M_{W,i\ell} \left(B_{W,ii} M_{W,ii}^{-1} + B_{W,\ell\ell} M_{W,\ell\ell}^{-1} \right)$. Thus $\hat{\theta}$ is a second-order U -statistic with sample-dependent kernel $C_{i\ell}$. Theorem 4 combines this representation with Appendix Lemma 3; relative to Kline, Saggio, and Solvsten (2020), the only additional term is the series approximation error, which is negligible under Assumption 1 and 3.

Define also

$$\tilde{w}_i := \sum_{\ell=1}^n M_{W,i\ell} \frac{B_{W,\ell\ell}}{1 - P_{W,\ell\ell}} w_{\ell}.$$

Theorem 4. *If Assumption 1, 2, and 3 hold, and the following conditions are satisfied,*

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

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

7. Simulation study

We use a Monte Carlo study to evaluate the estimator under controlled conditions that isolate the problems described in the theory. In the baseline design, the covariates z_i are drawn uniformly on a fixed cube, the linear regressors x_i (so called because they enter the outcome linearly) are generated as nonlinear functions of z_i , and the quadratic form of interest is $\theta = \beta' A \beta$, where A selects the first two coordinates of β , so it measures dispersion in two fixed effects. Unless otherwise noted, the nonlinear designs use $n = 500$, nuisance dimension $d = 4$, and $p = 90$ linear regressors. The rows are constructed as comparative statics: within a panel, the simulation holds the underlying random design fixed whenever dimensions permit and changes only the feature named in the row—the nonlinearity of $f(z)$, the error distribution, the nuisance dimension, the number of linear regressors, leverage, or sample size.

More specifically, each scenario first draws a fixed design $(z_i, x_i)_{i=1}^n$ and coefficient vector β , then holds that design fixed across Monte Carlo replications while redrawing only the error term e_i . The first panel of Table 1 uses the same z_i , x_i , β , and error draws and changes only the functional form of f . The second panel keeps the strongly nonlinear radial design as the reference case; the heteroskedastic and heavy-tailed rows change only the error process, while the remaining rows change one of the other named features using nested random draws where possible. The nuisance covariates satisfy $z_i \sim \text{Unif}([-1, 1]^d)$. Before standardization, the linear regressors are

$$x_{ij}^* = \exp\{0.30 \|z_i\|_2 + 0.50 \eta_{ij} + 0.10 \nu_j\}, \quad \eta_{ij}, \nu_j \sim \mathcal{N}(0, 1),$$

where η_{ij} is an idiosyncratic shock drawn separately for each observation and regressor, and ν_j is a single shock shared by every observation in regressor column j ; each column of x_i^* is centered and scaled. The shared term $0.30 \|z_i\|_2$ correlates every regressor with the nuisance covariate, entangling x_i and z_i just as the bias term $\mathcal{B}$ in (2) requires: without that correlation, a linear regression of y on x alone would already be unbiased for β , leaving nothing for the semiparametric correction to fix. The nuisance signal is also centered and scaled before entering the outcome. We use three functional forms:

$$f_{\text{linear}}(z_i) = 0.8z_{i1} - 0.5z_{i2} + 0.35z_{i3} - 0.2z_{i4}, \quad f_{\text{cubic}}(z_i) = 0.8z_{i1}^3 - 0.5z_{i2}^2 + 0.3z_{i3}z_{i4},$$

and the strongly nonlinear radial design

$$f_{\text{radial}}(z_i) = \|z_i\|_2^7.$$

The radial seventh-power function is the main nonlinear stress test: it differs sharply from any linear control because its slope rises quickly as observations move away from the center of the cube. The error term is standard Gaussian, heteroskedastic Gaussian with scale $0.5 + 0.5|x_{i1}|$ normalized by its fixed-design root mean square, or a variance-one Student- t distribution with five degrees of freedom.⁶ The many-regressor, higher-dimension, and larger-sample designs each raise one of p , d , or n above its baseline value, and the high-leverage design inflates a small share of extreme observations to create thinly supported, high-leverage points.⁷

We compare the plug-in estimator with leave-out estimators that use complete total-degree polynomial bases of degree 1, 3, and 5, including all monomials whose total degree does not exceed the stated degree. Cubic polynomials are common in applied work (Card et al. 2018); the degree-5 basis lets us check whether going beyond cubic matters. Table 1 reports the subset of designs that cover the main cases in the paper: approximation error and nonlinearity, robustness to heteroskedasticity and heavy tails, and finite-sample performance as nuisance dimension, the number of regressors, leverage, and sample size change one at a time. The first four columns report signed Monte Carlo bias for the plug-in (PI) and leave-out (LOO) estimators at each basis degree. The fifth column reports RMSE for LOO at degree 5. Appendix Table B1 reports nominal 95% coverage for the same estimator separately, with confidence intervals based on the estimated variance of the quadratic form computed from the leave-one-out variances $\hat{\sigma}_i^2$.

Approximation error and the smoothness condition. Assumption 1 predicts that series approximation error decays at rate $k^{-2\alpha_f}$, so higher-degree bases should reduce bias most when f departs strongly from linearity. The simulations confirm this. In the strong radial design, the bias falls in magnitude from about -0.04 for PI(1) and LOO(1) to -0.02 for degree 3 and -0.01 for degree 5. Under high leverage, the same pattern holds: the bias falls from about -0.02 for PI(1) and -0.03 for LOO(1) to zero after adding higher-degree basis terms. By contrast, in the linear benchmark and the mildly nonlinear cubic design, the bias is already close to zero, with the largest entry about 0.01. The first panel of Table 1 should therefore be read as evidence on where the smoothness condition matters in finite samples: adding basis terms removes systematic misspecification bias when nonlinearity is

6. These are population or fixed-design normalizations; we do not divide each Monte Carlo draw by its realized sample standard deviation.

7. In Table 1, $p = 90$ in all scenarios except the many-regressor design, where $p = 300$. The nuisance dimension is $d = 4$ in all scenarios except the higher-dimension design, where $d = 8$. The sample size is $n = 500$ in all scenarios except the larger-sample design, where $n = 1000$. In the high-leverage design, the regressors for the 5% of observations with the largest $|z_{i1}|$ —those farthest from the center of the cube along the first nuisance coordinate—are additionally scaled up by a factor of six in the first ten regressor columns (or all columns, if $p < 10$), creating a small number of extreme, thinly supported points that drive up their own leverage $P_{W,ii}$.

Table 1: Simulation study: bias and finite-sample performance

Scenario	PI(1)	LOO(1)	LOO(3)	LOO(5)	RMSE(5)
<i>Approximation and nonlinearity</i>					
Mild nonlinearity	0.01	0.00	0.00	0.00	0.05
Linear benchmark	0.00	0.00	0.00	0.00	0.05
Strong nonlinearity	-0.04	-0.04	-0.02	-0.01	0.05
<i>Strong radial design: one-feature changes</i>					
Heteroskedasticity	-0.03	-0.04	-0.02	-0.01	0.05
Heavy-tailed errors	-0.03	-0.04	-0.01	-0.01	0.05
Higher nuisance dimension	-0.05	-0.05	-0.02	NA	NA
Many regressors	-0.01	-0.03	-0.02	-0.01	0.06
High leverage	-0.02	-0.03	-0.01	0.00	0.02
Larger sample	0.02	0.01	-0.01	0.00	0.03

Notes: Columns PI(1), LOO(1), LOO(3), and LOO(5) report signed Monte Carlo bias, $\mathbb{E}[\hat{\theta} - \theta]$, for the plug-in (PI) estimator with a degree-1 basis and the leave-out (LOO) estimator with degree-1, degree-3, and degree-5 bases, respectively. Positive entries indicate overestimation of the quadratic form. RMSE(5) reports the root mean squared error of the degree-5 LOO estimator. Bias and RMSE are reported in the units of the quadratic-form target. Appendix Table B1 reports the corresponding nominal 95% coverage diagnostic. Intervals use the estimated variance of the quadratic form computed from the leave-one-out residual variances $\hat{\sigma}_i^2$. The first block varies only the functional form of $f(z)$. The second block holds the strong radial design fixed and changes only the feature named in the row, relative to the strong-nonlinearity row. Each estimable design uses 5,000 Monte Carlo replications. NA means that the complete basis is not estimable: in the higher-dimension row, $d = 8$ gives the degree-5 basis 1,287 columns, so $p + k > n$.

large, and costs little when it is not.

Heteroskedasticity and error tails. The leave-out correction in (3) uses observation-specific residual variances $\hat{\sigma}_i^2$, so it depends on the moment conditions in Assumption 2(i). The heteroskedastic and heavy-tailed designs show why the table reports both the bias and RMSE. Because these rows now keep the same radial design as the homoskedastic strong-nonlinearity row, the low-degree bias is similar: about -0.03 to -0.04 before the higher-degree basis removes most of it. Degree-5 RMSE is about 0.05 under both heteroskedasticity and Student- t_5 errors. These designs therefore create dispersion that is not captured by the bias alone; adding basis terms removes the approximation component, but residual-variance estimation remains a finite-sample issue.

Leverage and many regressors. Leverage measures how much an observation's own outcome drives its own fitted value; it is high when a worker or firm effect is pinned down by very little data, as with rarely observed workers or small firms. Assumption 2(ii) requires $\max_i P_{W,ii} < c < 1$, bounding individual leverage away from one as the combined regressor dimension $p + k$ grows. The many-regressor and high-leverage designs make this condition hard to satisfy. Point estimation remains reliable: in the many-regressor design, the degree-5 leave-out estimator has bias around -0.01 and RMSE of about 0.06. Appendix Table B1 reports the corresponding coverage diagnostic: nominal 95% coverage is about 0.86 in the high-leverage design, 0.97 in the many-regressor design, 0.90 under

heteroskedasticity, and 0.91 under Student- t_5 errors.⁸ In the higher-dimension design, degree 3 remains feasible, with bias around -0.02 , but the complete degree-5 basis is infeasible and is reported as “NA.” The pattern points to residual-variance estimation and leverage, rather than point-estimate bias alone, as the main finite-sample problem. Larger samples improve point accuracy: bias is near zero, RMSE falls to about 0.03, and intervals again cover close to 95% of draws.

Summary. For applied work, adding basis terms is most valuable when nonlinear misspecification is large, especially under high leverage. When those features are absent, additional flexibility costs little but gains little. The point estimator is reliable across most designs.

AKM-calibrated Monte Carlo. We adapt the calibration logic in Section 9.5 of Kline, Saggio, and Sølvesten (2020) to our application.⁹ We hold fixed a small connected mover graph (a set of workers and firms linked by job moves, within which worker and firm fixed effects are jointly identified) with 13,939 person-year observations, 2,393 workers, 639 firms, and 6,366 worker–firm matches, using only gender and age as controls. The OLS worker and firm effects on this graph are rescaled so that their variances match the leave-out estimates. Each match shock is drawn from a standardized Student- t distribution with five degrees of freedom, with a heteroskedastic variance. Following Kline, Saggio, and Sølvesten (2020), we let that variance depend on the firm-variance diagonal B_{gg} , the leverage P_{gg} , and the log employment of the worker’s current and other firms through an exponential function fit to the leave-out variance estimates.¹⁰ The design crosses linear and degree-5 nuisance DGPs with linear and degree-5 fitted nuisance specifications, holding firm assignments, covariates, and year assignments fixed across 1,000 draws.

Table 2: Portuguese-graph calibrated Monte Carlo for firm-effect variance

DGP	Nuisance fit	KSS bias (MC SD)	HO bias	PI bias
Linear DGP	Flexible	0.06% (0.50%)	1.24%	1.90%
Linear DGP	Linear	0.06% (0.48%)	1.25%	1.90%
Nonlinear DGP	Flexible	0.07% (0.62%)	1.31%	1.98%
Nonlinear DGP	Linear	0.18% (0.62%)	1.60%	2.37%

Notes: The true value is the calibrated firm-effect variance, and all bias entries are percentages relative to that value. KSS bias (MC SD) reports the relative bias of the KSS point estimate, followed in parentheses by the Monte Carlo standard deviation of the relative estimation error. The parenthesized Monte Carlo standard deviation measures dispersion across replications, not a statistical standard error. HO bias and PI bias report the relative bias of the homoskedasticity-only correction and the plug-in estimator, respectively. Linear and Flexible denote the linear and degree-5 nuisance fits. Each design uses 1,000 draws.

8. Every design here under-covers except many regressors, which over-covers: the estimated variance used to build the interval is conservative when p is large relative to n , so the interval is too wide rather than too narrow. Only the under-coverage cases are the concerning direction for applied inference.

9. We do not reproduce their first-differenced mover-pair design.

10. The exponential model is fit by nonlinear least squares in levels to all finite leave-out variance estimates, using the positive estimates only to initialize. Because the KSS regression premultiplies each match mean by the square root of its person-year count, the fitted variance is in weighted-match units; we divide it by the match count before assigning a common shock to the match’s person-year observations.

Table 2 evaluates the point correction. KSS relative bias is 0.06–0.18% across the four designs, compared with 1.24–1.60% for the homoskedastic correction and 1.90–2.37% for the plug-in estimator. The KSS Monte Carlo standard deviation is 0.48–0.62% of the true value. The KSS bias is largest in the Nonlinear DGP, Linear-fit design—the case where the true relationship is nonlinear but the fitted specification is linear—but it stays tiny throughout, because this parsimonious calibration carries only an age profile. The exercise therefore checks only that the leave-out point correction stays nearly unbiased on a realistic mobility graph. It does not speak to the functional-form sensitivity found in the application, which operates through the firm inputs that this design excludes.

This exercise checks point-estimate bias, not inference, and the feasible standard-error calculation is not validated by it: the calculation excludes about half of the match-level leave-out variance estimates, and the surviving estimates inflate the standard error six- to eightfold relative to the Monte Carlo standard deviation—so far that the nominal 95% interval covers in all 1,000 draws of every design, the signature of an interval too wide to be informative rather than a validated one. We treat those interval results as diagnostics, not as evidence for cluster-level inference.

8. Empirical application

We use Portuguese linked employer–employee data to study how alternative control specifications change the worker–firm wage decomposition. Adding worker and firm-input controls lowers all three variance components. Allowing those controls to enter nonlinearly produces smaller movements in firm variance and sorting, while the worker component moves more across specifications.

8.1. Data and sample selection

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, employment) for the universe of private-sector firms. The firm accounts dataset *Central de Balanços* (CB) provides annual balance-sheet and income-statement information, including fixed assets and intermediate inputs, for non-financial corporations. We link QP and CB at the firm level to obtain a panel with worker characteristics and firm inputs, following the same linkage of Portuguese wage records to firm accounts used by Card, Cardoso, and Kline (2016).

We restrict our sample to the period 2008–2018 and to full-time workers aged 20–65. The dependent variable is the log of individual hourly wages. Following the standard AKM literature, we decompose the cross-sectional variance of log wages into the variance of worker fixed effects (σ_α^2), the variance of firm fixed effects (σ_ψ^2), and their covariance $\text{cov}(\alpha, \psi)$, which captures sorting. Detailed sample descriptives, estimation-variable coverage, and sample-flow statistics are reported in Appendix Tables B2, B3, and B4.

8.2. Estimation

We study two versions of this empirical exercise. The first is a parsimonious specification that uses only worker-side controls: gender and age. The second specification adds worker education and qualification together with $\log$ firm employment, $\log(1 + \text{fixed assets})$, and $\log(1 + \text{intermediate inputs})$. About 7.6% of firm-years report exactly zero fixed assets, so the two monetary inputs enter as $\log(1 + x)$ rather than $\log(x)$ to retain these economically meaningful zeros.¹¹ The transformations also keep the raw monetary scales from dominating the polynomial basis, and negative accounting entries are treated as missing. In the empirical tables, we label these two exercises as Panel A (parsimonious controls) and Panel B (worker and firm-input controls). Because Panel B requires non-missing firm controls, the main empirical tables estimate both panels on the Panel B non-missing *leave-match-out* set: the largest connected worker–firm set (linked by job moves, so that worker and firm fixed effects are jointly identified) that supports the match-level correction.¹² This keeps the main cross-panel comparison about controls rather than sample composition; Appendix Table B4 reports the larger parsimonious candidate sample as well as the common main estimation sample.

Within each of these two control sets, we estimate the same nested five-step specification ladder, progressively relaxing the restrictions on how covariates enter the wage equation: (i) a pure AKM baseline with no observed controls beyond the fixed effects; (ii) linear additive controls, meaning common quadratic and cubic age terms centered at 50, with the common linear age direction excluded because it is not identified with worker and year fixed effects, together with education, qualification, and firm inputs entering linearly and additively; (iii) heterogeneous linear controls, which add demographic-group deviations in the age profile and group-specific slopes on the firm inputs; (iv) nonlinear homogeneous controls, which replace those terms with a common degree-5 polynomial basis; and (v) the full model, which interacts that basis with demographic groups. The ladder separates the roles of adding observables, allowing slope heterogeneity, and allowing nonlinearity. Because the linear additive rung already includes quadratic and cubic age terms, the move to the degree-5 basis adds higher-order age terms and richer functions of the firm inputs, including their multivariate interactions. The polynomial basis is interacted with a fully saturated set of worker categorical variables, so the full model allows group-specific nonlinear effects of each continuous covariate: in Panel A the interaction groups are defined by gender, while in Panel B they are defined by $\text{gender} \times \text{education} \times \text{qualification}$.

We use the polynomial basis for the main ladder and re-estimate its two nonlinear rungs

11. The transformation $\log(1 + x)$ is not invariant to the units in which x is measured (Chen and Roth 2024): rescaling the monetary inputs changes where the accounting zeros sit relative to the positive values. This bears on the linear specifications more than on the flexible ones, which approximate an arbitrary smooth function of each input and are therefore nearly unaffected by the choice of units. Fixed assets and intermediate inputs are held at their recorded scale throughout, and both enter only as nuisance controls whose coefficients we do not interpret.

12. The source panel spans 2008–2018, but fixed assets are unavailable in 2008–2009. After imposing the Panel B non-missing firm-control requirement, the common Panel A–Panel B estimation sample is therefore effectively drawn from 2010–2018; Appendix Table B5 reports the year-by-year joint availability of the firm controls.

with two alternatives. Probabilists’ Hermite polynomials provide a different coordinate system for degree-5 polynomial terms.¹³ Additive cubic B-splines provide a different functional form with quantile-spaced knots. We choose the spline counts so that each alternative has the same realized control dimension as the corresponding polynomial model: $k = 4$ and 9 in Panel A, and $k = 129$ and 3,004 in Panel B for the common and group-specific nonlinear specifications. Appendix Table B7 compares the resulting full-regressor corrected decompositions.

Mapping to theory. To map the empirical specification into the model in Section 2, let z_i^{cont} denote the continuous covariates that enter flexibly and let $s_i \in \{1, \dots, G\}$ denote the worker categorical group that defines the interaction cell. In Panel A, z_i^{cont} contains age and s_i is gender. In Panel B, z_i^{cont} contains age, log employment, $\log(1 + \text{fixed assets})$, and $\log(1 + \text{intermediate inputs})$, while s_i indexes the fixed collection of gender $\times$ education $\times$ qualification cells. Writing x_i for the regressors that remain outside the flexible nuisance term—the worker and firm fixed effects, year effects, and any controls that enter linearly at a given specification—the full specification can be written as

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

This is a special case of (1): the unknown function is simply evaluated on the enlarged argument $(z_i^{\text{cont}'}, s_i)'$, so the nuisance is allowed to be a different smooth function in each worker group.

Our implementation centers and scales the continuous inputs, approximates each group-specific function f_s with the same polynomial family, and lets the coefficients vary across groups. Concretely, if $\{p^j(z_i^{\text{cont}})\}_{j=1}^k$ is the polynomial basis, then the interacted series uses the terms $\{\mathbb{1}\{s_i = s\} p^j(z_i^{\text{cont}})\}_{s,j}$. Equivalently, the model fits a separate smooth profile of age in Panel A, and separate smooth profiles of age and the three transformed firm inputs in Panel B, within each worker category.

For the theory, these interactions do not increase the continuous dimension of the approximation problem. The indicators $\mathbb{1}\{s_i = s\}$ split the sample into a fixed number of groups, but smoothness is still imposed only with respect to the continuous arguments in z_i^{cont} . Therefore Assumption 1 applies within each group at rate $\alpha_f = s_f/d_{\text{cont}}$, where d_{cont} is the number of continuous covariates entering the nuisance and s_f is the number of their continuous derivatives. This is precisely the fixed-dimensional interacted-series case covered by Corollary 1: $d_{\text{cont}} = 1$ in Panel A and $d_{\text{cont}} = 4$ in Panel B.

The empirical point correction uses the same augmented regressors as the theory: worker effects, firm effects, year effects, and every control-function term enter one joint

13. A low-order term—such as the linear age profile, which is excluded because it is not separately identified from the fixed effects—can be absorbed differently by different bases: the fixed-effect normalization pins down which specific combination of basis terms counts as part of the fixed effects rather than the control function, so changing the basis (monomial versus Hermite) changes how that excluded, collinear variation is split between the two. The resulting decomposition can therefore differ slightly from the monomial specification, even though both bases approximate the same underlying function.

sparse regression, and the correction is computed from that full design rather than from a residualized fixed-effect regression. The usual two-step alternative — preadjusting wages for covariates and then applying the leave-out correction to the residualized outcome — introduces a higher-order bias because the first-step estimation error enters both the outcome and the residual used in the correction (Kline 2024); the joint design avoids this by estimating the control function alongside the fixed effects rather than partialling it out beforehand. The implementation differs in its deletion and dependence structure. The theory deletes independent observations one at a time. The application instead uses 20 outcome-independent cross-fit folds, treating an entire worker–firm match as one cluster for movers (workers observed at more than one firm) and individual person-years as singleton clusters for stayers (workers observed at only one firm); deleting a stayer’s only match would remove that worker’s column from the training design. This full-regressor cluster-fold correction aligns the empirical point estimand with the augmented-regressor construction in Section 3. Its cluster structure and fold-shared residuals remain outside the observation-level inference theory, so the results below report point estimates only, without theorem-based standard errors.

The interaction groups are defined from worker-side categorical variables, not firm identifiers. This matters for identification. Interacting the series with firm identifiers would produce columns that are nearly collinear with the firm fixed effects already contained in X , undermining the full-rank condition on $S_{xx} = X'MX$ required by Assumption 2.¹⁴

The firm-side controls that distinguish Panel B from Panel A are employment, fixed assets, and intermediate inputs. Appendix Table B5 reports their within-firm variation and missingness in the cleaned firm-year panel. On their estimation scales, their within-firm variance shares are about 7.9% for employment, 17.8% for fixed assets, and 9.8% for intermediate inputs, so most of their variation is cross-sectional. On the raw headcount and euro scales the corresponding within-firm shares are only 5.8%, 1.6%, and 5.0%: differences in firm size across firms dominate the year-to-year movements within a firm. Taking logs removes that size gap and raises the within-firm share of fixed assets more than tenfold, from 1.6% to 17.8%. Specification diagnostics for each specification are reported in Appendix Table B9.

8.3. Results

Table 3 reports the full-regressor bias-corrected variance components for each specification on the common Panel B leave-match-out sample. We report point estimates and do not attach standard errors because the cluster-fold correction lies outside the observation-level inference theory developed above. Cross-specification differences are descriptive. Table 4 reports $\text{var}(\tilde{y})/\text{var}(y)$, the ratio of residualized-wage variance to total wage variance after partialling out observables and year effects; values above one signal that near-collinear controls are amplifying rather than absorbing variance. Appendix Table B6 decomposes the

14. Age is normalized at 50 throughout, and the common linear age term is excluded because it is not identified in the presence of worker and year fixed effects. All variance components in Tables 3 and B6 are reported as shares of total log-wage variance on the same leave-match-out estimation set.

Table 3: Wage variance decomposition

Model	σ_α^2	σ_ψ^2	$\text{cov}(\alpha, \psi)$
<i>Panel A: Parsimonious controls (Panel B sample)</i>			
AKM baseline	0.5507	0.1435	0.0803
Linear additive controls	0.5326	0.1367	0.0738
Heterogeneous linear controls	0.5388	0.1366	0.0748
Nonlinear homogeneous controls	0.5483	0.1358	0.0775
Full model	0.5619	0.1371	0.0767
<i>Panel B: Worker and firm-input controls</i>			
AKM baseline	0.5507	0.1435	0.0803
Linear additive controls	0.4736	0.1214	0.0465
Heterogeneous linear controls	0.4506	0.1198	0.0446
Nonlinear homogeneous controls	0.4913	0.1199	0.0410
Full model	0.4904	0.1166	0.0420

Notes: Entries report full-regressor cluster-fold bias-corrected variance components as shares of total log-wage variance on the common Panel B leave-match-out estimation sample: worker-effect variance ($\sigma_\alpha^2 / \text{var}(y)$), firm-effect variance ($\sigma_\psi^2 / \text{var}(y)$), and sorting covariance ($\text{cov}(\alpha, \psi) / \text{var}(y)$). The table reports the covariance once; its contribution to the variance decomposition is twice this term. The correction uses 20 outcome-independent folds, treating mover matches as clusters and stayer person-years as singleton clusters. Panel A's control set is gender and age only. Panel B adds education, qualification, log employment, $\log(1 + \text{fixed assets})$, and $\log(1 + \text{intermediate inputs})$. Panel A is rerun on that same common Panel B leave-match-out sample, so the panels differ in controls rather than sample composition. Age is normalized at 50; the other continuous inputs are centered at their sample means, and all continuous inputs are scaled before the polynomial basis is constructed. The specification sequence is defined within each panel: *Linear additive controls* uses common quadratic and cubic age terms centered at 50, excludes the common linear age direction because it is not identified with worker and year fixed effects, and uses linear firm inputs in Panel B; *Heterogeneous linear controls* allows group-specific age deviations and, in Panel B, firm-input slopes; *Nonlinear homogeneous controls* uses a common degree-5 polynomial basis; and *Full model* interacts that basis with worker groups. The table reports point estimates only; no statistical standard errors or formal tests are reported. Bias-corrected components need not sum to one.

within-panel movements into stepwise changes along two paths—one adding heterogeneity first, the other nonlinearity first—both arriving at the full model. Appendix Table B7 compares the nonlinear estimates across polynomial, Hermite, and cubic B-spline bases at the same realized control dimension.

Panel A: parsimonious controls. Panel A is stable across the specification ladder. Worker variance ranges from 0.533 to 0.562, firm variance from 0.136 to 0.144, and sorting from 0.074 to 0.080. The residualized-wage variance ratio ranges from 0.956 to 0.999. Thus neither the standard age profile nor its heterogeneous and nonlinear extensions materially changes the decomposition on the common Panel B sample.

Panel B: worker and firm-input controls. Adding the common quadratic and cubic age terms, education, qualification, and the transformed firm inputs changes the decomposition already in the linear additive specification. Relative to the AKM baseline, the additive specification lowers worker variance from 0.551 to 0.474, firm variance from 0.144 to 0.121, and sorting from 0.080 to 0.047. Allowing heterogeneous linear slopes lowers worker

Table 4: Variance remaining after partialling out observables

Model	$\text{var}(\tilde{y}) / \text{var}(y)$
<i>Panel A: Parsimonious controls (Panel B sample)</i>	
AKM baseline	0.9985
Linear additive controls	0.9561
Heterogeneous linear controls	0.9642
Nonlinear homogeneous controls	0.9783
Full model	0.9913
<i>Panel B: Worker and firm-input controls</i>	
AKM baseline	0.9985
Linear additive controls	0.8271
Heterogeneous linear controls	0.7982
Nonlinear homogeneous controls	0.8319
Full model	0.8271

Notes: Entries report $\text{var}(\tilde{y}) / \text{var}(y)$ on the common Panel B leave-match-out estimation sample used in Table 3, where y is log hourly wage. Here $\tilde{y}$ is the AKM input outcome after partialling out observables and year effects, before the remaining variation is decomposed into worker, firm, and sorting components. This is a descriptive accounting statistic. It is not a bounded share of variance explained by the controls or the full residual from the joint worker–firm–controls–year regression. Values above one are possible when partialling out observables and year effects amplifies variance because of collinearity. No statistical standard errors are reported.

variance further to 0.451, while firm variance and sorting move only slightly, to 0.120 and 0.045.

A common degree-5 polynomial raises worker variance relative to the additive model, from 0.474 to 0.491, while slightly lowering firm variance to 0.120 and sorting to 0.041. Interacting the polynomial basis with worker groups yields worker variance of 0.490, firm variance of 0.117, and sorting of 0.042. The residualized-wage variance ratios are 0.827, 0.798, 0.832, and 0.827 for the additive, heterogeneous-linear, common-polynomial, and full specifications, respectively; the heterogeneous-linear specification has the smallest ratio.

What changes across specifications. Employment, fixed assets, and intermediate inputs are economically meaningful firm-side controls, but most of their transformed variation is cross-sectional. The largest movement in Panel B comes from adding observed controls. Relative to the baseline, the linear additive specification lowers worker variance by 0.077, firm variance by 0.022, and sorting by 0.034. Relative to that additive specification, the common polynomial raises worker variance by 0.018 while lowering firm variance by 0.002 and sorting by 0.006. Adding group interactions to the polynomial changes the three components by only -0.001 , -0.003 , and $+0.001$. The ladder therefore distinguishes a substantial observables adjustment from the smaller reallocation generated by functional-form flexibility.

Robustness to the choice of basis. The Panel B polynomial and Hermite estimates are nearly identical. Their common specifications give worker variance of 0.491 and 0.490, respectively, firm variance of 0.120 under both, and sorting of 0.041 under both. Their group-specific specifications also agree closely. The cubic B-spline gives a common decomposition of 0.492, 0.116, and 0.040, and a group-specific decomposition of 0.474, 0.114, and 0.041. Across the three group-specific bases, firm variance therefore lies between 0.114 and 0.117, and sorting between 0.041 and 0.042; worker variance ranges from 0.474 to 0.491.

Panel A shows the same distinction. Firm variance is 0.136 under every common basis and 0.137 under every group-specific basis, while sorting stays between 0.075 and 0.078. Worker variance moves more, ranging from 0.542 to 0.551 under the common bases and from 0.549 to 0.562 under the group-specific bases.

8.4. Interpreting the role of flexible controls

The within-panel ladder shows that adding observed characteristics materially changes the Panel B decomposition, while alternative functional forms produce smaller changes in its firm and sorting components. The bulk of the movement comes from including the firm inputs at all, not from how flexibly they enter: once they are in the model linearly, a degree-5 basis reallocates little. The basis comparison reinforces this conclusion for firm variance and sorting, but not for worker variance: the group-specific B-spline returns the worker component to the linear-additive level. For these covariates in Portuguese data, the standard linear-in-logs adjustment is therefore close to adequate for the firm and sorting components; flexible controls mainly reallocate the worker component.

The small changes in firm variance and sorting do not arise because the rich control design is mechanically unable to detect nonlinear wage schedules. Relative to heterogeneous linear controls, the fully interacted specification lowers mean squared error by 1.25% in sample and 1.08% out of sample, on a holdout fold drawn so that the training data preserve both regressor support and the connectedness of the worker–firm network. In a separate signal-recovery exercise, we inject a nonlinear component with variance equal to 5% of the restricted residual variance; the recovered incremental fitted component has a slope of 0.998 against the injected signal. The underlying fixed effects also move even when their aggregate second moments change little: worker effects covering 54.5% of observations and firm effects covering 24.0% move to a different effect decile. Appendix Table B8 reports these diagnostics. They indicate that incremental nonlinear wage variation is modest in the data and that coefficient reallocation need not translate one-for-one into firm variance or sorting.

The full specification allows the return to age and the transformed firm inputs to differ across demographic groups (gender in Panel A; gender $\times$ education $\times$ qualification in Panel B).

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. This is a different margin from the one studied by Card, Cardoso, and Kline (2016), who let the firm pay premium itself differ by worker group and find in Portuguese data that women receive about ninety percent of the premiums men receive; we let the *control function* vary by group while holding the firm effect common. Allowing firm-level interaction groups, or group-specific firm effects, would require a separate design. Third, our decomposition is static: time-varying firm policies or match dynamics can exist without affecting the estimated variance components.

9. Conclusions

AKM decompositions attach economic meaning to worker effects, firm effects, and sorting only after wages have been adjusted for observed characteristics. Generalized AKM estimates the original variance components while replacing the standard linear adjustment with an unknown smooth function that can interact with worker groups. The method therefore changes the treatment of observables without changing the decomposition that researchers want to interpret.

We establish when a semiparametric leave-out correction remains valid with heteroskedastic errors, many fixed effects, and a growing series basis. Because the targets are quadratic, approximation error enters both the plug-in decomposition and its bias correction. Controlling both terms requires more smoothness than estimation of a linear functional, but the condition is mild for the low-dimensional covariate functions common in wage applications. The fixed-rank and growing-rank limits cover the main variance-component cases, and the random-projection approximation makes the estimator feasible on large matched datasets. The simulations show that this flexibility removes bias under strong nonlinearity and costs little when the linear specification is adequate.

The Portuguese application separates two specification choices that are often combined. Adding education, qualification, and transformed firm inputs materially lowers all three variance components. Once those controls are present, nonlinearities and worker-group interactions produce smaller additional changes in firm variance and sorting: across the group-specific bases, firm variance remains between 0.114 and 0.117 and sorting between 0.041 and 0.042. Worker variance is more sensitive to the basis, ranging from 0.474 to 0.491. The parsimonious age-and-gender decomposition is stable throughout the specification ladder.

The scope of these conclusions is precise. The theory covers a fixed number of continuous covariates and observation-level leave-out inference. The application instead uses a cluster-fold point correction, so its cross-specification differences are descriptive and cluster-level inference remains open. The interactions vary across worker categories rather than firms, and the model does not estimate firm-specific pay schedules. Extensions to a growing covariate dimension or machine-learning nuisance functions require new arguments.

The empirical lesson is not that nonlinear controls overturn AKM. In this application, the main specification choice for measured firm variance and sorting is which observed characteristics are removed, not whether included characteristics enter linearly. Generalized AKM lets the data determine whether linearity is harmless or reallocates wage dispersion, while preserving the worker, firm, and sorting components at the center of the analysis.

References

- Abowd, J., F. Kramarz, and D. Margolis. 1999. “High wage workers and high wage firms.” *Econometrica* 67 (2): 251–333.
- Achlioptas, D. 2003. “Database-friendly random projections: Johnson-Lindenstrauss with binary coins.” *Journal of Computer and System Sciences* 66 (4): 671–687.
- 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.
- Bakirov, A., F. Del Prato, and P. Zacchia. 2026. *TWICE: Tree-based wage inference with clustering and estimation*. arXiv: 2601.00776 [econ.GN]. <https://arxiv.org/abs/2601.00776>.
- 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., K. Jochmans, and M. Weidner. 2025. *A Neyman-orthogonalization approach to the incidental parameter problem*. arXiv: 2412.10304 [econ.EM]. <https://arxiv.org/abs/2412.10304>.
- Bonhomme, S., T. Lamadon, and E. Manresa. 2019. “A distributional framework for matched employer employee data.” *Econometrica* 87 (3): 699–739.
- Breunig, C., and X. Chen. 2022. *Simple adaptive estimation of quadratic functionals in nonparametric IV models*. arXiv: 2101.12282 [math.ST]. <https://arxiv.org/abs/2101.12282>.
- 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., A. R. Cardoso, and P. Kline. 2016. “Bargaining, sorting, and the gender wage gap: quantifying the impact of firms on the relative pay of women.” *The Quarterly Journal of Economics* 131 (2): 633–686.
- 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, Jiafeng, and Jonathan Roth. 2024. “Logs with Zeros? Some Problems and Solutions.” *The Quarterly Journal of Economics* 139 (2): 891–936.

- Chen, X. 2007. “Large sample sieve estimation of semi-nonparametric models.” *Handbook of Econometrics* 6:5549–5632.
- Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. 2018. “Double/debiased machine learning for treatment and structural parameters.” *Econometrics Journal* 21 (1): 1–68.
- Donald, S. G., and W. K. Newey. 1994. “Series estimation of semilinear models.” *Journal of Multivariate Analysis* 50 (1): 30–40.
- Farrell, M. H., T. Liang, and S. Misra. 2021. “Deep neural networks for estimation and inference.” *Econometrica* 89 (1): 181–213.
- Fisher, R. 1928. *Statistical Methods for Research Workers*. 5th ed. Oliver / Boyd.
- Hirano, K., G. W. Imbens, and G. Ridder. 2003. “Efficient estimation of average treatment effects using the estimated propensity score.” *Econometrica* 71 (4): 1161–1189.
- Kline, P. 2024. “Firm wage effects.” In *Handbook of Labor Economics*, edited by C. Dustmann and T. Lemieux, 5:115–181. Elsevier. <https://doi.org/10.1016/bs.heslab.2024.11.005>.
- Kline, P., R. Saggio, and M. Sølvesten. 2020. “Leave-out estimation of variance components.” *Econometrica* 88 (5): 1859–1898.
- Mincer, J. 1974. *Schooling, Experience, and Earnings*. National Bureau of Economic Research, Inc.
- Scheffé, H. 1959. *The Analysis of Variance*. New York: John Wiley & Sons.
- Searle, S. R., G. Casella, and C. E. McCulloch. 2009. *Variance Components*. 2nd ed. Hoboken, NJ: John Wiley & Sons.
- Sølvesten, M. 2020. “Robust estimation with many instruments.” *Journal of Econometrics* 214 (2): 495–512.

A. Proofs

This appendix collects the proofs. We first record the leave-one-out residual-variance identity used throughout, then prove the consistency and limiting-distribution results in the order they appear in the paper: Lemma 1, Lemma 2, Theorem 3, and Theorem 4.

Leave-one-out residual variance. The representation in (5) holds by applying 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

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

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 (4), we have

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

Proof of Lemma 1. The variance of $\hat{\beta}$ is

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

under the assumption of fixed x_i and z_i .

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

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

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

$$\begin{aligned}\hat{\theta} - \theta &= \sum_{i=1}^n \sum_{j=1}^n M_{ij} (x_j' \beta + f(z_j) + e_j) x_i' S_{xx}^{-1} A S_{xx}^{-1} \sum_{a=1}^n \sum_{b=1}^n M_{ab} x_a (x_b' \beta + f(z_b) + e_b) - \beta' A \beta - \sum_{i=1}^n B_{ii} \hat{\sigma}_i^2 \\ &= \sum_{i=1}^n \sum_{j=1}^n M_{ij} x_j' \beta x_i' S_{xx}^{-1} A S_{xx}^{-1} \sum_{a=1}^n \sum_{b=1}^n M_{ab} x_a x_b' \beta + \sum_{i=1}^n \sum_{j=1}^n M_{ij} f(z_j) x_i' S_{xx}^{-1} A S_{xx}^{-1} \sum_{a=1}^n \sum_{b=1}^n M_{ab} x_a f(z_b) \\ &\quad + \sum_{i=1}^n \sum_{j=1}^n M_{ij} e_j x_i' S_{xx}^{-1} A S_{xx}^{-1} \sum_{a=1}^n \sum_{b=1}^n M_{ab} x_a e_b + \sum_{i=1}^n \sum_{j=1}^n M_{ij} x_j' \beta x_i' S_{xx}^{-1} A S_{xx}^{-1} \sum_{a=1}^n \sum_{b=1}^n M_{ab} x_a f(z_b) \\ &\quad + \sum_{i=1}^n \sum_{j=1}^n M_{ij} f(z_j) x_i' S_{xx}^{-1} A S_{xx}^{-1} \sum_{a=1}^n \sum_{b=1}^n M_{ab} x_a x_b' \beta + \sum_{i=1}^n \sum_{j=1}^n M_{ij} x_j' \beta x_i' S_{xx}^{-1} A S_{xx}^{-1} \sum_{a=1}^n \sum_{b=1}^n M_{ab} x_a e_b \\ &\quad + \sum_{i=1}^n \sum_{j=1}^n M_{ij} e_j x_i' S_{xx}^{-1} A S_{xx}^{-1} \sum_{a=1}^n \sum_{b=1}^n M_{ab} x_a x_b' \beta + \sum_{i=1}^n \sum_{j=1}^n M_{ij} f(z_j) x_i' S_{xx}^{-1} A S_{xx}^{-1} \sum_{a=1}^n \sum_{b=1}^n M_{ab} x_a e_b \\ &\quad + \sum_{i=1}^n \sum_{j=1}^n M_{ij} e_j x_i' S_{xx}^{-1} A S_{xx}^{-1} \sum_{a=1}^n \sum_{b=1}^n M_{ab} x_a f(z_b) - \beta' A \beta - \sum_{i=1}^n B_{ii} \hat{\sigma}_i^2.\end{aligned}$$

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

$$\begin{aligned}
\hat{\theta} - \theta &= \sum_{i=1}^n \sum_{\ell=1}^n B_{i\ell} f(z_i)^2 + \sum_{i=1}^n \sum_{\ell \neq i}^n B_{i\ell} e_i e_\ell + \sum_{i=1}^n B_{ii} (e_i^2 - \hat{\sigma}_i^2) \\
&\quad + 2 \sum_{i=1}^n \sum_{\ell=1}^n B_{i\ell} x'_\ell \beta f(z_i) + 2 \sum_{i=1}^n \sum_{\ell=1}^n B_{i\ell} x'_\ell \beta e_i + 2 \sum_{i=1}^n \sum_{\ell=1}^n B_{i\ell} e_\ell f(z_i).
\end{aligned} \tag{A1}$$

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 (A1) 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 $n \rightarrow \infty$, $k \rightarrow \infty$, and $p \rightarrow \infty$.

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

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

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-Schwarz inequality,

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

Thus, using the Assumption 1 and 3, we have

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

Now, for the second term in the main decomposition,

$$\begin{aligned}
\sum_{i=1}^n \sum_{\ell=1}^n B_{i\ell} x'_\ell \beta f(z_i) &= F' B X \beta \\
&= F' M X S_{xx}^{-1} A S_{xx}^{-1} X' M X \beta \\
&= F' M X S_{xx}^{-1} A \beta.
\end{aligned}$$

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

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

So that using the Assumption 1 and 3, we have

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

To bound the third term, we should show that

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

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

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

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

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

and, similarly,

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

so that using Assumption 1 and 3, we ultimately have:

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

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

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 $\lambda_{\max}(\tilde{A})$ is the largest eigenvalue of $\tilde{A}$. From this it follows that we can bound

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

use assumptions $\theta = \mathcal{O}(1)$, $\lambda_1 \leq \text{trace}(\tilde{A}^2)^{1/2} = o(1)$, 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 $i = 1, \dots, n$ 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

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

Thus, the variance is now expressed as:

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

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 $\max_i (x_i' \beta + f(z_i))^2 \leq 2 \max_i (x_i' \beta)^2 + 2 \max_i f(z_i)^2 = \mathcal{O}(1)$.

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 2. 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

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

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

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

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

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

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 $k \rightarrow \infty$,

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

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

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

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

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:

$$\begin{aligned} \text{var} \left[\sum_{i=1}^n \hat{\sigma}_i^2 \hat{B}_{ii} (\hat{a}_i^2 - \text{var}[\hat{a}_i]) \right] &= \sum_{i=1}^n \sum_{\ell=1}^n \mathbb{E}[\hat{B}_{ii} \hat{B}_{\ell\ell}] \mathbb{E}[\hat{\sigma}_i^2 \hat{\sigma}_\ell^2] \text{cov}[\hat{a}_i^2, \hat{a}_\ell^2] \\ &= \mathcal{O} \left(m^{-2} \text{trace}(\tilde{A}^2) + m^{-3} \text{trace}(\tilde{A}_1^2)^{1/2} \text{trace}(\tilde{A}_2^2)^{1/2} \right), \\ \text{var} \left[\sum_{i=1}^n \hat{\sigma}_i^2 (\hat{B}_{ii} - B_{ii}) \frac{2 \sum_{\ell \neq i}^n P_{W,i\ell}^4 - P_{W,ii}(1 - P_{W,ii})^2}{m(1 - P_{W,ii})^2} \right] &= \mathcal{O} \left(m^{-3} \text{trace}(\tilde{A}^2) \right), \\ \text{var} \left[\sum_{i=1}^n B_{ii} (\hat{\sigma}_i^2 - \sigma_i^2) \frac{2 \sum_{\ell \neq i}^n P_{W,i\ell}^4 - P_{W,ii}(1 - P_{W,ii})^2}{m(1 - P_{W,ii})^2} \right] &= \mathcal{O} \left(m^{-2} \text{var}[\hat{\theta}] \right). \end{aligned}$$

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

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:

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

□

Proof of Theorem 3. 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

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

We prove the theorem in three steps.

Approximation. Equivalently, represent $\hat{\theta}$ as

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

and below we show that the second term is asymptotically dominated by the variance of the estimator, $\text{var}[\hat{\theta}]$. 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 1 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

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

The variances of the first and the third terms are

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

and of the second and the fourth

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

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

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 $\text{var}[\hat{\theta}]^{-1} \sum_{i=1}^n B_{ii}^2 = o(1)$. It holds because

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

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

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

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

$$\delta(\vartheta) := \sum_{i=1}^n v_i(\vartheta)(\hat{\sigma}_i^2 - \sigma_i^2), \quad (\text{A2})$$

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-Schwarz inequality, and $|\mathbb{E}[\hat{\sigma}_i^2 - \sigma_i^2]| = o_p(1)$,

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

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

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

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

Asymptotic normality. Our objective is to prove that

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

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

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

Because we have that

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

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

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

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 $\text{var}[\vartheta' \hat{b}]^{-2} \leq \max_i \sigma_i^{-2} = \mathcal{O}(1)$. $\square$

Proof of Theorem 4. 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 Solvsten (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 $\text{var}[q_{n,i}] = \sigma_{n,i}^2$, let $\{\dot{w}_{n,i}\}_{i,n}$ be a triangular array of non-random weights that satisfy $\sum_{i=1}^n \dot{w}_{n,i} \sigma_{n,i}^2 = 1$ 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 $2 \sum_{i=1}^n \sum_{\ell \neq i} Q_{n,i\ell}^2 \sigma_{n,i}^2 \sigma_{n,\ell}^2 = 1$. 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 3. *If $\max_i \mathbb{E}[q_{n,i}^4] + \sigma_{n,i}^{-2} = O(1)$, (i) $\max_i \dot{w}_{n,i}^2 = o(1)$, and (ii) $\text{trace}(Q_n^4) = o(1)$, then $(\mathcal{S}_n, \mathcal{U}_n)' \xrightarrow{d} \mathcal{N}(0, I_2)$.*

Proof. See Appendix B in Kline, Saggio, and Solvsten (2020) and Appendix A2 in Solvsten (2020). $\square$

The U -statistic representation holds because

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

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

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

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

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

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

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

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

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

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

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

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

with $\omega_1 := \text{var}[\hat{\theta}]^{-1/2} \mathcal{V}_s^{1/2}$, $\omega_2 := \text{var}[\hat{\theta}]^{-1/2} \mathcal{V}_u^{1/2}$, 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 $\text{var}[\hat{\theta}]^{-1/2} (\hat{\theta} - \theta)$ is implied by asymptotic normality of $\mathcal{U}_n$. Using notation of Lemma 3, we have that $\check{w}_i = \mathcal{V}_s^{-1/2} (2\tilde{w}'_i \gamma - \check{w}'_i \gamma)$, and $Q_{i\ell} = \mathcal{V}_u^{-1/2} C_{i\ell}$.

Table B1: Finite-sample coverage diagnostic for the degree-5 leave-out estimator

Scenario	95% coverage (LOO(5))
<i>Approximation and nonlinearity</i>	
Mild nonlinearity	0.93
Linear benchmark	0.93
Strong nonlinearity	0.90
<i>Strong radial design: one-feature changes</i>	
Heteroskedasticity	0.90
Heavy-tailed errors	0.91
Higher nuisance dimension	NA
Many regressors	0.97
High leverage	0.86
Larger sample	0.95

Notes: Entries report the share of 5,000 Monte Carlo replications in which the nominal 95% confidence interval for the degree-5 leave-out estimator contains the quadratic-form target. The intervals use the estimated noncentral variance of the quadratic form and leave-one-out residual variances $\hat{\sigma}_i^2$. *NA* means that the complete degree-5 basis is not estimable. This table is a finite-sample diagnostic, not an applied standard-error claim.

To verify the condition (i) of Lemma 3, note that

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

where the last equality follows from Theorem 4 (i), and the nonzero limit of ω_1 .

For the condition (ii) of Lemma 3, denote $\tilde{A}_W := S_{ww}^{-1/2} \tilde{A} S_{ww}^{-1/2}$, 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 4 (ii). $\square$

B. Additional Tables

Table B1 reports the finite-sample coverage diagnostic for the degree-5 leave-out estimator. Tables B2–B4 document the data and sample restrictions behind the estimation sample. Table B5 documents the firm controls that drive the richer specification, and Tables B6–B9 report the specification ladder, its sensitivity to the nonlinear basis, the validation exercises, and numerical diagnostics.

Table B2: Descriptive statistics of the cleaned worker–firm panel

	Mean	SD	p25	Median	p75
<i>A. Worker-year panel</i>					
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
<i>B. Worker composition and mobility</i>					
Job-to-job moves (count)	1,803,534	–	–	–	–
Job-to-job movers (%)	9.85%	–	–	–	–
Female workers (%)	42.14%	–	–	–	–
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%	–	–	–	–
<i>C. Firm-year panel</i>					
Workers	13.35	119.97	3.00	4.00	9.00
Fixed assets (M€)	0.72	26.38	0.01	0.03	0.14
Intermediate inputs (M€)	0.72	22.49	0.03	0.06	0.19
<i>D. Panel counts</i>					
Unique workers	3,532,497	–	–	–	–
Unique firms	484,704	–	–	–	–
Worker-year observations	18,314,740	–	–	–	–

Notes: The table uses the cleaned full-time worker–firm panel for workers aged 20–65 in 2008–2018, before the non-missing-control, connected-component, and leave-match-out restrictions. Panels A and C report worker-year and firm-year statistics, respectively: means, standard deviations, and the 25th, 50th, and 75th percentiles. Units are shown in the row labels. Panel B reports counts and sample shares; its percentages use worker-year rows as the denominator, the job-to-job move share uses non-missing switch indicators, and categorical missing values are shown explicitly. Panel D reports counts.

Table B3: Missingness of estimation variables

Variable	Missing share
Age	0.00%
Education	0.32%
Firm ID	0.00%
Fixed assets	25.50%
Intermediate inputs	8.26%
Log hourly wage	0.01%
Qualification	11.88%
Gender	0.00%
Worker ID	0.00%
Workers	8.26%
Year	0.00%

Notes: Entries report the share of worker-year observations in the cleaned analysis sample (2008–2018) with missing values for each variable required by at least one empirical specification, including the outcome, identifiers, year, and controls. This table uses worker-year denominators and is computed before the non-missing-control, connected-component, and leave-match-out restrictions. The firm-control diagnostic in Table B5 uses firm-year denominators and an imputed employment field, so its missingness rates are not directly comparable. The main empirical tables use the Panel B leave-match-out sample described in Table B4.

Table B4: Sample flow into the leave-match-out estimation set

Stage	<i>N</i>	Workers	Firms	Movers	Matches
<i>Panel A: Parsimonious controls</i>					
Loaded QP rows	23,304,646	4,217,716	533,360	1,758,967	7,104,038
Cleaned analysis sample	18,314,740	3,532,497	484,704	1,231,386	5,291,365
Non-missing controls	18,312,694	3,532,350	484,696	1,231,231	5,290,982
Largest connected component	17,256,654	3,282,116	333,180	1,212,893	5,021,308
Leave-match-out set	15,232,458	2,432,843	193,487	1,075,443	3,969,503
<i>Panel B: Worker and firm-input controls</i>					
Loaded QP rows	23,304,646	4,217,716	533,360	1,758,967	7,104,038
Cleaned analysis sample	18,314,740	3,532,497	484,704	1,231,386	5,291,365
Non-missing controls	11,892,426	2,642,775	296,549	663,784	3,487,931
Largest connected component	10,891,978	2,392,780	180,478	653,930	3,227,750
Leave-match-out set	9,439,305	1,719,643	105,442	578,409	2,457,963

Notes: The table reports sample flow from loaded QP rows through the cleaned worker-year panel, non-missing controls, the largest worker–firm connected component, and the final leave-match-out set used by the full-regressor cluster-fold point correction. *N* is the number of observations at each stage; Workers and Firms are unique identifiers, Movers are workers observed at more than one firm, and Matches are unique worker–firm pairs. The Panel A rows describe the larger parsimonious candidate sample. The main empirical tables use the Panel B leave-match-out set for both panels.

Table B5: Firm controls: within-firm time variation and missingness

	Non- missing firm-year	Firms ($\geq 2y$)	Mean years (firms $\geq 2y$)	Within share (%)	Zero within SD (%)
<i>A. Firm-control time variation in estimation sample</i>					
Employment	2,397,757	383,679	5.99	7.88%	27.11%
Fixed assets	1,504,836	253,373	5.71	17.75%	5.99%
Intermediate in-puts	1,874,310	290,932	6.22	9.75%	0.72%
<i>B. Missingness by year (firm-year panel)</i>					
Year	Firm-years	Workers miss.	Fixed assets miss.	Intermed. inputs miss.	All controls obs.
2008	253,005	0.00%	100.00%	26.27%	0.00%
2009	245,077	0.00%	100.00%	25.37%	0.00%
2010	223,267	0.00%	23.66%	23.66%	76.34%
2011	218,229	0.00%	23.01%	23.01%	76.99%
2012	202,744	0.00%	22.47%	22.47%	77.53%
2013	198,597	0.00%	21.20%	21.20%	78.80%
2014	202,625	0.00%	20.36%	20.36%	79.64%
2015	207,037	0.00%	19.89%	19.89%	80.11%
2016	211,892	0.00%	19.39%	19.39%	80.61%
2017	215,908	0.00%	18.80%	18.80%	81.20%
2018	219,376	0.00%	18.24%	18.24%	81.76%

Notes: Panel A uses the cleaned 2008–2018 firm-year panel used to form the estimation sample, not the final leave-match-out sample. It reports within-firm time variation for the Panel B firm controls, restricting to firms observed for at least two years. Each row uses its own non-missing firm-years, so row counts differ across controls and need not equal the joint estimation sample. The within-firm share is the share of total variance on the estimation scale attributable to within-firm changes: log employment and $\log(1+x)$ for fixed assets and intermediate inputs. Zero within SD is the share of firms with at least two observed years and zero within-firm standard deviation of the transformed control. Panel B reports year-by-year missingness rates in the firm-year panel. The worker-count column has zero missingness because employment is imputed from the matched employer–employee data; fixed assets are unavailable in 2008–2009 and are about 18–24% missing thereafter; intermediate inputs are about 18–26% missing throughout the sample. The final column reports the share of firm-years with workers, fixed assets, and intermediate inputs all observed. These firm-year rates are not directly comparable to the worker-year rates in Table B3.

Table B6: Stepwise changes in wage variance decomposition

Change	$\Delta\sigma_\alpha^2$	$\Delta\sigma_\psi^2$	$\Delta\text{cov}(\alpha, \psi)$
<i>Panel A: Parsimonious controls (Panel B sample)</i>			
Add linear observables	-0.0181	-0.0068	-0.0065
Add heterogeneity at degree 1	0.0061	-0.0001	0.0010
Add nonlinearity without interactions	0.0157	-0.0009	0.0037
Add heterogeneity in nonlinear basis	0.0136	0.0013	-0.0008
Add nonlinearity with interactions	0.0232	0.0005	0.0018
<i>Panel B: Worker and firm-input controls</i>			
Add linear observables	-0.0771	-0.0221	-0.0338
Add heterogeneity at degree 1	-0.0230	-0.0016	-0.0019
Add nonlinearity without interactions	0.0177	-0.0016	-0.0056
Add heterogeneity in nonlinear basis	-0.0009	-0.0032	0.0010
Add nonlinearity with interactions	0.0398	-0.0032	-0.0026

Notes: Entries are next-specification minus comparison-specification differences in bias-corrected variance-component shares from the full-regressor cluster-fold point correction on the common Panel B leave-match-out estimation sample. The five rows are, respectively: linear additive controls minus the AKM baseline; heterogeneous linear controls minus linear additive controls; nonlinear homogeneous controls minus linear additive controls; the full model minus nonlinear homogeneous controls; and the full model minus heterogeneous linear controls. A positive value denotes an increase. Starting from the linear additive specification, one path adds heterogeneity at degree 1 and then nonlinearity with interactions; the other adds nonlinearity without interactions and then heterogeneity in the nonlinear basis. Both paths arrive at the full model. These are point-estimate differences; no statistical standard errors or formal tests are reported.

Table B7: Alternative series bases and the wage variance decomposition

Basis	Nonlinear specification	σ_α^2	σ_ψ^2	$\text{cov}(\alpha, \psi)$	$\text{var}(\tilde{y}) / \text{var}(y)$
<i>Panel A: Parsimonious controls (Panel B sample)</i>					
Degree-5 polynomial	Common	0.5483	0.1358	0.0775	0.9783
Degree-5 polynomial	Group-specific	0.5619	0.1371	0.0767	0.9913
Degree-5 Hermite	Common	0.5418	0.1358	0.0770	0.9709
Degree-5 Hermite	Group-specific	0.5491	0.1371	0.0757	0.9765
Cubic B-spline	Common	0.5505	0.1358	0.0776	0.9808
Cubic B-spline	Group-specific	0.5502	0.1371	0.0751	0.9764
<i>Panel B: Worker and firm-input controls</i>					
Degree-5 polynomial	Common	0.4913	0.1199	0.0410	0.8319
Degree-5 polynomial	Group-specific	0.4904	0.1166	0.0420	0.8271
Degree-5 Hermite	Common	0.4900	0.1199	0.0409	0.8306
Degree-5 Hermite	Group-specific	0.4908	0.1165	0.0424	0.8286
Cubic B-spline	Common	0.4925	0.1159	0.0400	0.8276
Cubic B-spline	Group-specific	0.4740	0.1143	0.0411	0.8070

Notes: Entries report full-regressor cluster-fold bias-corrected variance-component shares and the residualized-wage variance ratio on the common Panel B leave-match-out estimation sample. *Common* denotes the nonlinear homogeneous specification; *Group-specific* denotes the full specification that interacts the basis with worker groups. The residualized-wage ratio uses $\tilde{y}$ as defined in the note to Table 4. The polynomial and Hermite specifications use complete degree-5 multivariate bases. The B-spline specification is additive across continuous inputs and uses cubic bases with quantile-spaced knots; its spline counts match the polynomial model's realized control dimension. All specifications use the same 20 outcome-independent folds, treating mover matches as clusters and stayer person-years as singleton clusters. The randomized correction satisfies a numerical-precision rule: the Monte Carlo standard error of each correction term is at most 6×10^{-4} of total log-wage variance. This is a numerical-precision check, not a statistical inference statement. The table reports point estimates only.

Table B8: Validation of the flexible covariate specification

Diagnostic	Worker effect	Firm effect
<i>Panel A. Incremental wage fit</i>		
In-sample partial R^2 : homogeneous nonlinear vs. additive linear	0.39%	
In-sample partial R^2 : interacted nonlinear vs. heterogeneous linear	1.25%	
Held-out partial R^2 : interacted nonlinear vs. heterogeneous linear	1.08%	
Held-out observations	94,493	
<i>Panel B. Injected nonlinear signal</i>		
Signal variance relative to restricted residual variance	5.0%	
Recovery slope	0.998	
Recovery correlation	0.994	
<i>Panel C. Fixed-effect reallocation, heterogeneous linear to full model</i>		
RMS coefficient change	0.178	0.033
Coefficient correlation	0.956	0.994
Observation weight attached to effects changing decile	54.5%	24.0%

Notes: These diagnostics use the common Panel B leave-match-out sample. Panel A compares the nonlinear specifications with their nested linear benchmarks. Partial R^2 is one minus the ratio of the rich-model mean squared error to the restricted-model mean squared error. The held-out result uses 94,493 observations from one held-out fold of an outcome-independent, graph-aware 100-fold assignment; the assignment preserves regressor support and connectedness in the training sample. Panel B injects a nonlinear component that is orthogonal to the heterogeneous-linear design and has variance equal to 5% of that model's residual variance. The recovery statistics compare the rich model's incremental fitted component with the injected signal. Panel C compares the plug-in worker and firm coefficients from the heterogeneous-linear and full joint regressions. The root mean squared (RMS) changes and decile movements use observation weights. These diagnostics are descriptive and are not the bias-corrected variance components reported in the main tables; no statistical standard errors are reported.

Table B9: Control dimension and specification diagnostics

Model	k
<i>Panel A: Parsimonious controls (Panel B sample)</i>	
AKM baseline	0
Linear additive controls	2
Heterogeneous linear controls	3
Nonlinear homogeneous controls	4
Full model	9
<i>Panel B: Worker and firm-input controls</i>	
AKM baseline	0
Linear additive controls	10
Heterogeneous linear controls	102
Nonlinear homogeneous controls	129
Full model	3004

Notes: k counts the observed-control and basis columns in the joint augmented regression; worker, firm, and year fixed-effect columns are excluded. Both panels use the common Panel B leave-match-out sample; Panel A changes the controls only. This table reports control dimension only. Sample-flow counts are reported separately in Table B4.